Scientific Research Publishing

Journal of Data Analysis and Information Processing, 2026, 14(1), 40-48
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

On the Utility of Pose Estimation Models for Golf Swing Understanding

Alina Yuan[1], Bryant Ndongmo[2]

[1]Cranbrook School, Bloomfield Hills, MI, USA
[2]Massachusetts Institute of Technology, Cambridge, MA, USA
Email: alinayuan08@gmail.com

How to cite this paper: Yuan, A. and Ndongmo, B. (2026) On the Utility of Pose Estimation Models for Golf Swing Understanding. *Journal of Data Analysis and Information Processing*, **14**, 40-48.
https://doi.org/10.4236/jdaip.2026.141003

Received: October 22, 2025
Accepted: December 19, 2025
Published: December 22, 2025

Abstract

Human pose estimation has shown increasing potential in sports analytics, particularly for evaluating and improving athletic motion. However, existing models are typically optimized for general-purpose or stationary scenarios and struggle when applied to high-speed, occlusion-prone sports such as golf. This study investigates the performance of two state-of-the-art pose estimation models—YOLO Pose and MediaPipe Pose—in analyzing golf swings from video data. A custom dataset was developed, consisting of golf swing recordings across diverse players, backgrounds, and lighting conditions. Each video was segmented into frames, and a subset was manually annotated to create ground-truth keypoints for evaluation. Model performance was assessed using the Object Keypoint Similarity (OKS) metric. Results show that while MediaPipe Pose achieved higher average accuracy (mean OKS = 0.636) compared to YOLO Pose (mean OKS = 0.604), YOLO demonstrated more consistent predictions with lower variance. Qualitative analysis further revealed that MediaPipe better handles partial occlusions but is more sensitive to environmental factors. These findings highlight trade-offs between model precision, consistency, and robustness in dynamic sports contexts, suggesting the need for domain-specific adaptations to improve accuracy in golf swing analysis. Such insights further underscore how pose-based motion understanding can serve as a foundation for developing intelligent feedback systems, bridging the gap between traditional coaching and automated performance analytics.

Keywords

Pose Estimation, Golf Swing Analysis, Computer Vision, YOLO Pose, MediaPipe Pose, Sports Analytics, OKS Metric, Human Motion Analysis

1. Introduction

Golf swing performance analysis has traditionally relied on expert coaching, slow motion video review, and sensor-based simulators. While these methods can provide accurate and detailed feedback, their high costs and limited accessibility may prevent many from using them. Recent advances in computer vision technologies [1] [2] have inspired new possibilities for analyzing golf swings through videos and images. Particularly, the ability to capture and extract body landmarks using pose estimation models allows for further detailed analysis of the swing without special equipment.

However, the unique nature of golf swings presents challenges for such models. The high speed and complex rotational movement cause occlusion and motion blur when trying to use these models, causing inaccurate data and inefficient results. Current pose estimation models vary in their ability to handle these constraints. However, a systematic comparison of these models as they pertain to golf swings has not yet been discussed.

This project aims to address that gap by evaluating multiple pose estimation models under the context of golf swing datasets and identifying their strengths and limitations. We propose to analyze results using evaluation metrics and to identify the underlying causes of performance weakness. Furthermore, we aim to develop a tailored pose estimation model specifically designed for golf swing motion, addressing the limitations of existing pose estimation models.

2. Background

Human pose estimation relies on a variety of computational architectures designed to localize key points of the human body from visual input. Modern pose estimation models typically employ deep learning techniques, such as those employed in Mediapipe, YOLO, and OpenPose. However, many of these models are optimized for stationary or low-motion environments and often struggle when applied to complex body configurations or high-speed movements. When used in sports contexts like golf, they face additional challenges such as occlusion (body parts hidden from the camera) and motion blur, which significantly reduce detection accuracy—especially when models are trained on general-purpose datasets rather than domain-specific ones.

For example, BlazePose [3] employs a combination of heatmap prediction and coordinate regression to estimate body landmarks. The detector depends heavily on facial visibility for initialization and alignment, but in golf, the player's face may be partially obscured by hats, shadows, or lighting reflections. Although BlazePose's use of 33 key points makes it efficient and lightweight—well-suited for simple videos—it demonstrates limited 3D depth estimation and decreased performance when key points are occluded.

Similarly, YOLO, developed by Ultralytics, adopts a single-stage architecture optimized for efficiency and real-time inference. Unlike BlazePose, YOLO models are often trained with 17 key points, excluding many facial landmarks. While this

contributes to faster computation, YOLO also suffers from similar 3D estimation limitations and struggles under occlusion or rapid motion conditions.

Mediapipe, also developed by Google, tracks 33 key points with corresponding [x, y, z, visibility] coordinates, providing spatial depth and confidence for each landmark. Compared to the other two models, Mediaipipe supports 3D pose estimation natively, making it more suitable for sports analysis. However, its performance is highly sensitive to environmental conditions—particularly lighting variations and self-occlusions that commonly occur in dynamic golf swings.

Recent research demonstrates that the limitations of general-purpose pose estimation models in sports contexts arise not only from architectural constraints but also from dataset bias, motion dynamics, and the absence of temporal reasoning. The introduction of the SportsPose dataset by Ingwersen *et al.* [4] revealed that models trained on static, non-sport activities tend to misrepresent high-velocity motion, producing large localization errors when joints rotate out of the sagittal plane—a common occurrence in golf swings. In a domain-specific application, Ju *et al.* [5] showed that conventional 2D and 3D pose estimation systems calibrated on everyday postures struggled with golf-specific phenomena such as club-shaft occlusion, high angular velocity, and rapid camera-frame transitions, resulting in reduced biomechanical validity. From a motion-science perspective, precise localization of hip-shoulder separation, lead-arm flexion, and wrist hinge is critical for quantifying swing efficiency and power transfer; even small deviations at these joints can propagate into large analytical errors. Suo [6] emphasized that markerless vision-based motion capture still trails laboratory systems due to environmental sensitivity, motion blur, and partial occlusion, all of which are common in outdoor sports settings. Furthermore, real-time single-stage detectors such as YOLOv8 have improved inference speed but still degrade in rotational tasks where limbs overlap, as demonstrated by Dong *et al.* [7]. Tharatipyakul *et al.* [8] likewise concluded that deep learning models for athletic motion require not only spatial accuracy but temporal continuity to maintain biomechanical plausibility across frames. Collectively, these findings indicate that achieving meaningful sports analytics demands domain-specific fine-tuning, improved temporal modeling, and the inclusion of biomechanical priors—criteria that current off-the-shelf pose models meet only partially.

Overall, while existing models demonstrate impressive real-time performance and accessibility, their general-purpose training and environmental sensitivity limit their effectiveness for sports like golf. Collectively, these constraints highlight an urgent need for domain-specific pose estimation models trained on golf-oriented datasets that capture the biomechanics of rotation, balance, and post-impact stabilization. Establishing such specialized resources will enable more robust evaluation protocols and foster reproducibility across future research efforts.

3. Experimental Setup and Dataset

To evaluate the accuracy of pose estimation in golf swings, we constructed a ded-

icated dataset specifically for golf-swing analysis. This dataset captures a wide range of swing conditions to enable comprehensive comparison and to reflect the variability that pose estimation models may encounter in real-world settings and to provide a scalable framework for future fine-tuning and model retraining using newly collected motion data. The videos include recordings of PGA Tour professionals, amateur players, and beginners, featuring diverse swing types and body structures to ensure model generalization and fairness in evaluation. Recordings were collected under various backgrounds, camera angles, and lighting conditions to test model robustness across environments. This variability was intentional, simulating the diverse conditions—such as changing sunlight, camera quality, and frame rate—that characterize on-course golf footage. All swings in the dataset were performed by right-handed players, as there were insufficient samples of left-handed swings to support meaningful evaluation.

Each video was processed using a custom-built frame extraction program, which generated hundreds of still frames in image format for annotation. A subset of frames from each video was manually labeled to create ground-truth keypoint annotations, serving as benchmarks for assessing model accuracy. Annotated key points include the shoulders, elbows, wrist, hip, knee, and ankles—the joints most critical to evaluating golf swing mechanics and impact posture.

By incorporating a diverse range of players, environments, and motion types, the dataset ensures that performance metrics reflect a model's ability to generalize, rather than its strength in narrow conditions. This dataset not only supports the evaluation of existing pose estimation models but also provides a foundation for developing specialized models better suited to golf motion analysis.

Building on prior research, this study compares the performance of two leading pose estimation frameworks, YOLO Pose and MediaPipe Pose, in analyzing golf swing motion. The comparison focuses on each model's accuracy, efficiency, and robustness when tracking human movement specific to the golf swing. While both models employ deep learning architectures, they differ in design: MediaPipe Pose uses a two-stage pipeline with 33 landmarks, whereas YOLO Pose follows a single-stage architecture with 17 landmarks. Evaluating them side-by-side provides insight into trade-offs between speed and precision, as well as their adaptability to dynamic, high-speed movements.

To quantitatively assess model performance, we use the Object Keypoint Similarity (OKS) metric, introduced in [9]. Before understanding OKS, it is important to review the Intersection over Union (IoU) metric [10], also known as the Jaccard Index. IoU measures the overlap region between predicted and ground-truth bounding boxes, defined as the ratio of the intersection area to the union area, and ranges between 0 and 1. Similarly, OKS measures the similarity between predicted and ground-truth keypoints.

Keypoint similarity is calculated according to Equation (1),

$$KS = \exp\left(-\frac{d^2}{2s^2k^2}\right) \tag{1}$$

where d is the Euclidean distance between the ground truth and the predicted keypoint, k is the constant for the keypoint, and s^2 is the object's segmented area. KS values lie in the range [0, 1]. The OKS for an instance (*i.e.*, a single video frame) is then calculated as the weighted average of keypoint similarities. OKS metric is according to Equation (2),

$$OKS = \frac{\sum_{i=1}^{N} KS \cdot \delta(v_i > 0)}{\sum_{i=1}^{N} \delta(v_i > 0)} \tag{2}$$

where KS_i is the keypoint similarity for the i-th keypoint, and v_i is the ground truth visibility flag for the keypoint.

- When $v = 0$, the keypoint is not labeled.
- When $v = 1$, the keypoint is labeled but not visible.
- When $v = 2$, the keypoint is labeled and visible.

$\delta(v > 0)$ is the Dirac-delta function that yields 1 if the keypoint is labeled and 0 otherwise.

Each video was processed into frames and passed through both models for keypoint prediction. A subset of frames was manually annotated to serve as ground truth. Using the frames labeled with ground truth and the same frames with the predicted landmark, we will then be able to calculate an OKS score.

The OKS score is calculated with 12 essential keypoints to the golf swing, excluding facial keypoints and detailed keypoints, to ensure direct comparability between YOLO Pose and MediaPipe Pose and to highlight performance on joints most relevant to golf swing dynamics.

4. Results

We evaluated the OKS score of both models by comparing their predicted keypoints against the ground truth for 100 frames. OKS scores range from [0, 1], with a score of 1 indicating a perfect match, meaning that all predicted points match the ground truth. The distribution of OKS scores for both models was visualized using a box plot (Figure 1) for comparison.

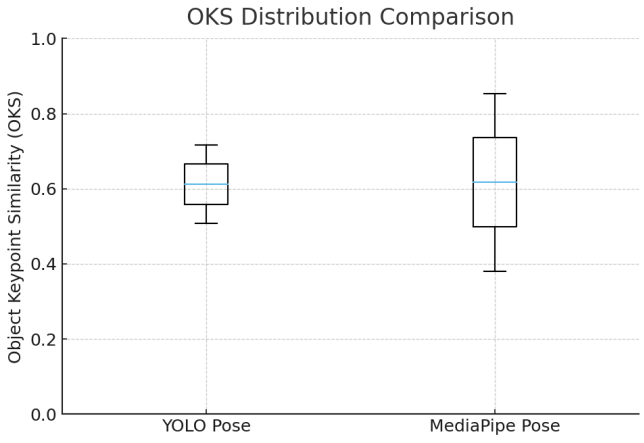

Figure 1. OKS Distribution Comparison.

The median OKS score for YOLO is 0.611, while the median OKS score for Mediapipe is 0.617. The mean OKS score for YOLO is 0.604, and 0.636 for Mediapipe. The interquartile range and range of YOLO were 0.108 and 0.209 respectively, while for MediaPipe, they were 0.238 and 0.474. YOLO demonstrated more consistent performance with lower variance, while. Mediapipe achieved a higher average accuracy but displayed greater variability across samples.

Qualitative inspections suggest that Mediapipe performs better during extended golf swings, particularly when multiple joint points are occluded. This indicates that MediaPipe is better able to recover or infer missing landmarks, although its performance tends to fluctuate under challenging lighting conditions and complex backgrounds. In contrast, YOLO Pose provides more stable predictions with lower frame-to-frame variance, making it more reliable in consistent environments, albeit at the cost of slightly reduced accuracy in complex scenes. This observation reinforces the hypothesis that model consistency is as critical as raw accuracy when translating pose estimation outputs into biomechanical insights applicable to athletic training.

Beyond numerical metrics, interpreting these results requires understanding how pose estimation error propagates into motion analysis and athletic feedback. Prior research emphasizes that frame-level accuracy alone is insufficient to assess the reliability of pose systems in dynamic sports contexts. Ingwersen *et al.* [4] observed that even models with similar mean keypoint errors can diverge dramatically when evaluated for temporal smoothness, a factor critical for modeling continuous movements like the golf swing. Ju *et al.* [5] likewise found that unstable landmark trajectories caused motion noise in derived biomechanical parameters such as hip-rotation velocity and swing-plane angle, which can mislead performance assessments. When examined through this lens, YOLO Pose's lower variance across frames may indicate superior temporal stability, whereas MediaPipe's higher OKS but greater fluctuation suggests susceptibility to environmental and motion-induced noise. Suo [6] further noted that the absence of built-in temporal filtering in most single-frame detectors limits their ability to enforce cross-frame consistency, underscoring why minor per-frame inaccuracies can accumulate into large analytical deviations.

The trade-off between precision and stability observed here also aligns with Dong *et al.* [7], who reported that single-stage YOLO-based estimators, while optimized for real-time use, exhibit reduced robustness in high-velocity limb motion, leading to intermittent joint swaps or phantom detections. Tharatipyakul *et al.* [8] extended this analysis by demonstrating that introducing recurrent refinement or temporal-attention modules significantly improves spatial coherence and mitigates jitter during fast sports actions. These findings suggest that golf-specific pose estimation should emphasize temporal integration as much as spatial accuracy, possibly through hybrid pipelines combining single-frame detection with frame-sequence smoothing. In practical terms, such integration could allow pose-based coaching applications to deliver more reliable feedback on swing tempo,

sequence timing, and follow-through mechanics. Hence, while YOLO and MediaPipe offer complementary strengths—speed versus granularity—neither achieves the temporal or biomechanical fidelity required for fully automated golf-swing evaluation without additional temporal modeling or fine-tuning on sports-specific datasets.

5. Conclusions and Future Work

This study evaluated the performance of YOLO Pose and MediaPipe Pose for analyzing golf swing motion using a custom dataset featuring diverse players, lighting conditions, and backgrounds. Both models demonstrated strong potential for sports motion analysis, with MediaPipe Pose achieving slightly higher mean accuracy (OKS = 0.636) and better handling of occluded joints, while YOLO Pose showed more consistent predictions with lower variance. These results highlight a trade-off between accuracy and stability, as MediaPipe offers greater precision in complex poses, whereas YOLO provides steadier performance in uniform settings. The findings also provide empirical evidence that model selection should depend on context—whether the goal is precise positional accuracy or reliable temporal stability.

Although both models performed reasonably well, they remain limited in addressing challenges unique to golf, such as high-speed motion, self-occlusion, and varying environmental conditions. Moving forward, future work will focus on improving model robustness through domain-specific training and incorporating additional samples, including left-handed players and varied recording environments. The development of a specialized golf pose estimation model could further enhance accuracy and reliability, enabling more practical applications in swing analysis and sports performance evaluation.

While the comparative evaluation between YOLO Pose and MediaPipe Pose offers a valuable benchmark for golf-specific motion analysis, the broader implications extend to the general design of pose estimation frameworks for dynamic sports. Recent research has emphasized that improving such models requires not only algorithmic refinement but also domain adaptation and multi-modal integration. Ingwersen *et al.* [5] argue that training with 3D sports-specific datasets substantially enhances generalization under motion blur and extreme joint rotation—conditions identical to those found in golf swings. Building upon that, Ju *et al.* [5] propose combining pose keypoints with embedding-based body-part representations to better capture rotational continuity and player individuality, which could enable more personalized swing-analysis feedback. Similarly, Suo [6] and Dong *et al.* [7] suggest that hybrid architectures integrating RGB video, inertial measurements, and depth data outperform vision-only pipelines by reducing occlusion ambiguity and enabling reliable 3D reconstruction even in outdoor lighting. For instance, inertial-vision fusion has been shown to reduce joint-angle estimation error in fast-moving limb segments [6].

Another emerging direction is the use of biomechanical priors and constraint-

based loss functions during training. Tharatipyakul *et al.* [8] demonstrate that embedding anatomical joint-limit constraints prevents implausible postures while maintaining inference speed. This principle could be directly applied to golf by enforcing realistic ranges for elbow flexion, hip rotation, and wrist deviation—variables closely tied to swing efficiency. Integrating these biomechanical priors would transform pose estimation from a purely geometric task into one that respects physiological plausibility. Furthermore, combining temporal-attention modules with constraint-based learning could stabilize predictions across entire swing sequences, supporting consistent downstream kinematic calculations such as angular momentum or torque distribution.

Taken together, these findings reinforce the view that progress in golf-swing pose estimation depends on interdisciplinary convergence between computer vision, sports biomechanics, and sensor fusion. Future systems should therefore move beyond accuracy benchmarks toward holistic evaluation frameworks that consider temporal coherence, biomechanical validity, and interpretability for end-users such as coaches or athletes. Incorporating these principles could ultimately bridge the gap between automated computer-vision analytics and the nuanced, expert-level feedback traditionally provided by human coaches.

In addition to advancing computer vision for sports, this research underscores the importance of interdisciplinary collaboration between data science, biomechanics, and sports coaching. By linking pose estimation metrics with measurable performance indicators such as swing tempo, angular velocity, and energy transfer, future studies could establish data-driven benchmarks for training efficiency. Furthermore, integrating pose-based analytics into accessible mobile platforms or wearable systems could democratize swing feedback for amateur golfers. Such developments would extend beyond golf, offering a transferable framework for sports where motion precision is critical, including tennis, baseball, and cricket. In parallel, collaborations with sports scientists and motion-analysis experts could accelerate the translation of computer-vision advances into actionable coaching tools. Thus, the insights gained here mark an essential step toward intelligent, automated, and athlete-centered performance assessment systems.

Acknowledgements

We are grateful to Prof. Greg Shakhnarovich for helpful discussions and his guidance throughout the project.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Lugaresi, C., *et al.* (2019) MediaPipe: A Framework for Building Perception Pipelines. arXiv: 1906.08172. https://arxiv.org/abs/1906.08172

[2] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2015) You Only Look Once: Unified, Real-Time Object Detection. arXiv: 1506.02640.

https://arxiv.org/abs/1506.02640

[3] Bazarevsky, V. and Grishchenko, I. (2020) On-Device, Real-Time Body Pose Tracking with MediaPipe BlazePose. Research Google. https://research.google/blog/on-device-real-time-body-pose-tracking-with-mediapipe-blazepose/

[4] Ingwersen, C.K., Møller Mikkelstrup, C., Jensen, J.N., Rieger Hannemose, M. and Dahl, A.B. (2023) Sportspose—A Dynamic 3D Sports Pose Dataset. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Vancouver, 17-24 June 2023, 5219-5228. https://doi.org/10.1109/cvprw59228.2023.00550

[5] Ju, C., Kim, J. and Lee, D. (2023) GolfMate: Enhanced Golf Swing Analysis Tool through Pose Refinement Network and Explainable Golf Swing Embedding for Self-training. *Applied Sciences*, **13**, Article 11227. https://doi.org/10.3390/app132011227

[6] Suo, X., Tang, W. and Li, Z. (2024) Motion Capture Technology in Sports Scenarios: A Survey. *Sensors*, **24**, Article 2947. https://doi.org/10.3390/s24092947

[7] Dong, C. and Du, G. (2024) An Enhanced Real-Time Human Pose Estimation Method Based on Modified YOLOv8 Framework. *Scientific Reports*, **14**, Article No. 8012. https://doi.org/10.1038/s41598-024-58146-z

[8] Tharatipyakul, A., Srikaewsiew, T. and Pongnumkul, S. (2024) Deep Learning-Based Human Body Pose Estimation in Providing Feedback for Physical Movement: A Review. *Heliyon*, **10**, e36589. https://doi.org/10.1016/j.heliyon.2024.e36589

[9] COCO—Common Objects in Context. cocodataset.org. https://cocodataset.org/#keypoints-eval

[10] Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J. and Zisserman, A. (2009) The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, **88**, 303-338. https://doi.org/10.1007/s11263-009-0275-4