Key Technologies for 3D Golf Swing Pose Recognition and Angle Analysis Using Multimodal Fusion

Sihan Yu, Yaojin Zhong, Zhiming Zhan*
School of Artificial Intelligence, Jianghan University, Wuhan 430056, China
*Corresponding email: umbrella1201@163.com

Abstract

To improve the accuracy and robustness of dynamic pose measurement in golf training, this study proposes a method that fuses vision and inertial measurement unit (IMU) data for three-dimensional (3D) pose recognition and key-angle analysis. The pipeline comprises data acquisition, preprocessing, feature fusion, and quantitative analysis. A part attention network (PAN) adaptively fuses features through dynamic weighting. Six key angle measures are defined based on golf swing biomechanics. The mean joint localization error was 3.1 mm, the key-angle measurement error was ±1.2°, and swing phase recognition accuracy reached 95.6%. Recognition accuracy remained above 92.0% under challenging conditions such as low light and occlusion. The fusion method improved accuracy and robustness compared with conventional single-modality approaches and supports quantitative assessment in golf training.

Keywords

Three-dimensional pose recognition, Multimodal fusion, Part attention network, Inertial measurement unit, Angle analysis

Introduction

Background and significance

Golf demands precise movement. Small differences in joint angles and body posture during a swing directly affect ball striking and performance, making accurate measurement of dynamic pose parameters important for evidence-based training. Conventional analysis relies on visual observation by coaches or video replay from a single source. These approaches introduce subjective error, offer limited means of measuring key angles, and make it difficult to reconstruct movement over time. Advances in intelligent sports technology have brought computer vision and inertial measurement units (IMUs) into motion analysis.

Each modality has limitations. Changes in lighting and occlusion by clothing can interrupt visual trajectories or displace detected landmarks, reducing robustness outdoors. IMUs are unaffected by these visual conditions but accumulate drift and cannot provide global spatial localization [1].

Prior research in China and abroad

(1) Single-modality pose recognition

a) Vision

Approaches based on the You Only Look Once (YOLO) family and MediaPipe enable real-time object detection and extraction of 33 human body landmarks. They are widely used in sports pose analysis. Rapid motion can produce false or missed landmark detections, while changes in lighting and occlusion substantially reduce accuracy in challenging conditions [2-4].

b) IMU

High sampling rates, such as 100 Hz, allow IMUs to capture transient motion details independently of visual conditions. They directly report kinematic quantities such as acceleration and angular velocity. Cumulative drift and the lack of global spatial localization make it necessary to use them with another modality [5].

(2) Multimodal pose recognition

Multimodal fusion addresses the limitations of individual modalities by combining sources such as vision and IMU data. Existing methods typically integrate data using filters, including Kalman and complementary filters, or simple weighting. Some studies use attention mechanisms to emphasize important features. Adaptive weighting across golf swing phases remains underdeveloped, and a mature fusion architecture for complex, high-speed motion is still lacking [6-8].

(3) Golf-specific pose recognition

Optical motion capture systems developed outside China, such as Vicon, can achieve joint-angle measurement accuracy on the order of 0.1°. They require multi-camera arrays and dedicated facilities, and a single system costs more than RMB 1 million, limiting their adoption in routine training.

Mainstream approaches in China generally use off-the-shelf pose recognition techniques such as YOLO and MediaPipe without golf-specific optimization. During swings that reach instantaneous speeds of 50 m/s, or under partial occlusion and changing lighting, these approaches can mislocalize joints and incur key-angle measurement errors exceeding ±3°. Limited golf-specific training data also reduce their ability to track the shoulders, elbows, and knees reliably, making them inadequate for quantitative training.

Rationale for fusing vision and IMU data

The characteristics of vision, IMUs, and PAN support multimodal fusion in the following respects.

(1) Complementary measurements

IMUs report local three-dimensional kinematics, while vision provides spatial coordinates for 33 body landmarks in the global pose. Together they address gaps in the information available from either modality alone.

(2) Complementary error characteristics

IMUs accumulate drift, while vision is subject to random error. Each modality can help compensate for errors in the other. Studies have reported substantially lower pose recognition errors after fusion, supporting its feasibility.

(3) Complementary performance across conditions

IMUs continue to operate despite changes in lighting or occlusion, while vision provides global localization. Their combination supports the range of conditions encountered during a golf swing.

Research objectives and contributions

(1) A golf-specific dynamic fusion strategy assigns weights according to feature confidence to address variations in single-modality accuracy across swing phases.

(2) A swing-angle analysis framework defines six key angle measures and links joint localization to golf-specific quantitative assessment across the entire swing.

(3) The system achieves a mean joint localization error of 3.1 mm, a key-angle error within ±1.5°, and a processing rate of 27.8 frames per second (FPS), meeting on-site real-time analysis requirements.

Technical background

Inertial measurement

An IMU contains components such as an accelerometer and a gyroscope. It estimates pose from kinematic measurements using physical laws and data fusion algorithms. The accelerometer measures acceleration in three dimensions, and the gyroscope measures angular velocity. These measurements can be used to estimate the orientation and direction of motion of an object. Figure 1 illustrates the internal structure of an IMU.

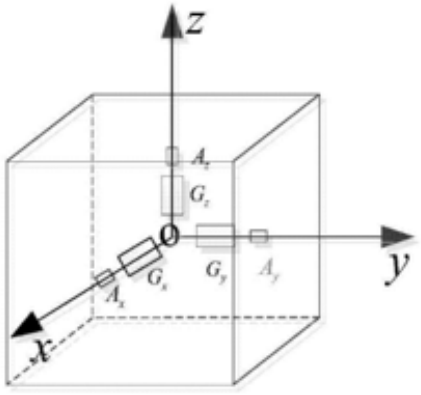


Figure 1. Illustrating the internal structure of an IMU.

IMUs continue to provide motion data when lighting or occlusion makes visual information unavailable or inaccurate. Their high sampling rates also capture rapid changes that may fall between video frames.

(1) Measurement principles

An IMU combines accelerometer and gyroscope measurements with an attitude estimation model to convert raw sensor data into an estimate of orientation.

a) Accelerometer

Under Newton's second law $F = ma$, an accelerometer detects changes in inertial force to measure three-dimensional linear acceleration $a = [a_x, a_y, a_z]^T$. The measured value includes both motion-induced and gravitational components: $a = a_{motion} + g \cdot n_g$. Here, $a_{motion} \in \Re^3$ denotes motion-induced acceleration, g=9.81 m/s^2 denotes gravitational accele-ration, and $n_g \in \Re^3$ denotes the unit vector in the direction of gravity.

b) Gyroscope

Based on conservation of angular momentum, the gyroscope measures angular velocity about three coordinate axes $\omega = [\omega_x, \omega_y, \omega_z]^T$, directly capturing limb rotation.

c) Attitude estimation

A Kalman filter fuses the data, and quaternions avoid Euler-angle gimbal lock. The filter has prediction and update stages, with the following state equations:

$$\begin{cases} \hat{x}_k^- = F\hat{x}_{k-1} + Gu_k \\ P_k^- = FP_{k-1}F^T + Q \\ K_k = P_k^- H^T (HP_k^- H^T + R)^{-1} \\ \hat{x}_k = \hat{x}_k^- + K_k(z_k - H\hat{x}_k^-) \\ P_k = (I - K_k H)P_k^- \end{cases} \quad (1)$$

Here, $\hat{x}$ is the state vector, which includes the quaternion, angular velocity, and acceleration. The matrices F and H are the state transition and observation matrices. The matrices Q and R are the process noise and measurement noise covariance matrices, respectively.

(2) Data processing workflow

Raw IMU data undergo several processing steps before fusion, as shown in Figure 2.

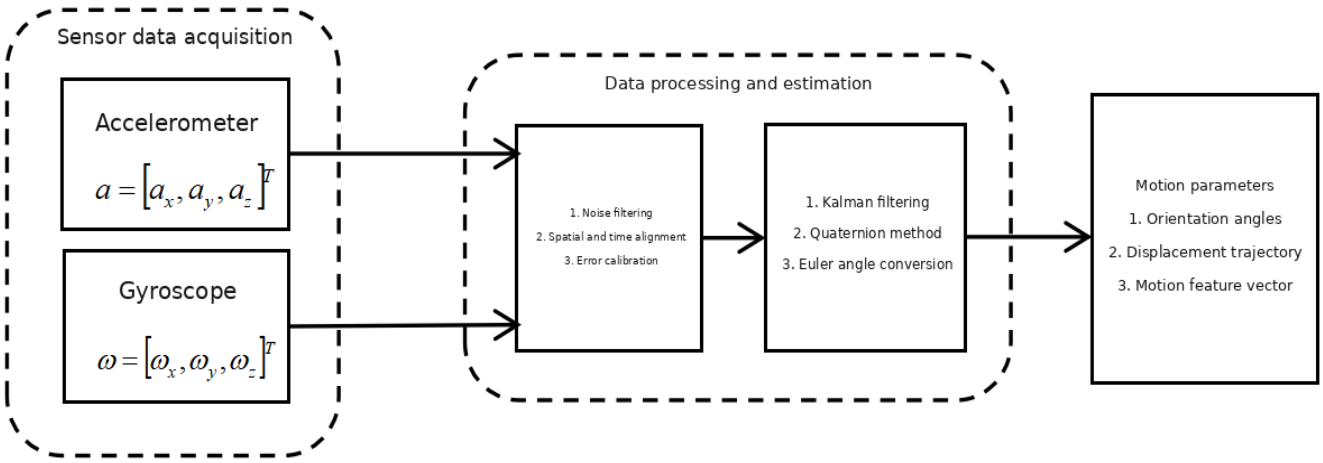


Figure 2. IMU data processing workflow.

Vision-based pose recognition

(1) Detection and landmark extraction

a) YOLOv8 object detection

The algorithm directly regresses the coordinates and confidence scores of human bounding boxes without generating region proposals. Graphics processing unit (GPU) acceleration enables real-time detection, reducing the likelihood of losing track of the golfer during rapid movement. The detected bounding box provides a precise region of interest (ROI) for landmark extraction and reduces interference from background noise.

b) MediaPipe pose recognition

This skeletal landmark detection framework is designed for dynamic human pose estimation and extracts the two-dimensional (2D) or 3D coordinates of 33 body landmarks in real time. Its end-to-end deep learning model incorporates pose priors and a feature pyramid to improve the continuity and accuracy of landmark detection. It is particularly suited to tracking the shoulders, elbows, and knees during a golf swing.

c) Combined workflow

YOLOv8 first locates the golfer and crops out the background. MediaPipe then extracts landmarks from the cropped image. This two-stage approach combines coarse localization and precise extraction to recognize high-speed golf swings in real time.

(2) Performance and suitability for golf

a) Real-time performance

YOLOv8 reaches up to 30 FPS on 4K inputs, and MediaPipe extracts landmarks at 25 FPS or higher. The combined architecture meets the processing requirement for dynamic golf swing analysis of at least 20 FPS.

b) Landmark coverage

The 33 landmarks cover key locations on the trunk, limbs, and head and support calculation of all six key golf swing angles.

c) Robustness

Adaptive image enhancement and threshold adjustment improve performance under bright light and shadows. MediaPipe's pose smoothing filter reduces landmark jitter during movement.

(3) Limitations and complementary measurements

The vision architecture is efficient and provides a baseline level of accuracy for capturing golfer posture from video frames.

Motion blurs during fast swings can shift detected landmarks, and complex outdoor lighting or occlusion by clothing can lead to missed detections and unstable pose outputs. This study integrates YOLOv8 and MediaPipe.

It adopts a PAN to fuse visual features and IMU measurements. This strategy improves recognition accuracy under golf-scene conditions. It also offsets the inherent drawbacks of pose estimation methods that rely solely on visual input.

Comparison of IMU and vision modalities

IMU sensors offer several advantages for golf pose recognition. Table 1 compares the two modalities.

Table 1. Technical comparison of IMU and vision.

Characteristic	IMU	Vision
Environmental dependence	Unaffected by lighting or occlusion	High. Requires adequate lighting and an unobstructed view.
Sampling rate	High (100 Hz), capturing brief swing details	Limited (≤60 Hz), potentially missing details of fast movement such as the downswing
Measurement output	Direct kinematic measurements (acceleration and angular velocity)	Requires multiple processing steps, including object detection and landmark regression
Error characteristics	Cumulative drift, which filtering can mitigate	Random error affected by scene complexity and image quality
Deployment cost	Low (approximately RMB 1,000 per sensor)	Moderate (approximately RMB 10,000 per industrial camera)

Part attention network

(1) PAN principles

The pyramid attention network described uses a multiscale feature pyramid and an attention mechanism to emphasize features of key body parts [9]. It dynamically assigns feature weights to highlight regions important to the task, while suppressing interfering background information that may degrade model performance. The algorithm consists of the following steps.

a) Multiscale feature extraction

A backbone such as ResNet extracts feature maps at different levels [10]. These hierarchical outputs capture low-level contours, mid-level textures and high-level semantic cues for body-related perception.

b) Attention weight generation

Feature response strength determines the part attention masks, which adaptively locate informative body regions without extra manual annotation. Feature enhancement. Attention weights are applied to feature maps to enhance relevant features, guiding the network to focus discriminative learning on target body parts.

(2) Loss function

A multi-objective loss function improves the model's ability to capture features of key body parts during the golf swing.

The cross-modal contrastive loss ($\mathscr{L}_{\text{contrast}}$) reduces the distance between features for the same class of movement. The attention sparsity loss ($\mathscr{L}_{\text{sparse}}$) encourages the model to focus on key body parts. The motion prior alignment loss ($\mathscr{L}_{\text{prior}}$) aligns attention weights with biomechanical principles.

Multimodal fusion system design for golf

Overall framework

The proposed method for dynamic 3D golf pose recognition and angle analysis comprises four modules, as shown in Figure 3, which are designed to jointly exploit visual and inertial sensing signals for reliable golf swing analysis. The acquisition module synchronously records images and IMU motion data, ensuring consistent timestamp alignment between two modalities. The preprocessing module filters noise and calibrates coordinates to mitigate sensor bias and reduce signal distortion. The feature fusion module uses attention to adaptively combine the two sources and estimate accurate 3D landmark coordinates, prioritizing high-quality cues during unstable motion. The angle analysis module uses a sports biomechanics model to calculate key golf swing angles for quantitative assessment and technical feedback. Figure 4 illustrates how visual and inertial features are aligned for 3D pose estimation.

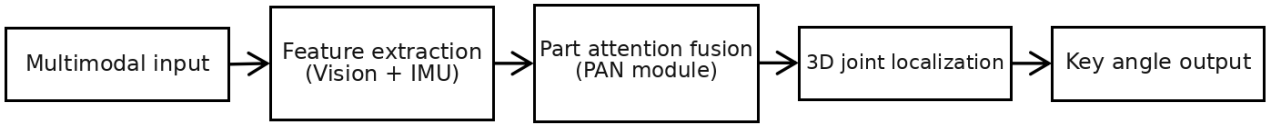


Figure 3. Overview of the proposed method.

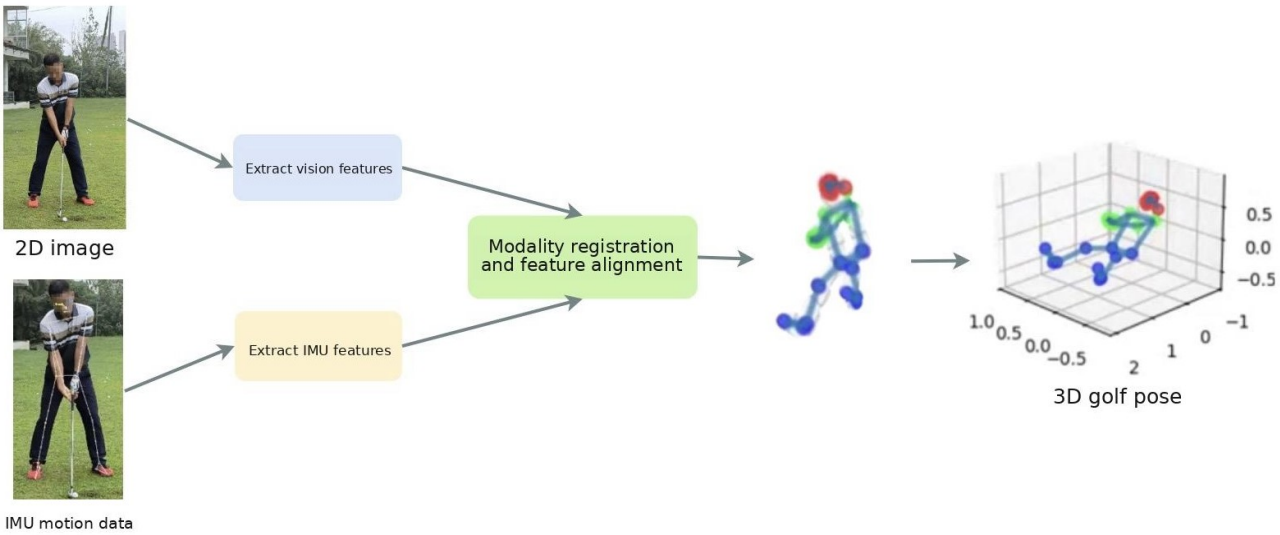


Figure 4. Golf pose estimation network.

Multimodal data preprocessing

The acquisition system comprises open source computer vision artificial intelligence (AI) Kit with Depth (OAK-D) industrial cameras and Xsens MTw Awinda IMU sensors. The cameras record at 4K resolution and 120 FPS to capture the swing.

The IMUs contain accelerometers and gyroscopes that record angular velocity and acceleration in real time at 100 Hz, independently of lighting or occlusion. Hardware synchronization aligns timestamps across the sources to ensure consistency in time and space for subsequent fusion. Figure 5 shows an example of the detected body landmarks and the skeletal overlay.

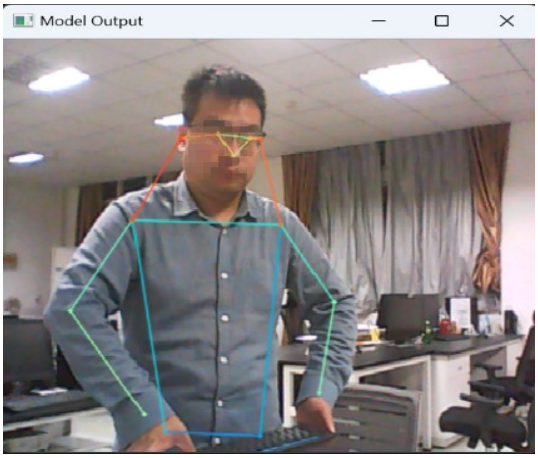


Figure 5. Example of human landmark detection and skeletal overlay.

(1) Visual data preprocessing

Contrast-limited adaptive histogram equalization (CLAHE) improves image contrast under varying lighting conditions on the golf course. OpenCV corrects image distortion using camera calibration parameters to remove perspective error [11,12].

YOLOv8 locates the golfer, allowing background regions to be cropped to reduce computation and improve landmark recognition accuracy. Figure 6 illustrates the visual inspection of person detections and landmark annotations.

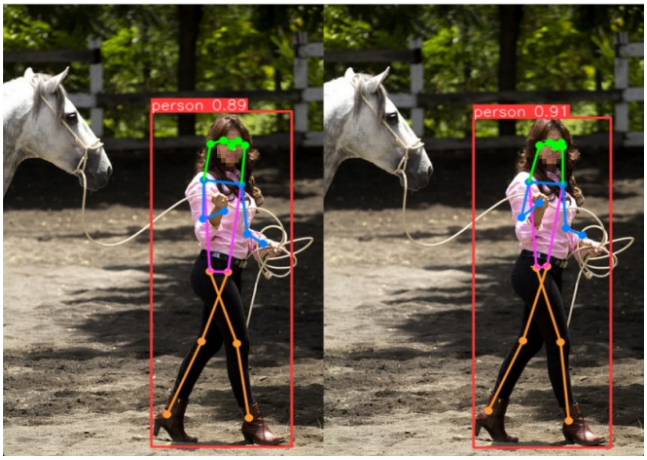


Figure 6. Example of person detection and visual inspection of landmark annotations.

The dataset was used for training, and the processed annotation files were checked against the corresponding images. Randomly selected images were also inspected with their annotation overlays.

(2) IMU data preprocessing

A complementary filter suppresses high-frequency noise in the IMU measurements while preserving the underlying motion signal. Kalman filtering synchronizes IMU and vision data and keeps the timing error within ±5 ms. A coordinate transformation maps local IMU measurements to the global coordinate system for spatial alignment with vision data.

Dynamic weighting model

The model combines a PAN with phase-adaptive

weighting to fuse vision and IMU features. The two-stage architecture considers variations in feature reliability across different swing phases. It also takes difficult sensing conditions into consideration. This design enhances the extraction and fusion of pose features tailored to golf movements.

Figure 7 illustrates the architecture of the dynamic weighting model.

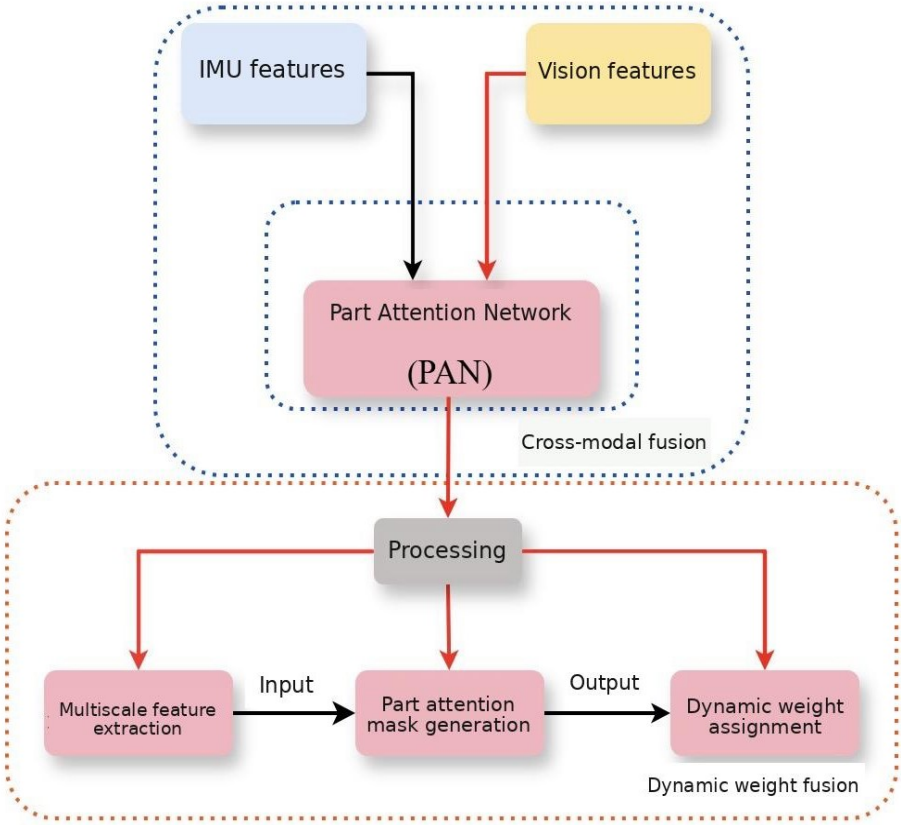


Figure 7. Dynamic weighting network.

(1) Model design

The model addresses the high speed of the swing, differences between phases, and the importance of key body parts through three design principles.

a) Feature selection

The PAN enhances features of the shoulders, elbows, and knees and suppresses noise from less relevant regions.

b) Dynamic adaptation

Fusion weights are adjusted in real time according to vision and IMU feature confidence, accounting for differences in modality performance during the backswing, downswing, and impact.

c) Error compensation

Vision provides accurate global localization, while IMUs offer a high sampling rate and operate independently of visual conditions. The two modalities compensate for each other's weaknesses and improve the robustness of the fused features. The weights assigned to vision and IMU data are adjusted dynamically across swing phases. Definitions: W_v is the vision feature weight, W_i is the IMU feature weight, S_v is the vision feature confidence, and S_i is the IMU feature confidence.

Normalized weights:

$$W_v = \frac{e^{S_v/\tau}}{\sum_{k\in\{v,i\}} e^{S_k/\tau}},\ W_i = \frac{e^{S_i/\tau}}{\sum_{k\in\{v,i\}} e^{S_k/\tau}} \quad (2)$$

Fused feature:

$$F_{fusion} = W_v \times F_v + W_i \times F_i \quad (3)$$

d) Downswing example

With S_v=0.85, S_i=0.92, and τ=0.1, the resulting weights are W_v=0.48 and W_i=0.52. The higher IMU weight compensates for motion blur in the images.

(2) Real-time performance metrics

Frame rate and average processing time per frame are used to determine whether the system meets the requirements for dynamic golf swing analysis. FPS is given by the following expression:

$$FPS = \frac{N}{T_{total}} \quad (4)$$

For example, for 1,000 frames and a total processing time of 36 s, the frame rate is FPS=27.8 and the average processing time per frame is D=1/FPS, around 35.9 ms, meeting the real-time requirement of at least 20 FPS.

Key golf swing angles

(1) Angle calculation model

The angle calculation model uses the fused 3D landmark coordinates.

a) Joint angles are calculated using the vector dot product. The coordinates of adjacent landmarks define vectors, and trigonometric functions yield the angle between them.

b) A temporal segmentation algorithm uses changes in angular velocity to identify five phases automatically: backswing, top of the swing, downswing, impact, and follow-through. The resulting segmentation supports phase-specific angle analysis.

c) The evaluation framework sets acceptable ranges for each key angle based on professional coaches' experience and sports biomechanics standards, providing a quantitative basis for improving technique.

(2) Elbow joint angle

The elbow joint angle illustrates how an angle is calculated from vectors defined by adjacent 3D landmarks. The same procedure applies to other joint angles.

a) Landmark coordinates

Upper-arm endpoint $A(x_1,\ y_1,\ z_1)$, elbow joint $B(x_2,\ y_2,\ z_2)$, and forearm endpoint $C(x_3,\ y_3,\ z_3)$.

b) Vectors

$\overrightarrow{BA} = (x_1 - x_2,\ y_1 - y_2,\ z_1 - z_2)$ and $\overrightarrow{BC} = (x_3 - x_2,\ y_3 - y_2,\ z_3 - z_2)$.

c) Angle calculation

$$\theta = arccos\left(\frac{\overrightarrow{BA}\cdot\overrightarrow{BC}}{\left|\overrightarrow{BA}\right|\times\left|\overrightarrow{BC}\right|}\right)\times\frac{180^\circ}{\pi} \quad (5)$$

d) Example

Substituting the shoulder A (685,307,150), elbow B (635,369,145), and wrist C (610,420,142) coordinates give an elbow joint angle at impact of θ=125.6°, which lies within the biomechanically optimal range of 120°~130°.

Experimental validation and results

Experimental setup and dataset

(1) Experimental setup

The hardware comprised a ThinkStation P7 workstation with a 16-core W5-3435X processor, eight 64 GB registered error-correcting code (ECC) double data rate 5 (DDR5) memory modules, and two A6000 GPUs. The setup also included three OAK-D industrial cameras, six Xsens MTw Awinda IMU sensors worn on both shoulders, elbows, and knees, and an NVIDIA Jetson autonomous machine accelerator (AGX) Orin edge computing platform. The software environment comprised Ubuntu 20.04, PyTorch 2.3.0 with compute unified device architecture (CUDA) 12.1, MediaPipe 0.9.3, and OpenCV 4.5.5.

(2) Experimental dataset

A multimodal golf swing dataset was collected from 10 professional golfers and 15 amateurs. Two professional golf coaches annotated the swing phases and manually measured key angles to provide a reference for model validation.

Evaluation metrics

Mean per joint position error (MPJPE) measures 3D joint localization accuracy [13]. Key-angle measurement accuracy is assessed using the absolute error between model estimates and manual measurements. FPS measures dynamic processing performance. Swing-phase recognition accuracy is the proportion of correctly identified instances of the backswing, downswing, and impact. Figure 8 shows an example of the reconstructed 3D body landmarks.

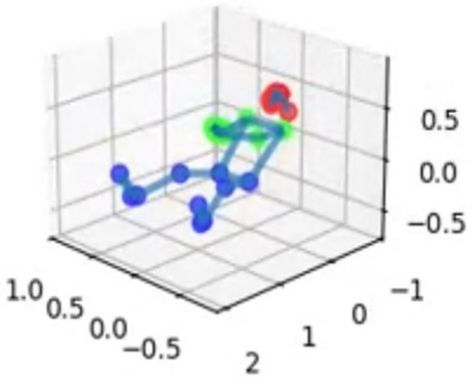


Figure 8. Example visualization of 3D body landmarks.

Results and analysis

(1) Baseline performance

The method reliably identifies 33 body landmarks, with a mean joint localization error of 3.1 mm during golf swings.

The absolute key-angle measurement error did not exceed 1.5°. The reported mean error was ±3.2° for vision alone and ±2.7° for IMU alone. The maximum processing rate was 27.8 FPS, and the minimum remained above 20 FPS, meeting the real-time requirements for swing analysis. Table 2 presents the comparison. Figure 9 compares the distributions of joint localization errors before and after training.

Table 2. Accuracy and processing speed by method.

Method	Shoulder rotation error	Elbow angle error	Knee flexion error	Mean error	FPS	Applicable conditions
Vision only (YOLOv8 + MediaPipe)	±3.2°	±2.8°	±3.5°	±3.2°	22.5	Static scenes or slow motion
IMU only	±2.8°	±2.5°	±2.9°	±2.7°	25.1	Dynamic scenes without occlusion
Proposed multimodal fusion	±1.2°	±1.0°	±1.3°	±1.2°	27.8	Complex dynamic scenes

Note: Values are averages from 50 sets of test swings performed by 10 professional golfers. The proposed method uses attention-based fusion to improve accuracy and processing speed.

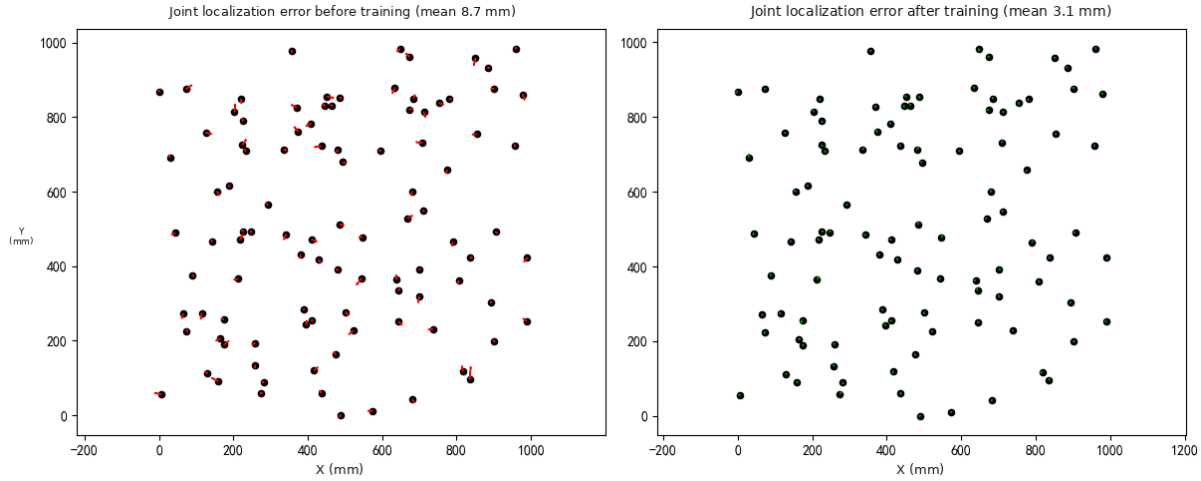


Figure 9. Joint localization error before and after training.

(2) Swing recognition and angle analysis

In tests with professional golfers, the method achieved over 95.0% pose recognition accuracy in each of three key phases: backswing, downswing, and impact. It reliably captured the distinct body configurations of each phase. Figure 10 shows the recognized swing pose and the corresponding key-angle output for quantitative analysis.



Figure 10. Golf pose recognition and key-angle analysis. (a) Swing pose recognition; (b) Key-angle analysis

Figure 11 shows how the key angles change over a complete golf swing cycle. At impact, professional golfers averaged 85.3° of shoulder rotation, 125.6° of elbow joint angle, and 142.8° of knee flexion, closely matching widely accepted recommended biomechanical values. Amateurs commonly showed insufficient shoulder rotation and excessive elbow flexion during this instant. This was consistent with the coaches' practical assessments and supported the effectiveness of the proposed method.

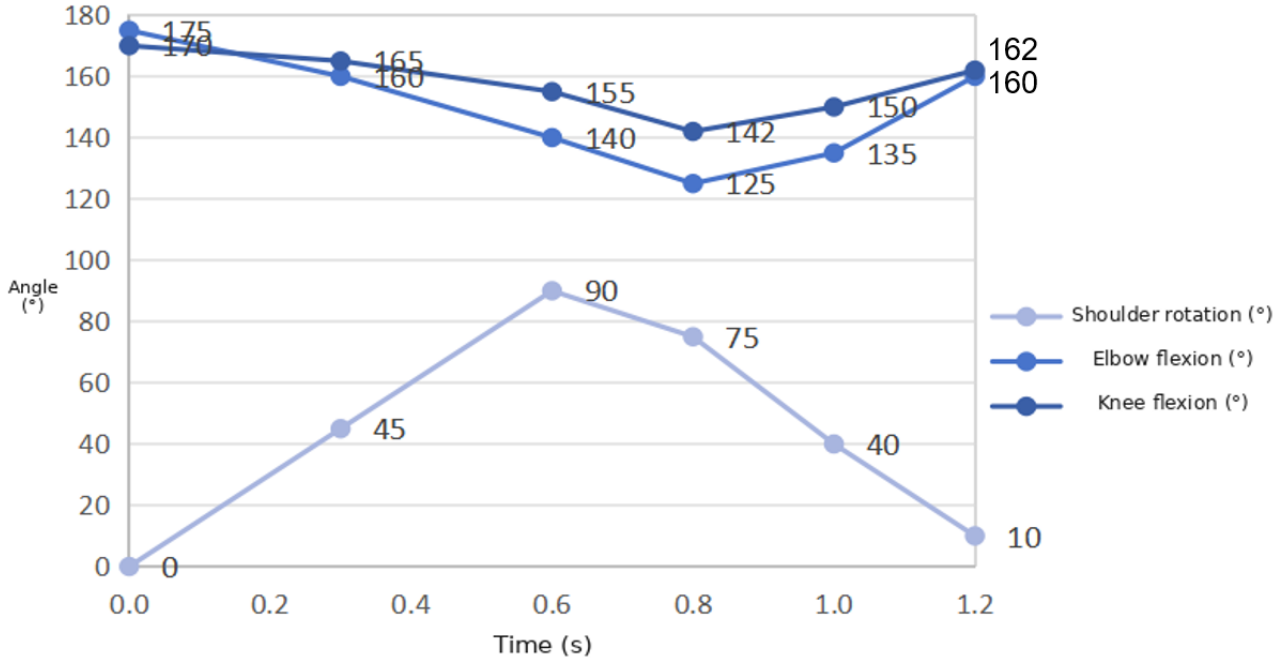


Figure 11. Key angles over the golf swing.

(3) Robustness under challenging conditions

Figure 12 compares results across venues and lighting conditions. Under overcast, low-light conditions, the proposed method's angle measurement error increased by only 0.3°. With spectator occlusion, it maintained pose recognition accuracy above 92.0%, while single-modality methods fell below 75.0%, demonstrating robustness under challenging conditions.

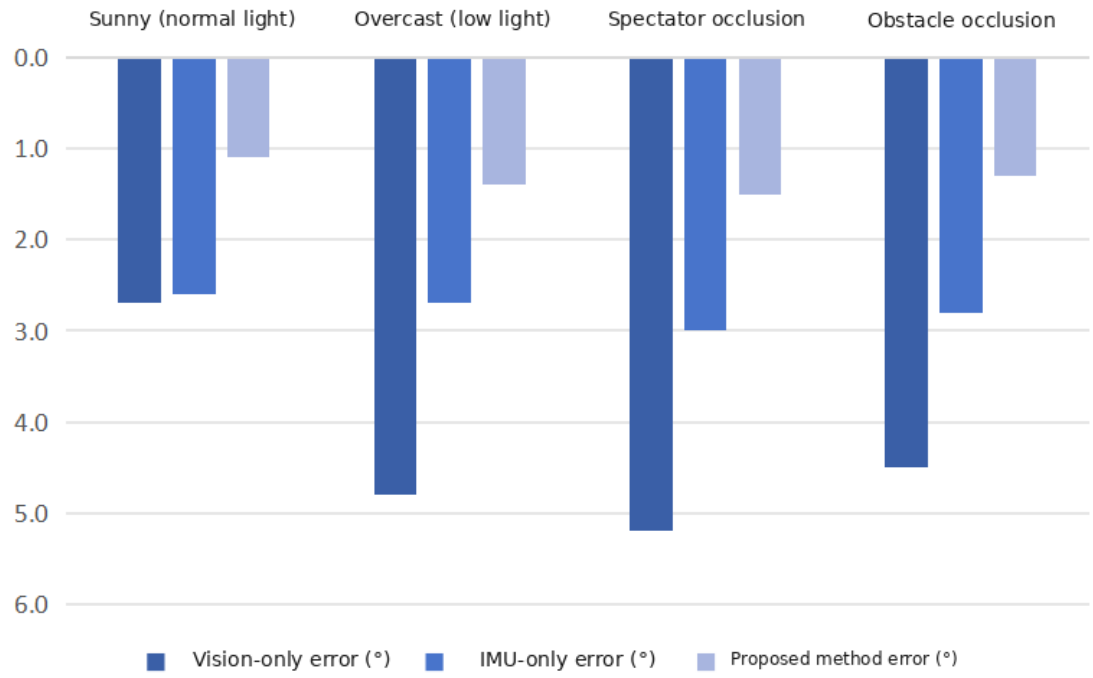


Figure 12. Angle measurement errors under different lighting and occlusion conditions.

Conclusion

This study proposes a multimodal method for dynamic 3D golf pose recognition and angle analysis. Attention-based fusion of vision and IMU data address the limited accuracy and robustness of conventional methods. The golf-specific angle model quantifies key swing parameters for training assessment. The absolute key-angle error did not exceed 1.5° and the processing rate reached 27.8 FPS, supporting technique refinement and evidence-based golf training. The study did not analyze the club trajectory, limiting its ability to characterize swing coordination fully. This remains a subject for further work.

Funding

This research was not supported by any funds.

Acknowledgements

The authors would like to show sincere thanks to those technicians who have contributed to this research.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Bourgain, M., Rouch, P., Rouillon, O., Thoreux, P., Sauret, C. (2022) Golf swing biomechanics: a systematic review and methodological recommendations for kinematics. *Sports*, 10(6), 91.

[2] Watson, M., Coughlan, D., Clement, N. D., Murray, I. R., Murray, A. D., Miller, S. C. (2023) Biomechanical parameters of the golf swing associated with lower back pain: a systematic review. *Journal of Sports Sciences*, 41(24), 2236-2250.

[3] Kim, S. E., Burket Koltsov, J. C., Richards, A. W., Zhou, J., Schadl, K., Ladd, A. L., Rose, J. (2023) Validation of inertial measurement units for analyzing golf swing rotational biomechanics. *Sensors*, 23(20), 8433.

[4] Ye, H., Zhu, W., Wang, C., Wu, R., Wang, Y. (2022) Faster voxelpose: real-time 3d human pose estimation by orthographic projection. *European Conference on Computer Vision*, 142-159.

[5] Yi, X., Zhou, Y., Xu, F. (2021) TransPose: real-time 3D human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics (TOG)*, 40(4), 1-13.

[6] Mahony, R., Hamel, T., Pflimlin, J. M. (2008) Nonlinear complementary filters on the special orthogonal group. *IEEE Transactions on Automatic Control*, 53(5), 1203-1218.

[7] Han, D., Wang, Z., Xia, Z., Han, Y., Pu, Y., Ge, C., Huang, G. (2024) Demystify mamba in vision: a linear attention perspective. *Advances in Neural Information Processing Systems*, 37, 127181-127203.

[8] Shuai, H., Wu, L., Liu, Q. (2022) Adaptive multi-view and temporal fusing transformer for 3D human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4), 4122-4135.

[9] Zhang, X., Zhou, Z., Han, Y., Meng, H., Yang, M., Rajasegarar, S. (2023) Deep learning-based real-time 3D human pose estimation. *Engineering Applications of Artificial Intelligence*, 119, 105813.

[10] Shafiq, M., Gu, Z. (2022) Deep residual learning for image recognition: a survey. *Applied Sciences*, 12(18), 8972.

[11] Chen, R. C., Dewi, C., Zhuang, Y. C., Chen, J. K. (2023) Contrast limited adaptive histogram equalization for recognizing road marking at night based on YOLO models. *IEEE Access*, 11, 92926-92942.

[12] Sun, P., Zhou, F., Wang, L., Zhang, W., Liu, Y., Yang, L. (2023) A novel global camera calibration method based on flexible multidirectional target. *IEEE Transactions on Instrumentation and Measurement*, 72, 1-13.

[13] Yamaguchi, H., Rizk, H., Amano, T., Hiromori, A., Ukyo, R., Yamada, S., Ohno, M. (2024) Towards intelligent environments: human sensing through 3d point cloud. *Journal of Reliable Intelligent Environments*, 10(3), 281-298.