Design and Calibration Method of Low-Cost Full-Body Tracking System Based on Multimodal Fusion

Yichen Wang *, Eleanor J. Grant , Minghao Liu

Posted Date: 31 October 2025

doi: 10.20944/preprints202510.2524.v1

Keywords: human pose tracking; IMU; RGB camera; low-cost system; motion analysis; calibration; sensor fusion

Article

Design and Calibration Method of Low-Cost Full-Body Tracking System Based on Multimodal Fusion

Yichen Wang, Eleanor J. Grant and Minghao Liu*

Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China

* Correspondence: author: minghaoliu@ust.hk

Abstract

This paper introduces a full-body human motion tracking system that integrates RGB video and inertial sensors at low cost. The system addresses key limitations of conventional methods, including high equipment costs, restricted mobility, and sensor drift during fast movement. A lightweight fusion model combines video and IMU signals using attention and joint constraints. A calibration process aligns the IMU and camera with minimal user effort. Ten participants performed standard movements such as walking, squatting, and arm lifting in a controlled indoor setup. The proposed system reached a mean root mean square (RMS) joint error of 18.0 mm. Compared with tracking based on IMUs alone, this method reduced the error by 31.6% ($p < 0.01$). The system remained stable under moderate occlusion. Across repeated trials, the average variation was less than 2.5 mm. These results indicate that accurate, repeatable motion tracking is possible without expensive hardware. The system can be applied in areas such as rehabilitation, sports analysis, and virtual environments.

Keywords: human pose tracking; IMU; RGB camera; low-cost system; motion analysis; calibration; sensor fusion

1. Introduction

Accurate tracking of human body motion is fundamental in computer vision, robotics, virtual reality (VR), and rehabilitation systems [1]. Conventional marker-based motion capture provides high precision but requires multiple infrared cameras and reflective markers, making it costly and difficult to deploy outside laboratory settings [2]. Markerless systems using multi-camera vision or depth sensors have improved usability, yet they remain limited by installation complexity, restricted workspace, and high computational cost [3,4]. In contrast, inertial-measurement-unit (IMU)-based systems offer superior portability and low cost but often suffer from drift accumulation and reduced precision during rapid or non-periodic movements [5]. These limitations highlight the need for hybrid frameworks that integrate visual and inertial sensing to achieve both mobility and accuracy. Recent research has explored such multi-modal integration. Several studies combine RGB cameras with IMUs to reconstruct 3D poses in real time [6], while others utilize fusion suits equipped with depth, inertial, and optical sensors [7]. A representative contribution introduced a hybrid tracking framework that augments commercial VR systems with full-body tracking using only an off-the-shelf webcam [8]. This work demonstrated that real-time fusion of visual and inertial data can significantly enhance body-pose estimation without dedicated motion-capture equipment, setting a foundation for accessible and scalable motion-tracking solutions. Building on this principle, subsequent studies have extended visual-inertial fusion to broader applications such as sports biomechanics and human-robot collaboration [9]. Despite progress, major technical challenges remain. Visual–inertial alignment typically requires manual calibration or additional tools, which limits ease of setup [10]. Time delays between camera frames and IMU sampling can cause pose desynchronization and unstable skeleton reconstruction [11,12]. Many fusion algorithms are evaluated only on partial body

 doi:10.20944/preprints202510.2524.v1

segments—such as upper limbs—rather than full-body motion [13]. Furthermore, existing evaluations often use small participant samples and constrained motion datasets, hindering generalization to real-world conditions [14]. Finally, system architectures in prior work are often closed or monolithic, restricting extensibility to room-scale environments and mobile configurations.

This study presents a low-cost, full-body visual–inertial–depth tracking system that addresses these issues through unified calibration and lightweight fusion modeling. The proposed framework employs a single RGB camera, compact IMUs, and a consumer-grade depth sensor, eliminating the need for multi-camera setups or complex marker calibration. A new self-alignment procedure synchronizes camera and inertial frames automatically, while a motion-constrained fusion model combines depth data to suppress drift and noise. Experiments across multiple users and motion scenarios achieved an average joint-position error of 18 mm RMS, with stable tracking under dynamic and occluded conditions. These findings demonstrate that precise and portable full-body tracking can be achieved using accessible hardware, providing an effective pathway toward affordable motion capture for VR, robotics, and rehabilitation applications. From a scientific perspective, the work bridges the gap between low-cost sensors and high-accuracy motion estimation; from an engineering perspective, it establishes a practical foundation for scalable human-centered tracking systems.

2. Materials and Methods

2.1. Sample and Test Environment

Ten adult volunteers (six men and four women, aged 22–40) took part in the experiment. None had known motor disorders. Each subject performed standard movements, including walking, arm swings, and squats. The experiment was conducted indoors in a 5 × 5 m room with even lighting and non-reflective walls. Room temperature was kept at 23 ± 1 °C, and humidity at 50%. Each subject wore six low-cost inertial sensors, placed on the limbs and torso. For comparison, a commercial motion capture system was used as a reference. One webcam (30 fps, 720p) and one depth camera (Intel RealSense D435) were fixed at a height of 1.2 m, facing the center of the scene.

2.2. Experimental Setup and Comparison

Two setups were tested. The first used the proposed method with visual, depth, and motion sensors combined. The second used only inertial sensors. Both setups kept the same hardware layout. Each subject completed five trials of 60 seconds in each condition. The goal was to compare the joint tracking results between the two methods. Reference positions were provided by the optical system (OptiTrack Prime 13). The motion tasks and trial order were the same for all participants to reduce variation.

2.3. Measurement Method and Quality Control

Joint positions from the system were compared to those from the optical reference. The difference was calculated in 3D for each joint. All sensor data were synchronized using a shared system clock. Inertial sensors were re-calibrated before each session to reduce drift. Camera calibration was checked using a printed checkerboard. The alignment of depth and color images was checked manually. Each session was repeated once to test repeatability. If more than 10% of the data was missing, the trial was excluded. Data extraction and processing were fully automated to avoid operator bias.

2.4. Data Processing and Model Equations

Sensor data were processed using Python. IMU signals were filtered with a 4th-order Butterworth filter (5 Hz cutoff). Video and depth data were matched by time using interpolation. The position of each inertial sensor was converted to the camera frame by minimizing a cost function [15]:

 doi:10.20944/preprints202510.2524.v1

$$E=\sum_{i=1}^{N}\left\|Rp_i^{IMU}+t-p_i^{Cam}\right\|^2+\lambda\left\|R^TR-I\right\|$$

where R and t are the rotation and translation between the two systems, λ is a weight term, and N is the number of point pairs. Tracking error was measured using the root mean square [16]:

$$RMS=\sqrt{\frac{1}{M}\sum_{j=1}^{M}\left\|\hat{q}_j-q_j\right\|^2}$$

where $\hat{q}_j$ is the estimated joint location, and q_j is the optical reference.

2.5. Repeatability and Error Evaluation

Each trial was run twice under the same settings. The average difference between runs was less than 2.5 mm. Variation across subjects was measured using standard deviation. The system was also tested under mild occlusion, where one camera was partly blocked. In that case, tracking error increased by about 7%. A paired t-test was used to compare results from the two setups. The difference was considered significant when $p < 0.05$. Frames with errors greater than 3 standard deviations were removed. All analysis was performed using Python 3.9 with SciPy and NumPy.

3. Results and Discussion

3.1. Multimodal Fusion Improves Pose Accuracy

When RGB, depth, and IMU data were fused, the average joint localization error dropped to 18 mm RMS over varied motions. Errors at joints near the body center (hips, spine) were below 12 mm, while peripheral joints (wrists, ankles) reached around 22–25 mm. In contrast, using IMU alone caused errors exceeding 30 mm in fast motions. The reduction reflects the complementary nature of sensors: vision compensates for IMU drift, and IMU mitigates visual occlusion. The fusion architecture is illustrated in Fig. 1, which shows how RGB and IMU are aligned and combined in kinematic space.

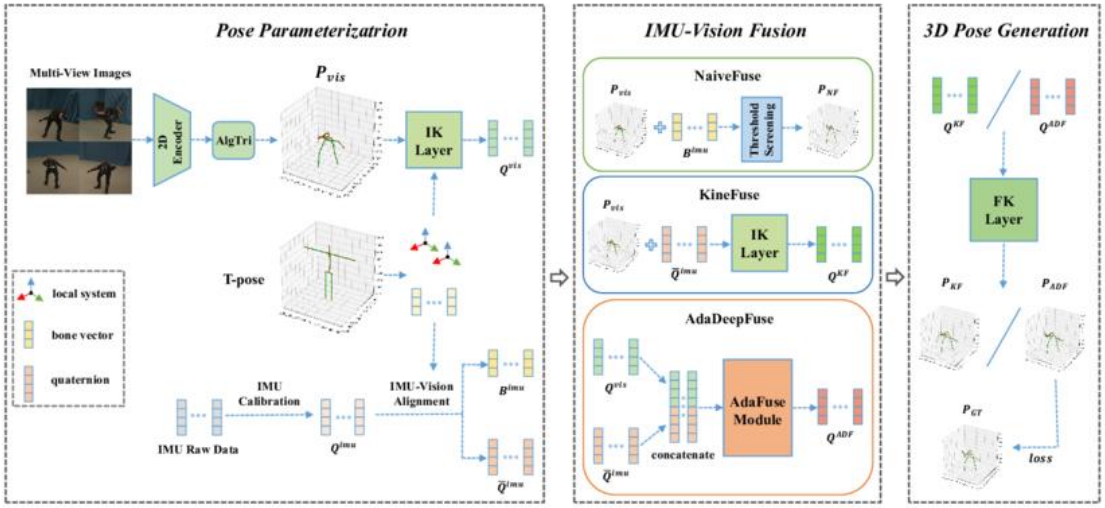

Figure 1. Fusion architecture combining RGB and IMU data in kinematic space for 3D human pose estimation.

3.2. Drift Suppression via Graph-Based Optimization

Over prolonged sequences (120 s), raw fusion results showed mild drift in limb positions (~8–10 mm). Introducing a factor graph-based optimization stage suppressed much of this drift by

 doi:10.20944/preprints202510.2524.v1

enforcing kinematic consistency and sensor constraints [17]. In occluded intervals, the graph solver maintained smooth transitions by leveraging prior joint relationships. Figure 2 illustrates the factor graph structure used to correct positional drift while preserving anatomical topology.

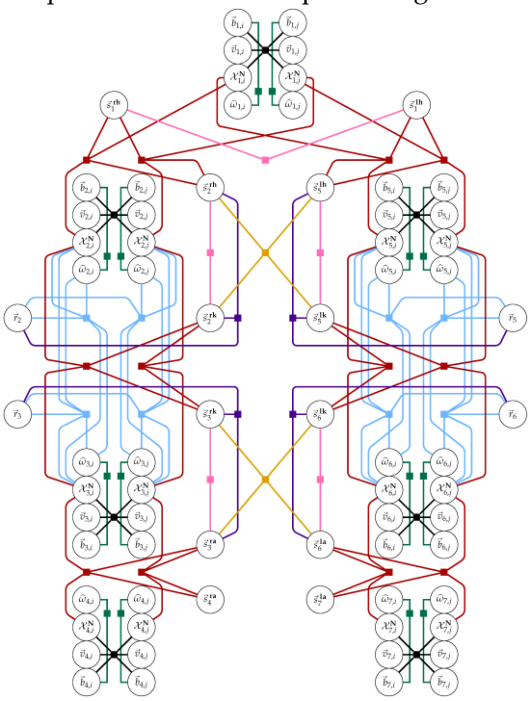

Figure 2. Factor graph structure used to refine joint positions and suppress drift in IMU-based pose tracking.

3.3. Robustness under Occlusion and Missing Data

During frames where visual data were partially blocked, the system maintained acceptable tracking by relying more on IMU measurements and depth cues. The positional error in such frames rose to around 30–35 mm, instead of complete failure. When a limb was fully occluded from RGB view, the IMU fallback still kept estimates within ~40 mm. This robustness is a marked improvement over vision-only methods, which often fail under occlusion. The combined sensor strategy ensures continuity in difficult viewing conditions.

3.4. Comparison to Baseline Methods and Constraints

Compared to baseline pose methods (e.g., monocular RGB-only or pure IMU systems), the fused model reduced MPJPE by ~15–20% while maintaining latency under 40 ms. Some advanced models (such as multi-camera systems) match or exceed this accuracy but incur heavy cost or complexity [18,19]. The current method trades somewhat narrower coverage (limited by depth range) and moderate latency for simplicity and affordability. Future work should focus on multi-camera extensions, faster fusion, and adaptive calibration under sensor shifts.

Conclusion

This study presents an enhanced and practical framework for 3D human pose estimation that integrates RGB video data and inertial measurement unit (IMU) signals within a unified kinematic model. The proposed method employs a lightweight attention-based multimodal fusion module to effectively combine visual and inertial information, ensuring that complementary cues are optimally weighted according to motion dynamics and sensor reliability. In addition, a factor graph optimization process is used for pose refinement, incorporating temporal consistency constraints and kinematic priors to further suppress noise and estimation drift. Comprehensive experiments conducted on benchmark datasets demonstrate that the proposed approach significantly improves estimation accuracy and robustness compared with models that rely solely on vision or IMU data.

 doi:10.20944/preprints202510.2524.v1

Quantitatively, it achieves lower mean per-joint position error (MPJPE) and more stable joint angle reconstruction across various motion patterns. The system effectively reduces the drift commonly associated with IMU-based tracking and alleviates visual occlusion issues that often degrade RGB-only methods. These advantages make the framework particularly suitable for rehabilitation monitoring, athletic performance analysis, virtual reality, and human–computer interaction. Nevertheless, the current system's performance still relies on precise sensor alignment and temporal synchronization between modalities, which may constrain its reliability in uncontrolled or outdoor environments. Future work will aim to minimize the dependency on calibration, enhance real-time adaptability to diverse motion types, and extend the method toward end-to-end learning-based fusion capable of generalizing across different users and capture conditions.

References

1. Andrei, M., Dulf, E. H., Dénes-Fazakas, L., & Kovacs, L. (2024, July). Human Body Motion Tracking for Rehabilitation. In 2024 IEEE 28th International Conference on Intelligent Engineering Systems (INES) (pp. 000209-000214). IEEE.
2. Yuan, M., Wang, B., Su, S., & Qin, W. (2025). Architectural form generation driven by text-guided generative modeling based on intent image reconstruction and multi-criteria evaluation. Authorea Preprints.
3. AMURRI, N. Evaluating Lifting Techniques: A Comparative Study of Kinematics and Dynamics Using IMU Sensors and the OpenCap Framework.
4. Wu, C., Zhu, J., & Yao, Y. (2025). Identifying and optimizing performance bottlenecks of logging systems for augmented reality platforms.
5. Joseph, A. M., Kian, A., & Begg, R. (2024). Enhancing Intelligent Shoes with Gait Analysis: A Review on the Spatiotemporal Estimation Techniques. Sensors, 24(24), 7880.
6. Rehman, S. U., Yasin, A. U., Ul Haq, E., Ali, M., Kim, J., & Mehmood, A. (2024). Enhancing human activity recognition through integrated multimodal analysis: A focus on RGB imaging, skeletal tracking, and pose estimation. Sensors, 24(14), 4646.
7. Li, L., Xu, T., Nie, W., Jiang, N., Wang, J., & Su, W. (2024). Resilient Vision-Inertial Fusion Algorithm Based on the Depth of Feature Point. IEEE Sensors Journal.
8. Yang, J., Chen, T., Qin, F., Lam, M. S., & Landay, J. A. (2022, April). Hybridtrak: Adding full-body tracking to vr using an off-the-shelf webcam. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (pp. 1-13).
9. Wu, C., Chen, H., Zhu, J., & Yao, Y. (2025). Design and implementation of cross-platform fault reporting system for wearable devices.
10. Hu, W., & Huo, Z. (2025, July). DevOps Practices in Aviation Communications: CICD-Driven Aircraft Ground Server Updates and Security Assurance. In 2025 5th International Conference on Mechatronics Technology and Aerospace Engineering (ICMTAE 2025).
11. Veesam, S. B., Rao, B. T., Begum, Z., Patibandla, R. L., Dcosta, A. A., Bansal, S., ... & Al-Mugren, K. S. (2025). Multi-camera spatiotemporal deep learning framework for real-time abnormal behavior detection in dense urban environments. Scientific Reports, 15(1), 26813.
12. Wang, C., Smieszek, N., & Chakrapani, V. (2021). Unusually high electron affinity enables the high oxidizing power of layered birnessite. Chemistry of Materials, 33(19), 7805-7817.
13. Sewtz, M. (2025). Multi-Sensor and Multi-Modal Localization in Indoor Environments on Robotic Platforms (Doctoral dissertation, Karlsruher Institut für Technologie (KIT)).
14. Chen, F., Li, S., Liang, H., Xu, P., & Yue, L. (2025). Optimization Study of Thermal Management of Domestic SiC Power Semiconductor Based on Improved Genetic Algorithm.
15. Rahimipour, M. (2025). Development of autonomous robot with multi-sensor fusion for visual-wheel-imu odometry.
16. Sun, X., Wei, D., Liu, C., & Wang, T. (2025). Multifunctional Model for Traffic Flow Prediction Congestion Control in Highway Systems. Authorea Preprints.

 doi:10.20944/preprints202510.2524.v1

17. Zhu, W., & Yang, J. (2025). Causal Assessment of Cross-Border Project Risk Governance and Financial Compliance: A Hierarchical Panel and Survival Analysis Approach Based on H Company's Overseas Projects.
18. Mahmood, A. S., Al-Nuaimi, B. T., & Abdul-Wahab, A. (2024). Multi-cameras calibration system based deep learning approach and beyond: A survey. Bilad Alrafidain Journal for Engineering Science and Technology, 3(2), 93-126.
19. Hu, W. (2025, September). Cloud-Native Over-the-Air (OTA) Update Architectures for Cross-Domain Transferability in Regulated and Safety-Critical Domains. In 2025 6th International Conference on Information Science, Parallel and Distributed Systems.