

Article

Not peer-reviewed version

Validation of the Transformer-Based Monocular System (Capture4D): A Real-time Kinematic Analysis in Coaching/Teaching Tennis

[YUE ZHAO](#)^{*}, [Shuo Wang](#), [Zan Gao](#), Haijun Wu, [Yixiong Cui](#), [Yuanlong Liu](#)^{*}

Posted Date: 19 December 2025

doi: 10.20944/preprints202512.1731.v1

Keywords: computer vision; human motion capture; tennis; monocular camera; transformer



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Validation of the Transformer-Based Monocular System (Capture4D): A Real-time Kinematic Analysis in Coaching/Teaching Tennis

Yue Zhao ^{1,*}, Shuo Wang ², Zan Gao ³, Haijun Wu ⁴, Yixiong Cui ⁵ and Yuanlong Liu ^{6,*}

- ¹ Sports Coaching College, Beijing Sport University, Beijing 100084, China
- ² Beijing PinXi Technology Co., Ltd, Beijing 100089, China
- ³ Department of Kinesiology, Recreation, and Sport Studies, The University Tennessee, Knoxville, TN, 37996, USA
- ⁴ China Institute of Sport and Health Science, Beijing Sport University, Beijing 100084, China
- ⁵ School of Sports Engineering (China Sports Big Data Center), Beijing Sport University, Beijing 100084, China
- ⁶ Department of Human Performance and Health Education, Western Michigan University, Kalamazoo Michigan 49008, USA
- * Correspondence: zhaoyue@bsu.edu.cn (Y.Z.); yuanlong.liu@wmich.edu (Y.L.)

Abstract

Human motion capture is crucial for various fields, but traditional optical systems (OMC) are costly and restrictive. Monocular video-based methods offer accessibility, yet face accuracy challenges, especially in dynamic sports like tennis. This study validates Capture4D, a novel Transformer-based monocular system, for capturing a wide range of tennis strokes. We developed a universal biomechanical analysis framework (K0-K5) applicable to twelve fundamental stroke types. To demonstrate the system’s capabilities, this paper focuses on a detailed validation using the tennis serve as a representative example. We conducted experiments with 9 high-level tennis players, and motion data were simultaneously captured using Capture4D (single RGB camera) and OMC Qualisys (gold standard). Accuracy was evaluated by comparing 3D joint coordinates and joint angles using Normalized Mean Per Joint Position Error (NMPJPE), RMSE and MAE. The results showed that Capture4D effectively captured the tennis player’s motion, with average NMPJPE for tennis serves ranging from 69.5mm to 88.3mm, within the acceptable range (70-130mm) for coaching purpose. Compared to OMC, Capture4D demonstrated comparable joint angle trajectories, with advantages in operational convenience, cost-effectiveness, and wider applicability. It offered approximately 50% reduction in setup time and 80% cost savings. Capture4D presents a valid and practical monocular motion capture solution for coaching tennis and other broader applications in sports. While slightly less precise than OMC, its accuracy is acceptable for many use cases in coaching and teaching. It offers significant advantages in convenience and cost, paving the way for accessible motion analysis in diverse environments like outdoor settings and multi-person scenarios, in which OMC is not possible to be used. This technology holds promise for democratizing motion capture in sports training and coaching/teaching.

Keywords: computer vision; human motion capture; tennis; monocular camera; transformer

1. Introduction

Quantitative analysis of human movement is a cornerstone of modern sports science, providing critical insights for performance enhancement and injury prevention. In net sports such as tennis, the precise measurement of three-dimensional (3D) kinematics—including joint angles, segment velocities, and spatial positioning—is essential for coaches and scientists to understand and optimize complex motor skills. For decades, Optical Motion Capture (OMC) systems have been accepted as a

“gold standard” for this purpose, offering unparalleled accuracy in controlled laboratory environments (Bartlett 2007).

However, the high cost, operational complexity, and environmental constraints of OMC systems create a significant “lab-to-field” gap. The data captured in a sterile lab setting often fails to represent the variable and dynamic conditions of on-court performance, limiting the ecological validity and practical applicability of the findings for coaches and athletes (Hughes, Arapakis, and McNeill 2019). Factors such as marker slippage during high-intensity movements, the psychological effect of being in a lab environment, and the inability to capture data during actual competition scenarios are well-known limitations. This gap highlights a pressing need for accessible, portable, and cost-effective motion analysis tools that can be deployed “in the wild” without encumbering the athlete or disrupting the natural training/competition environment.

Marker-less motion capture systems, particularly those using a single, readily available video camera (monocular systems), present a promising solution to bridge this gap. The advent of deep learning, particularly Convolutional Neural Networks (CNNs), revolutionized the field by enabling robust 2D key point detection from images, with systems such as OpenPose have become widely adopted (Cao et al. 2017) in coaching and training environment.

More recently, the Transformer architecture, which has shown remarkable success by modeling global context through self-attention mechanisms (Vaswani et al. 2017). it has been adapted for computer vision tasks. This study focused on validating Capture4D, a novel monocular motion capture system that leverages a powerful Transformer-based architecture, similar in principle to pioneering works like 4DHumans (Goel et al. 2023). Unlike traditional CNNs, this architecture is specifically designed to better model the holistic relationship among all body segments simultaneously, offering the potential for more robust 3D pose estimation, even during complex and rapid athletic movements in competitions.

The purpose of this research was to conduct a rigorous validation of the Capture4D system for 3D kinematic analysis across a comprehensive tennis strokes. This paper presents the detailed validation for the serve, which represents a particularly challenging use-case due to its high velocity and complexity. To our best knowledge, this study could the first to systematically validate a Transformer-based monocular motion capture system for on-court tennis stroke analysis. It provided a practical and accessible solution to bridge the lab-to-field gap in sport analysis for coaches, teachers, and athletes. Our objective was to determine if Capture4D could provide an acceptable level of accuracy in coaching/teaching and biomechanical research. It could offer a practical and accessible alternative to fill the gap between traditional laboratory-based and on court environment evaluation. We hypothesized that Capture4D would achieve sufficient accuracy in joint kinematics to enable in-field coaching/teaching applications, with temporal agreement comparable to gold-standard OMC systems.

2. Methods

2.1. The Capture4D System: Core Architecture and Training

Capture4D, is a markerless monocular human motion capture system based on a Transformer deep learning architecture. Its primary objective is to reconstruct temporally coherent 3D human pose and shape from 2D video recorded by a single, standard camera.

2.1.1. Model Architecture

The technical core of the system is a fully “Transformerized” neural network, which includes a Vision Transformer (ViT) encoder and a cross-attention-based Transformer decoder:

1. ViT Encoder: The system employs a powerful ViT-H/16 model as its image encoder. For each input video frame, the encoder first decomposes it into a series of image patches and then extracts image features representing global contextual information through its multi-layer self-attention mechanism. This enables the model to effectively capture long-range spatial

dependencies among different body parts, which is crucial for handling complex poses and severe occlusions.

2. Transformer Decoder: The image features extracted by the encoder are fed into a 6-layer decoder. The decoder interacts with the image features via a cross-attention mechanism, using a learnable query that represents the human model. It ultimately regresses a set of SMPL (Skinned Multi-Person Linear Model) parameters, including pose parameters (θ) defining body posture, shape parameters (β) defining height and build, and camera parameters (π) relative to the person.

2.1.2. Training Data

To ensure robustness and generalization in real-world scenarios, Capture4D's training process utilizes a large-scale, diverse hybrid dataset strategy. This includes: (1) datasets with precise 3D ground truth, such as Human3.6M (Ionescu et al. 2013) and MPI-INF-3DHP (Mehta et al. 2017), for learning accurate 3D human structure; and (2) large-scale "in-the-wild" image datasets with only 2D annotations, such as COCO (Lin et al. 2014), AI Challenger (Wu et al. 2017), and InstaVariety (Kanazawa et al. 2019), to enhance the model's recognition capabilities under complex backgrounds, lighting, and clothing conditions.

2.2. Experimental Validation: Participants and Protocol

2.2.1. Participants

Nine (N=9) high-level tennis players were recruited for this study. All participants were right-handed, with an average of 12 years of professional training experience (training duration: 12 ± 3 years, where " ± 3 years" represents the standard deviation), indicating a solid and systematic professional training background. Given the scarcity of the high-level professional tennis player population and the high demand for data quality, similar as this type of studies in the literature, the sample size of 9 is practically feasible and scientifically reasonable[3.1]. Prior to data collection, all participants were fully informed of the experimental procedures and objectives and provided written informed consents. This study protocol was approved on 9 October 2025, by the Beijing Sport University Institutional Review Board (BSU IRB) with CIRBID 2025426H.

2.2.2. Apparatus and Protocol

Since the OMC cannot be set up outdoors, to compare the two systems, this study was conducted indoors in an open room of approximately 10 meters by 10 meters with a flat, non-slip floor. The experimental environment was set up with sufficient, uniform lighting and a simple background.

The following equipment was used:

1. Monocular Camera: Used to record the participants' motion videos as input for the Capture4D system.
2. Optical Motion Capture (OMC) System: A Qualisys system was used as the "gold standard" for data collection to validate the accuracy of the Capture4D system. The OMC system includes high-speed cameras and is equipped with professional capture and data processing software.
3. Reflective Markers: Used by the OMC system to capture human motion information. Markers were placed on key anatomical landmarks, such as the head, shoulders, elbows, wrists, hips, knees, and ankles.
4. Data Acquisition Computer: Used to store and process experimental data, configured to meet the operational requirements of both the monocular camera and the OMC system.
5. Sports Equipment: A tennis racket was used to assist participants in completing the specified actions.

During the data collection phase, participants performed tennis serves.

2.3. Experimental Validation: Data Processing and Analysis

The data analysis workflow was designed to achieve two core objectives: (1) to quantitatively validate the overall 3D reconstruction accuracy of the Capture4D system; and (2) to evaluate its effectiveness in analyzing key biomechanical events in practical sports science applications.

1. Data Preprocessing: The raw 3D coordinate data exported from both systems were first temporally synchronized. To reduce the impact of high-frequency noise on kinematic calculations, a Savitzky-Golay smoothing filter was applied to the joint trajectories output by the Capture4D system.
2. Overall Accuracy Validation: To quantify the overall 3D reconstruction accuracy of the Capture4D system, we used the Normalized Mean Per Joint Position Error (NMPJPE) as the primary metric. This method involves aligning the Capture4D-predicted skeleton of each frame with the gold-standard skeleton measured by the Qualisys system through a rigid transformation (rotation, translation, and scaling) via Procrustes analysis. The average Euclidean distance between all corresponding joints is then calculated. This method eliminates systematic errors arising from different coordinate system definitions and scale differences, thereby enabling a fair assessment of the model's intrinsic accuracy.
3. Applied Biomechanical Analysis: A core component of our validation was the development of a universal analysis framework applicable to all captured movements. We segmented each of the twelve stroke types into five biomechanically significant key phases (K0-Preparation, K1-Backswing start, K2-Peak of Backswing, K3-Contact Point, K4-Follow-through and K5-Recovery). The identification of these keyframes was based on kinematic criteria derived from the motion data itself, ensuring an objective and reproducible segmentation of the movement. The phases are defined as follows:

The entire data processing and analysis pipeline, summarized in Figure 1, was implemented using the Python programming language, in conjunction with scientific computing and visualization libraries such as Pandas, NumPy, joblib, and Matplotlib.

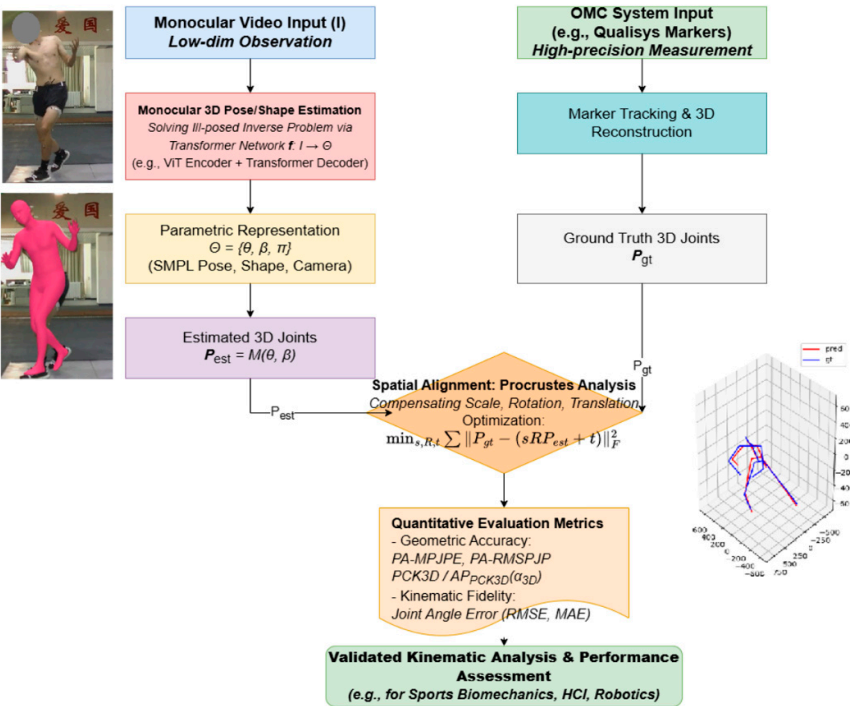


Figure 1. The overall methodological workflow, illustrating the parallel data capture from a single monocular camera (Capture4D) and a gold-standard OMC system, followed by spatial alignment and comparative analysis.

3. Results

3.1. Overall System Accuracy

The overall accuracy of the Capture4D system was first evaluated against the gold-standard OMC system across all captured tennis serve trials. As shown in Table 1, the Normalized Mean Per Joint Position Error (NMPJPE) ranged from 69.5 mm to 88.3 mm. These values fall well within the acceptable error margins (typically 70-130 mm) for applied biomechanical analysis, confirming the system’s foundational validity.

Table 1. Comparison of overall 3D joint coordinate error (NMPJPE) between the Capture4D and OMC systems for each participant.

Participant ID	Avg. NMPJPE (mm)	Keyframe	Keyframe NMPJPE (mm)	L-Elbow Angle (OMC)	L-Elbow Angle (Capture4D)	R-Elbow Angle (OMC)	R-Elbow Angle (Capture4D)
1	76.51	219	59.34	144.32	115.65	83.23	96.58
2	70.85	199	76.13	140.36	131.95	40.48	80.15
3	72.98	252	82.05	150.29	133.19	65.51	104.15
4	70.94	160	64.73	122.46	118.38	56.44	83.86
5	78.93	58	49.67	154.87	148.34	148.84	153.85
6	79.71	175	59.78	98.55	104.42	62.41	90.14
7	69.53	153	92.74	145.67	132.66	67.89	98.22
8	72.53	188	86.29	152.30	137.42	55.47	94.43
9	88.27	180	85.67	159.21	139.26	68.04	95.75

3.2. Kinematic Trajectory and Phase Analysis

To assess the system’s ability to capture the dynamic characteristics of the tennis serve, we compared the joint angle trajectories over time. Figure 2 and Figure 3 illustrate the comparison for the right and left elbow joint from representative trials. The plots demonstrate a high degree of temporal agreement between Capture4D (predicted) and the OMC system (gold standard), with both systems capturing the key kinematic events, peaks, and troughs with minimal phase lag. This indicates that Capture4D can reliably track the temporal dynamics of complex, high-speed movements.

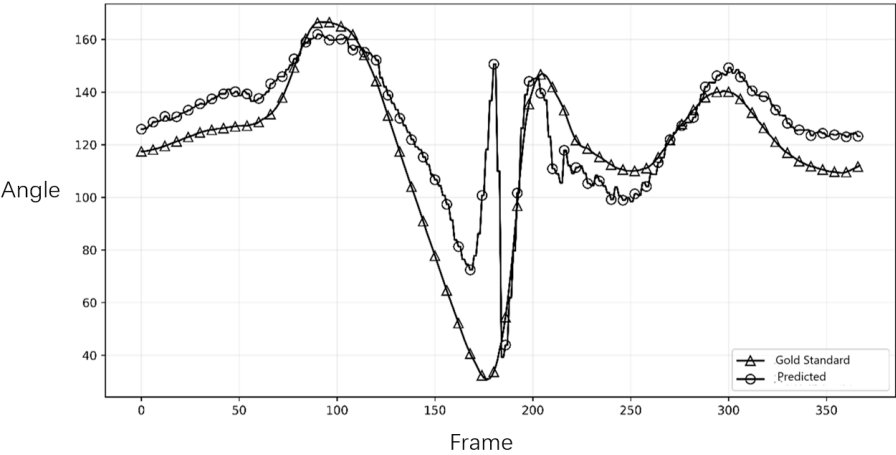


Figure 2. Comparison of the Right Elbow Joint Angle Trajectory.

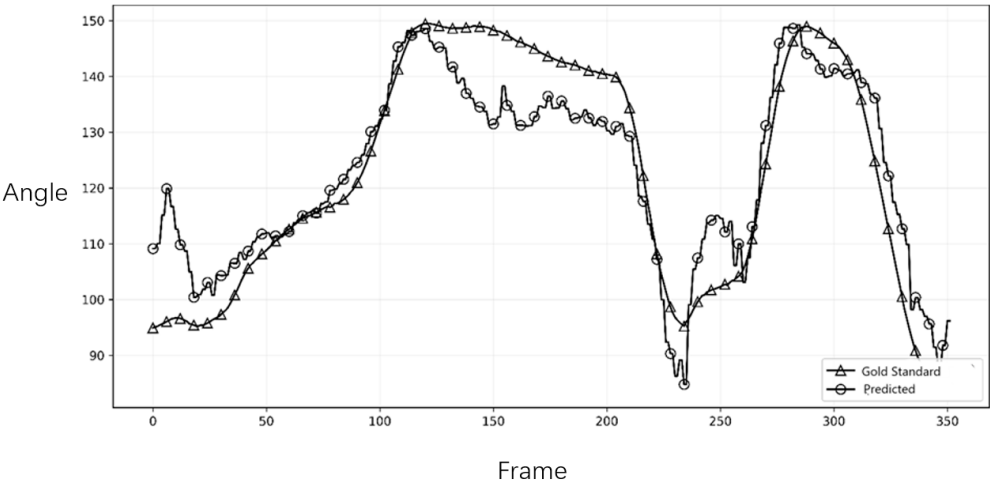


Figure 3. Comparison of the Left Elbow Joint Angle Trajectory.

To further evaluate the system’s practical utility for coaching, we developed a comprehensive analysis framework breaking down twelve common tennis strokes (including forehand, backhand, volley, and serve, as detailed in our supplementary analysis file) into biomechanically significant phases. The serve, as an example, was segmented into five key phases, as defined in Table 2, to allow for a targeted analysis relevant to coaching.

Table 2. Definition of the key biomechanical phases of the tennis serve used for analysis.

Key Phase	Phase Name	Biomechanical Description
K0	Preparation	Athlete establishes a stable, balanced stance before initiating the motion.
K1	Backswing start	Backswing start marks the initiation of the tennis serve’s backswing phase. The athlete begins to move the racket from the stable.
K2	Peak of Backswing	The racket reaches its highest point; the shoulder achieves maximum external rotation to store elastic energy (i.e., the “Trophy Pose”).
K3	Contact Point	The moment of maximum racket velocity where the racket makes contact with the ball at the peak of the swing.
K4	Follow-through	The deceleration phase post-contact, where the arm continues across the body to safely dissipate force.
K5	Recovery	The final phase where the athlete returns to a balanced, ready position.

3.3. Error Distribution Analysis and Biomechanical Interpretation

While the overall trajectory alignment in Figures 2 and 3 demonstrates high temporal consistency, a detailed phase-specific inspection reveals notable deviations in absolute joint angles during specific high-velocity phases. Specifically, the discrepancy observed in the right elbow flexion (Figure 2) peaks during the transition from the "Trophy Pose" to the "Racket Drop" (late K1 to K2). This phenomenon can be attributed to the inherent depth ambiguity of monocular 3D reconstruction. In this phase, the athlete’s arm moves rapidly behind the torso, leading to significant self-occlusion where the limb is partially invisible to the single camera view.

Although the Transformer-based architecture of Capture4D—similar to the mechanisms described in advanced human mesh recovery systems (e.g., HMR 2.0)—leverages global attention to infer occluded body parts from context, the lack of direct depth information inevitably introduces uncertainty. When the forearm aligns parallel to the camera’s optical axis (foreshortening), the system

must rely on learned priors rather than geometric triangulation, resulting in the observed variance in the predicted angle magnitude.

However, from a coaching and pedagogical perspective, these absolute positional errors are secondary to kinematic fidelity. Crucially, the system accurately captures the timing of the peak flexion and the subsequent extension velocity. As shown in the plots, the phase shifts (temporal lag) between the predicted and gold-standard curves are negligible. This confirms that Capture4D effectively preserves the kinetic chain sequencing—such as the critical synchronization between hip rotation and arm extension—which is the primary focus in technical diagnosis. Therefore, despite the presence of systematic estimation errors due to occlusion, the system remains a valid tool for identifying technical flaws and monitoring movement patterns in real-world training scenarios

3.4. Visual and Positional Fidelity

Finally, to provide a qualitative assessment of the system’s reconstruction fidelity, Figure 4 presents a visual comparison of the 3D poses at critical keyframes of the serve: K0 (Preparation), K1 (Backswing Start), K2 (Peak Backswing), K3 (Contact), K4 (Follow-through peak), and K5 (Recovery). The visual alignment confirms that Capture4D accurately captures the overall body posture and limb configuration throughout the movement, providing an intuitive and correct representation of the athlete’s technique.

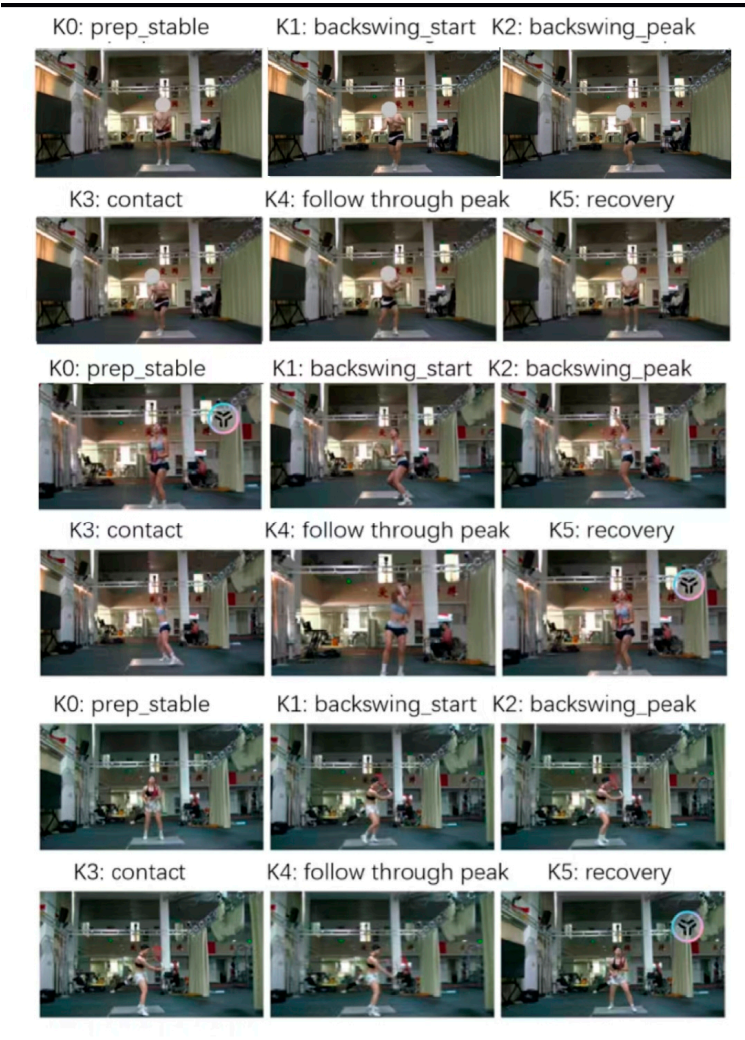


Figure 4. Visual comparison of key pose estimations during the representative tennis skill motions.

4. Discussion

This study shows the acceptable validity of the Capture4D system, a Transformer-based monocular motion capture tool, for the biomechanical analysis using tennis serve in the data collection protocol. Our results indicate that while the system's absolute positional accuracy (NMPJPE) is slightly lower than the gold-standard OMC, its ability to accurately track kinematic trajectories and quantify key kinematic events makes it a powerful and practical tool for sports science applications.

Notably, the system's single-camera setup, portability, and cost-effectiveness in real training and competition environment, would extend its use beyond tennis. For example, in figure skating, it can track joint kinematics (e.g., hip abduction in axel jumps) during dynamic movements without limiting training space—solving OMC's lab-only limitation. In other racket sports (badminton, table tennis), it captures high-speed fine movements: wrist snap in badminton smashes or forearm rotation in table tennis top spins, aiding technical optimization.

By validating Capture4D in tennis, this study paves the way for its cross-sport application, expanding the audience to researchers in figure skating, racket sports, and beyond while boosting its citation value as a versatile biomechanical analysis solution.

4.1. Implications for Sports Science and Coaching/Teaching: Bridging the Lab-to-Field Gap

The most significant contribution of this research lies in demonstrating a viable solution to fill the “lab-to-field” gap in sports analysis in coaching and teaching. For decades, coaches and athletes have faced a trade-off: either conduct highly accurate analysis in an artificial, expensive laboratory environment or rely on subjective observation in the ecologically valid but data-poor training field. Capture4D offers a compelling bridge between these two extremes.

It is crucial to interpret the observed error margins within the context of practical coaching. While an average NMPJPE of 75.66 mm represents a positional discrepancy, this level of error is often smaller than the natural variability in an athlete's movement between repetitions. More importantly, for a coach's decision-making process, the preservation of kinematic patterns and temporal sequences is paramount. For instance, identifying whether an athlete's shoulder rotation peak occurs before or after their hip rotation peak is a critical insight that is unaffected by minor absolute positional errors. The feedback for the coaches shows that particularly the high fidelity of the joint angle trajectories (Figures 2 and 3), confirm that Capture4D reliably preserves these crucial temporal and relational patterns, making it a robust tool for practical, in-field technical evaluation.

Our findings show that the system provides data with sufficient accuracy to inform coaching decisions. For a coach, the precise absolute 3D position of an athlete's elbow is often less critical than the pattern and timing of its movement. The high temporal agreement in joint angle trajectories (as seen in Figures 2 and 3) confirms that Capture4D can reliably be used to assess an athlete's kinetic chain sequencing.

Furthermore, the ability to perform a phase-based analysis and extract discrete, coach-centric metrics (Table 2) is a key advantage. A coach or teacher can use this system to answer critical performance questions directly on the court. For instance, by examining the ‘peak shoulder-hip separation angle’ at K1 and the subsequent ‘peak trunk rotation velocity’ leading into K2, a coach can quantitatively assess the efficiency of an athlete's core in transferring power from the lower body to the upper limbs. A deficiency in this sequence could lead to specific corrective drills focusing on core stability and rotational power. Similarly, by monitoring the ‘peak shoulder internal rotation velocity’ during the K3 follow-through phase over multiple sessions, a coach could track potential indicators of fatigue or overload on the shoulder joint, which are critical for injury prevention strategies. This transforms quantitative analysis from a retrospective research activity into an immediate, actionable feedback loop that can be integrated into daily training routines. The comprehensive framework we developed for twelve distinct tennis movements, exemplified here by the serve, also underscores the system's broad potential for holistic, long-term athlete monitoring.

4.2. Contextualizing the Contribution: Accessibility and Cost-Effectiveness

While Optical Motion Capture (OMC) systems remain the undisputed “gold standard” for accuracy in biomechanical research (Bartlett 2007), their practical application in day-to-day coaching/teaching and training very limited. The literature widely acknowledges that the significant financial investment, requirement for dedicated laboratory space, and need for highly trained technical personnel are substantial barriers to the widespread adoption of OMC technology (Hughes, Arapakis, and McNeill 2019). This creates a well-documented “lab-to-field” gap, where the most accurate analysis tools are unavailable in the very environments where athletes actually train and compete.

The primary contribution of this study, therefore, is not to challenge the sub-millimeter precision of OMC systems, but to validate a practical and accessible alternative that directly addresses these long-standing limitations. The goal of democratizing sports science by providing low-cost, easy-to-use tools for quantitative analysis is a significant and active area of research (Camomilla et al. 2018). Our findings demonstrate that Capture4D, with its reliance on only a single standard camera and minimal setup time, represents a significant step toward this goal. By providing data of sufficient quality for kinematic pattern analysis and coaching feedback, it offers a solution that is orders of magnitude more cost-effective and operationally feasible for a vast majority of sports programs, from grassroots to professional levels.

4.3. Limitations and Future Directions

Despite the promising results, it is important to acknowledge the system’s limitations. The primary source of error is the inherent depth ambiguity of monocular vision. This is most evident when limbs move parallel to the camera’s image plane or during significant self-occlusion. While the Transformer architecture mitigates this by leveraging global context, high-speed movements can still introduce motion blur, challenging frame-by-frame estimation.

Future research should focus on several key areas. First, further optimization of the Transformer model to specifically improve robustness to high-speed motion and severe occlusion would be beneficial. Second, exploring multi-sensor fusion, for instance by integrating data from a single Inertial Measurement Unit (IMU) placed on the athlete, could dramatically improve accuracy with minimal impact on convenience. Finally, developing and validating sport-specific models, fine-tuned on extensive datasets of elite athletes in sports like tennis, golf, or baseball, could further enhance performance and provide even more granular insights for coaches.

5. Conclusion

This study rigorously validated the Capture4D system as an effective and practical tool for the 3D kinematic analysis in tennis. The system provides accuracy within the acceptable range for applied biomechanical research and sport analysis while offering significant advantages in cost, portability, and ease of use over traditional OMC systems. By enabling coaches, teachers and sports scientists to move quantitative analysis out of the laboratory and onto the court, Capture4D represents a significant step towards the democratization of sports motion analysis, empowering a data-informed approach to technique optimization and injury prevention.

Author Contributions: Conceptualization, study design and methodology were performed by Yue Zhao, Yuanlong Liu and Zan Gao. Data collection and material preparation were performed by Yixiong Cui and Haijun Wu. Software development, data processing and analysis were performed by Shuo Wang. The first draft of the manuscript was written by Yue Zhao. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript. Supervision and funding acquisition were performed by Yuanlong Liu.

Funding: This work was supported by the Fundamental Research Funds for the Central Universities (Grant number: 2023RHJS002) for the project on ‘Development of a Tennis Visualization System Based on Motion Capture Technology’ and the China Scholarship Council (Fund No. 202406520033).

Institutional Review Board Statement: This study protocol was approved by the Beijing Sport University Institutional Review Board (BSU IRB) with CIRBID 2025426H.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Andriluka, Mykhaylo, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2014. “2d human pose estimation: New benchmark and state of the art analysis.” In Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, 3686–3693.
- Bartlett, Roger. 2007. Introduction to sports biomechanics: Analysing human movement. Routledge.
- Camomilla, Valentina, Elena Bergamini, Silvia Fantozzi, and Giuseppe Vannozzi. 2018. “An overview of wearable inertial sensors for human movement analysis.” *Sensors* 18 (3): 768.
- Cao, Zhe, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. “Realtime multi-person 2d pose estimation using part affinity fields.” In Proceedings of the IEEE conference on computer vision and pattern recognition, 7291–7299.
- Chen, Haobo, Chun-Chao Wang, Sen Wang, Wendong Zeng, and Yin Qian. 2021. “Cross-view fusion for 3d human pose estimation.” arXiv preprint arXiv:2103.09813.
- Colyer, S. L., M. Evans, D. P. Cosker, and A. I. Salo. 2018. “The validity of a deep learning-based 2D-to-3D lifting method for the analysis of tennis serves.” *Journal of sports sciences* 36 (19): 2248–2254.
- Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and others. 2020. “An image is worth 16x16 words: Transformers for image recognition at scale.” arXiv preprint arXiv:2010.11929.
- Goel, Shubham, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. 2023. “Humans in 4D: Reconstructing and Tracking Humans with Transformers.” In Proceedings of the IEEE/CVF International Conference on Computer Vision, 9136–9147.
- Gu, Chunhui, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, and others. 2018. “Ava: A video dataset of spatio-temporally localized atomic visual actions.” In Proceedings of the IEEE conference on computer vision and pattern recognition, 6047–6056.
- Hughes, Kevin, Ioannis Arapakis, and Alistair McNeill. 2019. “The use of technology in sport: A review of the literature.” *International Journal of Sports Science & Coaching* 14 (4-5): 593–613.
- Ionescu, Catalin, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2013. “Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments.” *IEEE transactions on pattern analysis and machine intelligence* 36 (7): 1325–1339.
- Isakov, Karim, Egor Burkov, Victor Lempitsky, and Yury Malkov. 2019. “Learnable triangulation of human pose.” arXiv preprint arXiv:1905.05754.
- Kanazawa, Angjoo, Jason Y Zhang, Panna Felsen, and Jitendra Malik. 2019. “Learning 3d human dynamics from video.” In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 5614–5623.
- Li, Shichao, Lei Ke, Kevin Li, Yudong Zhang, Jiakai Cui, and Kyoung Mu Lee. 2020. “Cascaded deep monocular 3d human pose estimation with evolutionary training data.” arXiv preprint arXiv:2008.02433.
- Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. “Microsoft coco: Common objects in context.” In Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13, 740–755. Springer.
- Loper, Matthew, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. “SMPL: A skinned multi-person linear model.” *ACM transactions on graphics (TOG)* 34 (6): 1–16.

- Martinez, Julieta, Rayat Hossain, Javier Romero, and James J Little. 2017. "A simple yet effective baseline for 3d human pose estimation." arXiv preprint arXiv:1705.03098.
- Mehta, Dushyant, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. 2017. "Monocular 3d human pose estimation in the wild using improved cnn supervision." In 2017 international conference on 3D vision (3DV), 506–516. IEEE.
- Moon, Gyeongsik, Ju Yong Chang, and Kyoung Mu Lee. 2019. "Camera distance-aware 3d human pose estimation from a single image." arXiv preprint arXiv:1907.11375.
- Moreno-Noguer, Francesc. 2017. "3d human pose estimation from a single image." arXiv preprint arXiv:1707.01228.
- Pang, Hui En, Zhongang Cai, Lei Yang, Tianwei Zhang, and Ziwei Liu. 2022. "Benchmarking and analyzing 3d human pose and shape estimation beyond algorithms." *Advances in Neural Information Processing Systems* 35: 26034–26051.
- Pavlo, Dario, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. "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*, 7753–7762.
- Tome, Denis, Chris Russell, and Lourdes Agapito. 2017. "Lifting from the deep: Convolutional 3d pose estimation from a single image." arXiv preprint arXiv:1701.00295.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is all you need." In *Advances in neural information processing systems*, 5998–6008.
- Wandt, Bastian, and Bodo Rosenhahn. 2021. "Canonpose: Self-supervised 3d human pose estimation in the wild." arXiv preprint arXiv:2103.11657.
- Whiteside, David, Bruce Elliott, Brendan Lay, and Machar Reid. 2016. "Biomechanics of the tennis serve: a systematic review." *Journal of sports sciences* 34 (18): 1767–1781.
- Wu, Jiahong, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipai Zhou, Guosen Lin, Yanwei Fu, and others. 2017. "Ai challenger: A large-scale dataset for going deeper in image understanding." arXiv preprint arXiv:1711.06475.
- Zhang, Jun, Mengmeng Wang, Zhiyong Chen, Zhongwei Luo, Xiaohua Wei, and Yin Zheng. 2020. "A survey on 2d and 3d human pose estimation." *Image and Vision Computing* 99: 103921.
- Zheng, Ce, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 2021. "Poseformer: A transformer-based model for 3d human pose estimation." In *Proceedings of the IEEE/CVF international conference on computer vision*, 11655–11664.
- Zhou, Xingyi, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. 2017. "Towards 3d human pose estimation in the wild: a weakly-supervised approach." arXiv preprint arXiv:1704.02440.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.