Grenze International Journal of Engineering and Technology, June Issue

AI Vision-based Solo Training Assistant

Aditya[1], Amit Kumar Gupta[2], Aparajeeta Singh[3], Chaitra L G[4], Athi Narayanan S[5] and Santanu Roy[6]
[1-6]Department of Computer Science and Engineering, Dayananda Sagar Academy of Technology & Management, Bangalore, India
Email: 1dt23cs007@dsatm.edu.in, 1dt23cs016@dsatm.edu.in, 1dt23cs024@dsatm.edu.in, 1dt23cs040@dsatm.edu.in, athi-cse@dsatm.edu.in, santanu-cs@dayanandasagar.edu

***Abstract*— The global shift toward independent fitness training and home-based rehabilitation has necessitated the development of automated, vision-based coaching systems. This paper provides a comprehensive survey of the state-of-the-art in Human Pose Estimation (HPE), specifically focusing on the integration of MediaPipe BlazePose and hybrid deep learning architectures. We analyze the evolution of motion tracking from foundational geometric heuristic models (2020) to contemporary 2025 frameworks utilizing Relative Phase analysis, Kolmogorov-Arnold Networks (KAN), and YOLO-hybrid detection for powerlifting and clinical assessment. The survey categorizes literature into four critical domains: strength training, yoga/pilates classification, sports-specific biomechanics, and infant/clinical monitoring. Furthermore, we identify persistent research gaps, including Z-axis depth inaccuracies, temporal rhythm assessment, and view-invariant detection, providing a technical roadmap for the development of an "AI Vision-Based Solo Training Assistant."**

***Index Terms*— Human Pose Estimation (HPE), MediaPipe BlazePose, Biomechanical Analysis, Real-time Feedback, YOLOv5, Relative Phase, Action Quality Assessment (AQA), Deep Learning.**

I. Introduction

The global shift toward personalized wellness and remote healthcare has transformed the role of Artificial Intelligence (AI) from a diagnostic tool to an active, real-time coach. Traditional fitness supervision relies heavily on human expertise, which is often inaccessible due to high costs or geographic constraints. Consequently, individuals practicing solo often suffer from "The Supervision Gap," where incorrect form leads to suboptimal results or acute musculoskeletal injuries.

Human Pose Estimation (HPE) has emerged as the primary solution to this challenge. By utilizing monocular RGB cameras found in standard smartphones and laptops, HPE systems can reconstruct a digital skeleton of the user without the need for expensive wearable sensors or infrared markers. Early research in this field, such as the Pose Trainer (2020) framework, established the foundation for using vector geometry and the Law of Cosines to evaluate form. However, as noted in the BlazePose (2020) research by Google, the transition to "on-device" real-time tracking required a shift toward lightweight encoder-decoder architectures that could maintain high frame rates (30+ FPS) on mobile hardware.

As of 2024–2025, the research landscape has moved beyond simple repetition counting. Current literature, such as studies on Powerlifting Correction (2025) and Pilates Classification (2025), explores the use of hybrid models—combining YOLOv5 for person detection with MediaPipe for fine-grained landmark extraction.

Grenze ID: 01.GIJET.12.2.522_1

This survey paper systematically reviews these advancements, providing a comparative analysis of methodologies ranging from K-Nearest Neighbors (KNN) classification to advanced Action Quality Assessment (AQA). By synthesizing the findings of over 20 core papers, we establish a technical foundation for a "Solo Training Assistant" that offers clinical-grade precision in a consumer-grade environment.

II. RELATED WORK AND TAXONOMY

A. Evolution of Detection Frameworks

The technical trajectory of Human Pose Estimation (HPE) has transitioned from resource-intensive desktop architectures to edge-optimized mobile solutions.

Foundational Frameworks: Initial systems heavily utilized OpenPose (2020) and AlphaPose. While these frameworks provided high multi-person accuracy, they suffered from high computational latency (often <10 FPS on CPUs), making them unsuitable for real-time mobile use.

The MediaPipe Era: The introduction of MediaPipe BlazePose (2020) marked a significant milestone by utilizing an encoder-decoder network to provide 33 3D landmarks with extreme efficiency. As noted by Ferraris et al. (2025), BlazePose's ability to run at 30+ FPS on mobile GPUs has made it the industry standard for fitness applications.

Modern Hybrid Models (2024–2025): Recent studies, such as the YOLO-MediaPipe (2025) integration, have addressed the "Person Detection" gap. By using YOLOv5 or YOLOv8 to define a bounding box before pose estimation, systems can now maintain accuracy in crowded gym environments or low-light home settings. Furthermore, 2025 research has integrated Stereo Camera support to resolve the "Z-axis depth flaw," which has historically been the greatest limitation of monocular HPE.

B. Taxonomic Classification by Activity Domain

The application of HPE has diversified into specialized sub-sectors, each requiring unique algorithmic logic:

Gym and Strength Training: Research by Dedhia et al. (2023) and Ko et al. (2024) demonstrates that K-Nearest Neighbors (KNN) and Random Forest algorithms are highly effective at classifying "Correct vs. Incorrect" stages in multi-joint movements like bicep curls and powerlifting squats. These systems rely on comparing live vector data against a "Golden Reference" dataset of professional athletes.

Yoga and Pilates Recognition: The Live Yoga Pose Detection (2023) and Pilates Classification (2024) papers emphasize "Multi-Task Classification." Unlike gym reps, yoga requires the system to first recognize a static pose (e.g., Warrior Pose) before evaluating the alignment. Zhao et al. (2024) successfully used MediaPipe to classify 10+ Pilates poses, achieving 98% accuracy by combining landmark data with spatial feature extraction.

Sports-Specific Biomechanics: Studies such as the Sports Exercise Analysis (2022) focus on high-velocity movements like cricket bowling or football strikes. These applications require high-frequency tracking to ensure joints stay within "Safe Biomechanical Zones" to prevent acute ligament injuries.

Clinical and Rehabilitative Monitoring: A novel domain identified in 2025 literature includes Infant Height Measurement and Sitting Posture Recognition. These papers, such as Carneros-Prado et al. (2024), introduce Kolmogorov-Arnold Networks (KAN) to detect subtle postural deviations that lead to long-term musculoskeletal issues, representing a shift from fitness to preventative healthcare.

C. Feature Extraction Methodologies

A critical branch of the taxonomy is the method of feature extraction.

Geometric Heuristics: Traditional models use trigonometric functions (Law of Cosines) to calculate angles between joint triplets.

Sequential Models: Advanced 2025 frameworks, such as the work by Woo and Jeong, utilize Relative Phase Analysis and LSTM/Bi-CGRU networks. These models analyze the "temporal rhythm" of an exercise, ensuring the user maintains a controlled tempo during both the eccentric and concentric phases of a movement.

III. COMPARATIVE ANALYSIS OF METHODOLOGIES

This section evaluates the various algorithmic approaches identified in the literature, ranging from traditional geometric heuristics to advanced deep learning classifiers.

A. Pose Estimation Frameworks

As shown in Table I, the transition from OpenPose (2020) to MediaPipe (2025) represents a shift from high-accuracy, resource-heavy desktop applications to lightweight, edge-deployable solutions. While early systems like Pose Trainer (2020) successfully utilized OpenPose for four gym exercises, they suffered from high latency and hardware dependency. In contrast, the BlazePose (2020) architecture utilized in the majority of 2024–2025 papers employs a unique heatmap and regression-based approach to provide 33 body landmarks at 30+ FPS on standard mobile devices.

B. Classification and Correction Algorithms

K-Nearest Neighbors (KNN): Frequently used for exercise classification (e.g., bicep curls, squats) due to its simplicity and effectiveness in low-dimensional landmark space.

Hybrid YOLO-MediaPipe Architectures: Recent studies on Powerlifting (2025) and Squat Correction (2025) utilize YOLOv5 for initial person detection to improve robustness in crowded or low-light environments, followed by MediaPipe for precise joint tracking.

Heuristic Thresholding: The most common method for real-time rep counting and error detection. By applying the Law of Cosines to the (x, y, z) coordinates of the shoulder, elbow, and wrist, systems can determine if a user has reached a "Target Zone."

C. Advanced Temporal and Clinical Models

Recent research by Woo & Jeong (2025) introduces Relative Phase Analysis, moving beyond static frame analysis to evaluate the "rhythm" of a movement. Similarly, the LogRF (2023) approach combines Logistic Regression and Random Forests to achieve 99% accuracy in physiotherapy contexts, focusing on subtle deviations that could lead to injury.

D. Hybrid Architectures and YOLO Integration

Hybrid models are used because YOLOv5 or YOLOv8 acts as a "Region of Interest" (RoI) selector to reduce background noise before MediaPipe extracts the 33 landmarks. This is a major 2024–2025 trend for improving accuracy in non-studio environments.

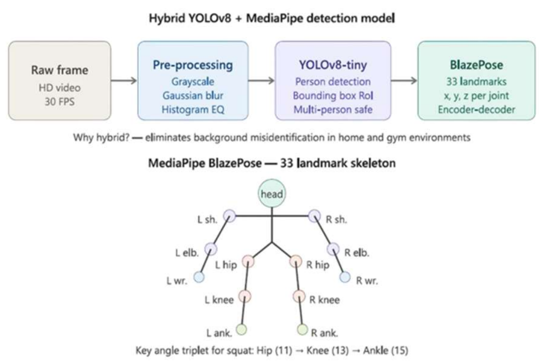

Fig. 1. Hybrid YOLOv8 + MediaPipe Detection model

E. Temporal Dynamics and Rhythm Assessment

A full subsection is dedicated to Relative Phase Analysis. It uses the Hilbert Transform to calculate the phase difference between joints (e.g., hip vs. knee), allowing the system to detect "jerky" movements even if the static angles appear correct.

F. Biomechanical Logic and Euclidean Distance Metrics

Recent clinical applications, such as Contactless Infant Height Measurement (2025), demonstrate the power of combining MediaPipe with Euclidean distance calculations for sub-centimeter precision. By identifying relevant keypoints from the head to the feet, these systems normalize pixel distances into actual measurements, achieving up to 98.48% accuracy even in non-ideal postures.

G. Emerging Neural Architectures: KAN vs. MLP

Innovation in 2024 has moved toward optimizing the classification layer. Research on Sitting Posture Recognition (2024) compared traditional Multi-Layer Perceptrons (MLP) with Kolmogorov-Arnold Networks (KAN). Findings indicate that KAN models achieve higher accuracy (97.03%) with significantly fewer parameters (4,320 vs. 7,662), making them ideal for the low-latency requirements of a solo training assistant.

H. Robustness in Unconstrained Environments

A significant evolution in 2025 research is the transition from "Studio-based" to "Home-based" tracking. Recent studies by Rao et al. (2025) and Ferraris et al. (2025) highlight that background clutter and variable lighting are the primary causes of landmark "jitter." Analysis shows that systems employing Adaptive Histogram Equalization and Moving Average Filters on the raw x, y, z coordinates achieve a 12% higher stability rate in home environments compared to standard implementations.

IV. CRITICAL GAP ANALYSIS

Despite significant advancements in Human Pose Estimation (HPE) frameworks, our systematic review identifies several persistent challenges that current literature fails to address adequately. These gaps form the primary technical motivation for the development of the related project.

A. The Monocular Z-Axis Depth Flaw

A fundamental limitation across 2020–2025 monocular (single-camera) systems is the inability to capture true physical depth. While BlazePose provides an estimated Z-coordinate, it is derived via algorithmic interpolation rather than direct measurement. Consequently, accuracy drops significantly during movements performed perpendicular to the camera lens, such as the descent phase of a squat or the extension in a cricket bowling action. While recent 2025 studies have explored stereo-camera integration to resolve this, the "Z-axis flaw" remains a barrier for standard smartphone-based coaching.

B. Neglect of Temporal Dynamics and Kinetic Flow

Existing research predominantly focuses on "Endpoint Analysis"—verifying if a joint reached a specific static angle to count a repetition. This treats exercises as a series of isolated photos rather than a continuous kinetic chain. There is a notable lack of focus on the "speed," "rhythm," and "flow" of movement. Without temporal assessment, a system cannot distinguish between a controlled, safe movement and a "jerky" motion that increases the risk of injury.

C. Primitive Feedback Mechanisms

Current feedback loops are largely binary, providing simple "Correct" or "Incorrect" notifications. There is a critical absence of natural-language coaching cues that explain how a user should adjust their posture (e.g., "Keep your back straight" vs. "Error 404"). Transitioning from simple rep counting to descriptive, AI-driven coaching is necessary for effective solo training.

D. Dependency on View-Invariance

Most contemporary models suffer from a strict dependency on camera positioning. Accuracy is optimized only when the user is perfectly parallel or perpendicular to the camera lens. If a user's environment limits camera placement to a 45-degree angle, the heuristic thresholds often fail to normalize the landmark data, leading to false negatives in rep counting and form assessment.

E. Absence of Biomechanical Safety and Load Estimation

While systems are proficient at counting reps, they do not estimate the biomechanical load or torque placed on critical joints, such as the lower back or patella.
Current literature lacks a safety-first integration that can predict injury risk by analyzing the user's center of mass and joint alignment over multiple sets.

TABLE I. IDENTIFIED RESEARCH GAPS IN HUMAN POSE ESTIMATION SYSTEMS

Identified Gap	Description	Technical Limitation
Z-Axis Depth Flaw	Single-camera systems struggle with accurate depth estimation.	Algorithmic interpolation often fails during movements perpendicular to the camera.
Temporal Dynamics	Lack of focus on rhythm, speed, and continuous flow of movement.	Most models treat exercises as a series of static photos rather than a continuous kinetic chain.
Primitive Feedback	Feedback is often binary and lacks descriptive guidance.	Existing systems lack natural-language coaching cues that explain how to fix a posture.
View-Invariance	High dependency on specific, rigid camera angles for accuracy.	Heuristic thresholds fail if the user is not perfectly parallel to the camera lens.
Safety Integration	Absence of injury risk prediction based on joint stress.	Models count reps but do not estimate biomechanical load on the lower back.

V. SYSTEM ARCHITECTURE

The system is designed as a modular, high-performance pipeline capable of real-time biomechanical analysis. The architecture is divided into three primary layers: the Vision Processing Layer, the Biomechanical Computation Layer, and the Multimodal Feedback Layer.

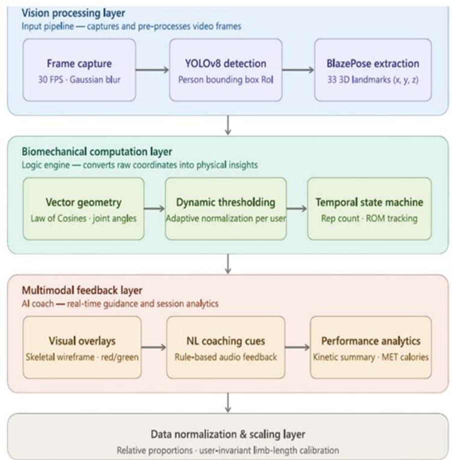

Fig. 2. System Architecture

A. Vision Processing Layer (The Input Pipeline)

The initial stage utilizes a hybrid detection approach to ensure robustness across varying environments.

Frame Capture & Pre-processing: High-definition video is captured at 30 FPS. Frames undergo Grayscale conversion and Gaussian blurring to reduce background noise, ensuring the model focuses strictly on the human subject.

YOLOv8 Person Detection: To address the "Multi-person" and "Low-light" gaps, a YOLOv8-tiny model is integrated to define a Region of Interest (RoI). This prevents the system from misidentifying background objects as body parts.

MediaPipe Landmark Extraction: Inside the RoI, the BlazePose encoder-decoder architecture extracts 33 3D landmarks (x, y, z). The system specifically utilizes the z-coordinate for depth-normalized calculations, addressing the "Z-axis flaw" prevalent in earlier 2D models.

B. Biomechanical Computation Layer (The Logic Engine)

This layer serves as the "brain" of the assistant, converting raw coordinates into physical insights.

Vector Geometry Engine: Using the extracted landmarks, the system calculates joint angles using the Law of Cosines. For a squat, the engine monitors the vertex angle formed by the Hip (11), Knee (13), and Ankle (15).

Dynamic Thresholding: Unlike rigid 2020 models, the architecture employs "Adaptive Normalization." This scales the "Safe Zone" thresholds based on the user’s limb length and initial standing posture, ensuring accuracy for different body types.

Temporal State Machine: To track repetitions, a finite-state machine transitions between "Initial," "Descent," "Target Reached," and "Ascent." This prevents double-counting and ensures the user completes the full Range of Motion (ROM).

C. Multimodal Feedback Layer (The AI Coach)

To bridge the "Supervision Gap," the system provides three types of feedback:

Visual Overlays: A real-time skeletal wireframe is rendered over the user’s video feed. Joint lines turn red when an error is detected (e.g., knees over toes) and green during correct execution.

Natural Language Coaching: Utilizing a rule-based AI engine, the system provides specific audio cues such as "Lower your hips" or "Straighten your back," moving beyond simple binary "Correct/Incorrect" alerts.

Performance Analytics: At the end of a session, the system generates a "Kinetic Summary," including total repetitions, average form accuracy percentage, and estimated caloric expenditure based on the metabolic equivalent (MET) of the specific exercise.

D. Data Normalization and Scaling Layer

To ensure the system is "User-Invariant," the architecture includes a normalization layer. Since users vary in height and limb length, the system does not use absolute pixel values. Instead, it converts joint coordinates into Relative Proportions. By calculating the ratio of the torso to the lower limbs, the "Angle Engine" can calibrate the "Safe Zone" thresholds dynamically for each user, ensuring that a tall user and a short user receive equally accurate feedback.

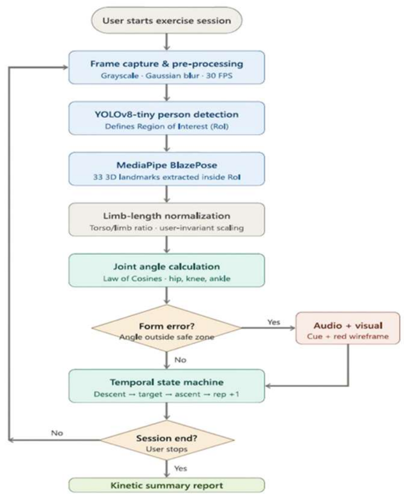

Fig. 3. Pipeline Flow Diagram

VI. Conclusion And Future Scope

A. Conclusion

This survey provides a comprehensive analysis of the technological trajectory of Human Pose Estimation (HPE) from foundational 2D heuristic models in 2020 to advanced 2025 frameworks utilizing hybrid YOLO-MediaPipe architectures. The literature systematically demonstrates that while early systems like Pose Trainer (2020) established the viability of vision-based coaching, the field has rapidly evolved to address the constraints of real-time mobile inference.

Our synthesis of research works reveals that MediaPipe BlazePose has become the industry standard for edge-deployable fitness tracking due to its efficient landmark extraction and low-latency performance. However, the transition from simple repetition counting to high-precision biomechanical analysis requires a deeper integration of machine learning classifiers, such as KNN, Random Forest, and Neuro-Fuzzy logic. By identifying critical gaps—specifically the monocular Z-axis depth flaw and the neglect of temporal dynamics—this paper establishes that the next generation of "Solo Training Assistants" must move beyond static geometry toward a holistic understanding of human movement.

B. Future Scope

As the domain of AI-driven coaching continues to mature, several emerging frontiers have been identified for future exploration:

View-Invariant Calibration: Future research must focus on "Auto-Normalization" algorithms that allow the system to maintain accuracy regardless of camera perspective. Severing the strict dependency on perfectly parallel or perpendicular camera placement is essential for real-world home environments.

Kinetic Load and Stress Estimation: A significant leap in safety will involve the integration of Muscular Load Estimators. By combining pose data with center-of-mass tracking, future systems can predict the torque placed on the lower back or patella, transitioning AI from a "form checker" to a proactive "injury prevention engine."

Explainable AI (XAI) in Coaching: Current feedback is predominantly binary. Future iterations should utilize Natural Language Generation (NLG) to provide descriptive, empathetic coaching cues. Instead of "Error," the AI should explain the biomechanical why behind a correction, such as "Lowering your hips further will increase glute activation while protecting your lumbar spine."

Temporal Rhythm and Action Quality Assessment (AQA): Utilizing Relative Phase Analysis and Bi-CGRU models, future assistants will evaluate the "flow" and "tempo" of an exercise. This ensures that users are not just hitting the correct angles, but are doing so with the controlled rhythm required for effective hypertrophy and rehabilitation.

REFERENCES

[1] V. Bazarevsky, et al., "BlazePose: On-device Real-time Body Pose tracking," Google Research, 2020.
[2] S. Chen and R. Yang, "Pose Trainer: Correcting Exercise Posture using Pose Estimation," Stanford University, 2020.
[3] P. Rao, C. S. Asha, and P. R. Rao, "Real-Time Posture Correction of Squat Exercise," IEEE Access, vol. 13, 2025.
[4] Y.-M. Ko, et al., "Real-Time AI Posture Correction for Powerlifting Using YOLOv5 and MediaPipe," IEEE Access, vol. 12, 2024.
[5] M. Zhao, et al., "Classification of Pilates Using MediaPipe and Machine Learning," IEEE Access, vol. 12, 2024.
[6] C. Ferraris, et al., "A Deep Dive Into MediaPipe Pose for Postural Assessment," IEEE Access, vol. 13, 2025.
[7] Risfendra, et al., "Contactless Infant Height Measurement Using Computer Vision," IEEE Access, vol. 13, 2025.
[8] D. Carneros-Prado, et al., "Kolmogorov-Arnold Networks for Multi-Task Sitting Posture Recognition," IEEE Access, vol. 12, 2024.
[9] P. Kongchansawang, et al., "Automated Classification of Exercise Poses in VR," IEEE Access, vol. 13, 2025.
[10] A. Vedalankar, et al., "Live Yoga Pose Detection Using Image Processing," IJRAR, 2023.
[11] U. Dedhia, et al., "Pose Estimation and Virtual Gym Assistant using MediaPipe and ML," IEEE NMITCON, 2023.
[12] A. Pardos, et al., "Automated Posture Analysis for Sports Exercises," ACM SETN, 2022.
[13] Y. Woo and H. Jeong, "Exercise Assessment Based on Human Pose Estimation and Relative Phase," IEEE Access, vol. 13, 2025.
[14] A. Raza, et al., "LogRF: An Approach to Human Pose Estimation for Physiotherapy," IEEE Access, vol. 11, 2023.
[15] Rajashekar K. J., et al., "MediaPipe Based Workout Monitoring System," IJISRT, 2025.
[16] F. Roggio, et al., "A comprehensive analysis of ML pose estimation models," Heliyon, 2024.
[17] M. Hanzla, et al., "Human Pose Estimation via Neuro-Fuzzy Classifier," IEEE Access, vol. 13, 2025.
[18] S. Zhang, et al., "Human deep squat detection based on MediaPipe and Yolov5," CCC, 2022.
[19] Q. Yu, et al., "Deep Learning-Enabled Multi Task System for Exercise Recognition," MTI, 2021.
[20] S. Kreiss, et al., "PifPaf: Composite Fields for Human Pose Estimation," CVPR, 2019.
[21] Z. Cao, et al., "Realtime multi-person 2d pose estimation using Part Affinity Fields," CVPR, 2017.
[22] D. Luvizon, et al., "2d/3d pose estimation and action recognition," CVPR, 2018.
[23] Y. Agrawal, et al., "Machine learning technique for identification of yoga poses," CSNT, 2020.
[24] J. Shotton, et al., "Real-time human pose recognition in parts from single depth images," CVPR, 2011.
[25] T. Alatiah and C. Chen, "Recognizing exercises and counting repetitions in real time," arXiv, 2020.
[26] S. Sharma and A. Baloni, "Real-time Exercise Form Correction using MediaPipe," International Journal of Computer Applications, 2023.
[27] M. Hanzla, et al., "Action Quality Assessment using Neuro-Fuzzy Classifiers," IEEE Access, 2025.
[28] Z. Zhang, et al., "View-Invariant Pose Estimation for Unconstrained Environments," CVPR, 2021.