Real-Time Hand Pose Tracking using 6-Axis IMUs

Anik Sarker*
Meta Reality Labs Research
Redmond, Washington, USA
aniksarker@meta.com

Ziyi Kou
Meta Reality Labs Research
Redmond, Washington, USA
zkou@meta.com

Ergys Ristani
Meta Reality Labs Research
Redmond, Washington, USA
ristani@meta.com

Li Guan
Meta Reality Labs Research
Redmond, Washington, USA
liguan@meta.com

Taylor Niehues
Meta Reality Labs Research
Redmond, Washington, USA
taylor.niehues@meta.com

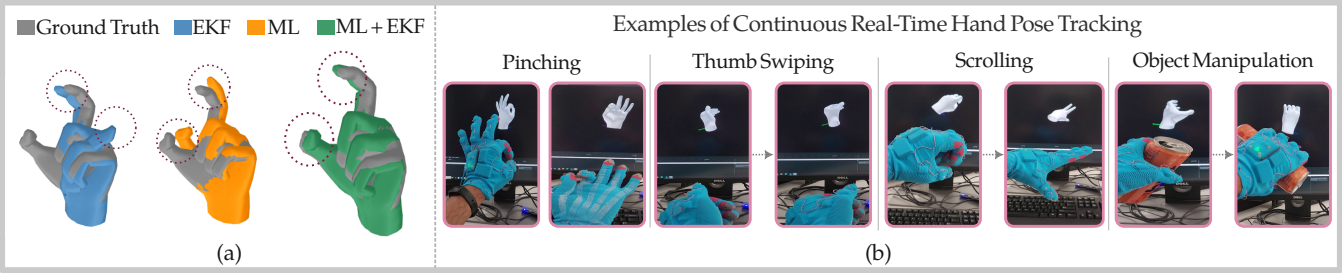

Figure 1: Hand pose estimation from a 6-axis IMU-based glove. (a) Comparative performance of hand pose estimation using EKF (blue), ML-only (orange), and the proposed ML+EKF (green) methods. The circled region highlights the thumb and index finger positions, showing that ML+EKF predictions are much better aligned with the ground truth than either EKF or ML alone. (b) Real-time implementation of ML+EKF. Sample hand poses are shown during live tracking.

Abstract

We introduce a real-time system for tracking hand pose using 6-axis inertial measurement units (IMUs) without requiring magnetometers or external sensors. Accurate hand pose tracking with only 6-axis IMUs is known to be fundamentally challenging due to the absence of a shared heading reference, leading to severe drift and inter-sensor misalignment. To overcome these limitations, we propose a hybrid method that combines a learning-based pose estimation approach followed by a late-stage Extended Kalman Filter (EKF). The learning-based model estimates noisy yet reasonable hand poses and is trained with drift-insensitive features like gravity vectors and wrist-relative gyroscope signals. On the other hand the EKF can appropriately filter the noise from pose estimates leading to robust tracking. Evaluated on a 12-hour dataset spanning 23 interaction tasks across 10 participants, our system improves joint angle accuracy by 40% over an EKF-only baseline and by 18% over a learning-only approach, achieving a mean joint error below 10°. The resulting framework enables real-time hand tracking invariant to magnetic perturbations, occlusion, or lighting changes, and is well suited for robotics, human–robot interaction (HRI), and human-computer interaction (HCI) applications.

CCS Concepts

• **Human-centered computing** → **Gestural input**; • **Computing methodologies** → *Motion capture*; Neural networks; • **Hardware** → *Sensors and actuators*.

Keywords

Hand Pose Tracking, 6-axis IMU, IMU glove tracking

ACM Reference Format:
Anik Sarker, Ziyi Kou, Ergys Ristani, Li Guan, and Taylor Niehues. 2026. Real-Time Hand Pose Tracking using 6-Axis IMUs. In *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI '26), March 16–19, 2026, Edinburgh, Scotland, UK*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3757279.3785628

1 Introduction

High-fidelity hand pose estimation is a key enabler of naturalistic interactions—from virtual object manipulation to expressive gestures—and a critical interface for immersive systems [10, 28, 40]. Reliable hand articulation tracking supports HRI, teleoperation, VR/AR, and rehabilitation. Data gloves are particularly effective for this setting: by placing sensors directly on the hand, they provide fine-grained, low-latency joint kinematics and have therefore seen broad adoption across HRI, VR/AR, biomechanics, and motion capture [25].

*This work was conducted while the author was a Ph.D. student at Virginia Tech and an intern at Meta Reality Labs Research.

HRI '26, Edinburgh, Scotland, UK

ACM ISBN 979-8-4007-2128-1/2026/03
https://doi.org/10.1145/3757279.3785628

Glove trackers often combine multiple sensing modalities, including 9-axis IMUs, 6-axis IMUs, EMG, and acoustic, tactile, or bend sensors, sometimes augmented by head-mounted or external cameras. While multimodal designs can improve accuracy, they introduce trade-offs in comfort, cost, power, calibration burden, and robustness to occlusion, lighting, and magnetic disturbance. In broader HRI settings, the tracker must be accurate, low-latency, comfortable, and reliable across diverse environments, which is difficult to achieve consistently with vision-augmented or magnetometer-dependent designs. Continuous tracking under fast, unconstrained hand motion is especially challenging for teleoperation and in-hand manipulation, where compounding drift, transient contacts, and frequent self-occlusion can break both camera-only and inertial-only pipelines.

We therefore target a wearable, magnetometer-free glove with twelve 6-axis IMUs (accelerometers and gyroscopes) for real-time, continuous hand kinematics tracking without external sensors. Avoiding magnetometers is particularly important in HRI: glove haptics (e.g., voice-coil motors, LRAs) and nearby ferrous objects can induce heading drift, forcing frequent re-calibration and reducing usability.

Prior 6-axis IMU gloves have been explored in constrained rehabilitation settings [19], but typically focus on slow, discrete motions (e.g., grip, thumb press) rather than the dynamic, diverse hand movements common in interactive systems and everyday manipulation. Robust continuous tracking in this regime is challenging: unlike 9-axis IMUs, 6-axis sensors lack a shared heading reference, so filter-based joint estimation becomes highly susceptible to drift and inter-sensor misalignment [10, 27]. We instead formulate 6-axis hand pose estimation as a learning problem. Although the mapping from a single inertial sample to pose can be one-to-many, short temporal windows concentrate hypotheses around the true pose. This design explicitly targets the long-horizon failure mode of magnetometer-free inertial tracking: small gyro bias and heading errors compound over time, so accurate short-horizon pose observations are essential to repeatedly re-anchor the kinematic state.

We introduce a hybrid tracking pipeline that combines learning-based inference with an explicit kinematic hand model. The system extracts drift-insensitive features, including gravity vectors and root-relative gyroscope signals, and feeds them to a low-latency Transformer to predict joint angles. To enforce temporal consistency and physical plausibility, we fuse these predictions with a hand-model-based Extended Kalman Filter (EKF), reducing noise and discontinuities relative to a purely learned solution. Before live tracking, the user performs a 10 second calibration gesture; thereafter, the tracker runs continuously without re-calibration.

We evaluate on a dataset of 12 hours from 10 participants performing 23 interaction tasks spanning fine gestures (e.g., pinching, typing) and dynamic actions (e.g., swiping, rotations, translations). The hybrid method improves joint-angle accuracy by 40% over an EKF-only baseline and 18% over ML-only, producing smoother and more temporally consistent estimates. These results support magnetometer-free, 6-axis IMU tracking as a practical component for robust deployment in HRI/HCI and as *complementary* to vision in occluded or in-hand interaction scenarios where camera tracking can degrade.

Our contributions are as follows: (1) demonstration of robust, continuous, real-time hand tracking using only 6-axis IMUs[1]; (2) design and insight of drift-insensitive features for magnetometer-free tracking; and (3) a hybrid ML+EKF framework optimized for standalone deployment.

2 Related Work

Hand tracking has been studied across optical motion capture, vision-based methods, and wearable gloves. Each paradigm trades off accuracy, cost, robustness to environment, and mobility. We review prior approaches and motivate a standalone, magnetometer-free 6-axis IMU glove for continuous tracking.

Optical motion capture. Systems such as OptiTrack [24] are a gold standard for high-precision 3D hand tracking using infrared cameras and markers [12, 17]. Paired with inverse kinematics, they yield low spatial error and are commonly used as ground truth [14, 36]. However, fixed infrastructure and controlled setups limit large-scale deployment.

Vision-based hand tracking for HRI/teleoperation. Deep learning has enabled 2D and monocular 3D hand pose estimation from RGB/RGB-D [28, 40], including widely used systems such as OpenPose [6] and MediaPipe [20]. Large-scale datasets (e.g., AssemblyHands) support learning under industrial hand activities [23]. Hand tracking is also central to robot teleoperation and human-to-robot hand motion mapping, with extensive surveys outlining sensing and retargeting design choices [21]. While vision methods can be accurate, performance remains setup-dependent and degrades under occlusion, poor lighting, or out-of-view motion [41]. This is especially pronounced in teleoperation and object manipulation where hands frequently self-occlude or are occluded by the object, and where camera placement may be constrained by the task. Clinical validation studies further emphasize that interpretation depends on protocol, task constraints, and measurement context [1]. These limitations motivate wearable sensing. Importantly for HRI, approaches that do not rely on cameras are advantageous in privacy-sensitive or occlusion-heavy interactions.

Multimodal tracking and visual-inertial fusion. Hybrid systems fuse inertial data with visual or magnetic cues to improve robustness [12, 16, 17]. Approaches such as MI-Poser [2] and SmartPoser [9] improve temporal consistency and coverage, but still depend on external sensing or calibration procedures that can limit scalability and generalization.

Wearable gloves and inertial tracking. Wearable gloves embed IMUs, flex sensors, capacitive, or optical elements to directly measure articulation [8, 15]. Systems such as UltraGlove [39] and MouseRing [33] demonstrate compact form factors, yet many rely on 9-axis IMUs with magnetometers that are sensitive to magnetic disturbance [13, 26, 38]. Magnetometer-free designs reduce power and improve robustness to EMI, but continuous tracking suffers from drift and inter-sensor misalignment, often requiring routine re-calibration.

[1] https://drive.google.com/file/d/1OVAY_TZ_AJblUskCt-YH4YXAc_OtTeCy/view?usp=drive_link

Despite the progress, existing inertial and hybrid solutions rarely satisfy key operational criteria *simultaneously*. These include continuous real-time operation, standalone deployment, resilience to occlusion, robustness to magnetically disturbed environments, and low-cost scalability. Recent examples include FSGlove [18] (sixteen 9-axis IMUs), which reports promising results but focuses on discrete pose sets under controlled motions rather than continuous frame-wise reconstruction, and ASTRA Glove [22] (sixteen 6-axis IMUs), which requires routine re-calibration and is not evaluated on continuous unconstrained tracking. As a result, recalibration-free, continuous hand tracking using only 6-axis IMUs across natural hand dynamics remains largely under-addressed.

To address this gap, we present a real-time, continuous hand tracking system that relies exclusively on twelve 6-axis IMUs. We eliminate magnetometers and external cameras by combining drift-insensitive features (gravity and wrist-relative gyroscope signals) with a lightweight Transformer and a late-fusion EKF. We avoid a pure 6-axis EKF formulation because yaw is unobservable and gyro biases integrate to drift; instead, a small network trained on short ($\approx$ 84 ms) windows produces a stable per-frame learned observation. The EKF treats this estimate as a measurement and fuses it with the motion model and raw inertial signals; correcting drift when ML is strong and defaulting to dynamics otherwise. This architecture is designed for standalone glove operation, maintaining continuous tracking in occlusion-heavy or magnetically noisy settings. At the same time, it can effectively enable future vision-inertial fusion methods: The inertial backbone ensures uninterrupted tracking when vision loses track, while vision can provide global translation and geometric constraints.

3 Experiment Setup

This section outlines the experimental setup used for real-time hand pose tracking, including the design of the IMU glove, the underlying hand model, and the procedure for data collection and synchronization.

3.1 Hand Model

We model the hand as an articulated kinematic chain parameterized by *wrist rotation* (3D rotation matrix), *wrist translation* (global wrist position), and *finger joint angles*. In this work we estimate only finger joint angles, holding the wrist fixed during training and evaluation to reduce complexity without sacrificing fidelity for the interaction tasks under study. For live demos we visualize an EKF-propagated wrist orientation to improve alignment; wrist pose is not trained, benchmarked, or claimed, and all reported results pertain to finger angles.

The kinematic model includes the following joints: *DIP* (distal–middle phalanges; flexion/extension), *PIP* (middle–proximal; flexion/extension), *MCP* (proximal–palm; flexion/extension, abduction/adduction), *CMC* at the thumb base (flexion/extension, abduction, opposition), and a *wrist/root* reference. Figure 2 depicts the model and functional IMU placements selected to optimize fidelity and reduce inter-sensor interference.

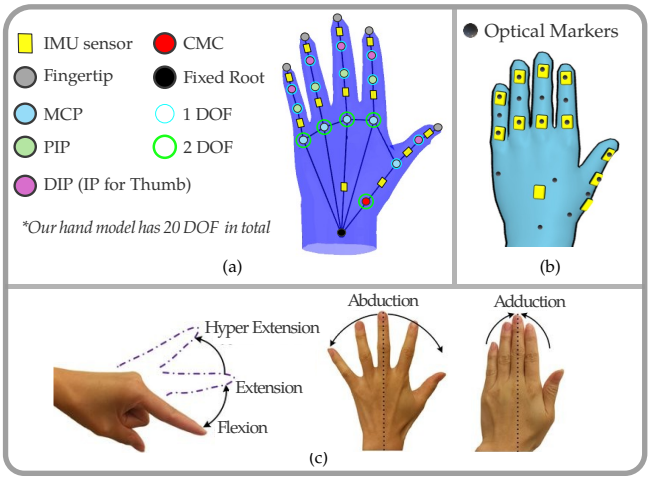

Figure 2: Hand model representation and IMU sensor placement on the glove. (a) The hand model with all joint angles and their respective degrees of freedom (DOF) is illustrated. (b) IMU sensor locations and optical marker placements are shown on the glove. (c) Different finger motion angles are depicted, including flexion/extension and abduction/adduction. Best viewed in color.

3.2 IMU Glove

The glove integrates twelve commercially-off-the-shelf 6-axis IMUs[2] (3-axis accelerometer + 3-axis gyroscope): three on the thumb (metacarpal, proximal, distal), one on the dorsum as a root reference, and two per index/middle/ring/pinky finger (proximal, distal phalanges). All sensors are hard-wired to a custom microcontroller for synchronized 200 Hz sampling and real-time streaming. A concealed wiring architecture minimizes electrical noise and maintains consistency over extended use.

3.3 Dataset

To acquire ground truth data, we utilized the OptiTrack [24] motion capture system, selected for its sub-millimeter accuracy and sub-10 ms latency. These capabilities are well-validated in prior research on fine motor control and hand tracking [3–5, 7, 11, 29, 37].

Nineteen 3 mm retro-reflective markers were affixed to the glove surface, covering key anatomical joints and segments. These placements were aligned with our hand model to ensure compatibility with inverse kinematics (IK) computations. The markers were minimally invasive and preserved natural hand motion. The OptiTrack system captured 3D trajectories of the markers at 60 Hz. These were processed with a custom IK solver to compute the full hand pose and joint angles, forming the ground truth.

IMU data, sampled at 200 Hz, was synchronized with OptiTrack data using internal timestamp alignment and resampling strategies. We evaluated two synchronization protocols: (1) 3:1 temporal pairing (IMU-to-OptiTrack), and (2) downsampling IMU data to 60 Hz. The second method, offering a 1:1 correspondence, yielded improved performance during training and inference and was thus adopted.

[2]ICM-45688 https://invensense.tdk.com/products/motion-tracking/6-axis/icm-456xy/

Data was collected from 10 participants, each performing 23 unique interaction tasks, with an average duration of approximately 11 seconds per task. In total the dataset spans approximately 12 hours of recorded motion data. Activities were selected to ensure comprehensive coverage across diverse categories: fine motor gestures such as various pinches, dynamic gestures such as swipes, symbolic communication like ASL gestures, targeted interactions such as poking, manipulative gestures, and free-form inputs.

A full description of the 23 task categories and associated data statistics is provided in Appendix 1 (supplementary material). The dataset was designed to capture a broad range of hand dynamics—from subtle postures to expressive, high-velocity gestures—enabling robust evaluation in realistic usage scenarios.

4 Methods

4.1 Overview

Our goal is to estimate root-relative hand joint angles using a wearable glove embedded with twelve 6-axis IMU sensors. Unlike traditional systems that rely on external sensors or 9-axis IMUs with magnetometers, we exclusively utilize 6-axis IMUs (accelerometers and gyroscopes).

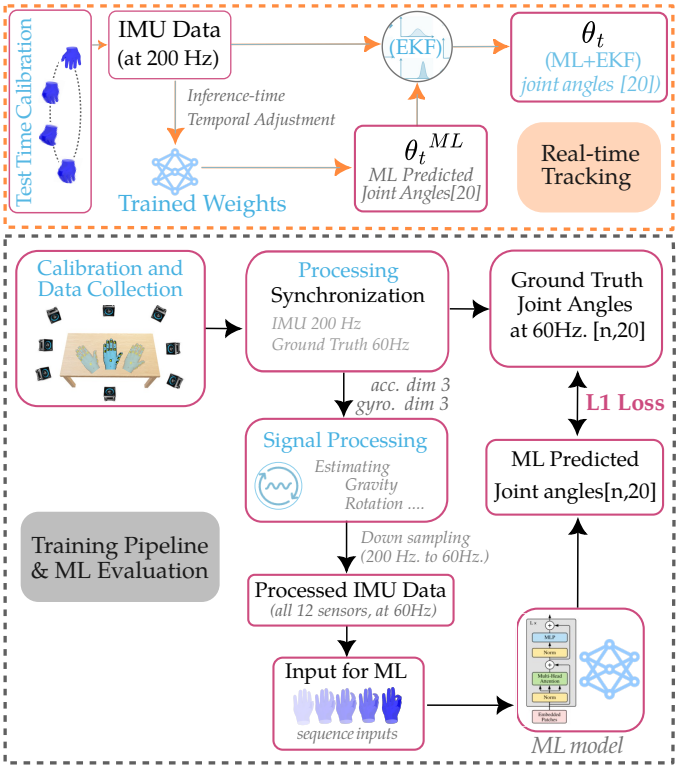

Figure 3: Overview of the proposed system architecture.

As a **baseline**, we employ an EKF-based tracker using only 6-axis IMU data. This baseline has access solely to gyroscope and accelerometer measurements, without any heading information from a magnetometer. Consequently, the tracker is highly susceptible to drift, particularly in the heading direction as discussed in Section 5. The absence of a shared heading reference across IMUs, combined with inherent sensor drift, represents the primary challenge for the baseline EKF and makes consistent, accurate pose estimation especially difficult.

We introduce a hybrid algorithm to address these challenges. Our method integrates a Machine Learning (ML) model with a model-based Extended Kalman Filter (EKF). The ML model is trained to infer finger joint angles from drift-resistant IMU features, while the EKF refines these predictions using temporal consistency and a motion model. The complete pipeline includes orientation calibration, IMU data synchronization and preprocessing, ML-based pose estimation, and EKF-based late fusion. Figure 3 illustrates the overall system architecture.

4.2 Data and Signal Processing

From the raw IMU data sampled at 200 Hz, we first perform necessary signal processing steps such as estimating gravity vectors, computing sensor orientations, and deriving additional features such as relative orientations (depending on the chosen model input configuration). Following this signal processing stage, during training we synchronize the IMU data with available ground-truth measurements.

Synchronization is critical for multi-IMU systems. Our setup includes twelve IMUs sampled at 200 Hz, while ground truth labels from the OptiTrack system are captured at 60 Hz. We explored two synchronization strategies: (1) using a 3:1 sample-to-label ratio (IMU to ground truth), and (2) directly downsampling IMU data to 60 Hz. The second approach proved superior, as it provided cleaner label alignment and eliminated the need for chunked inputs in the ML model, thereby simplifying the input pipeline. Moreover, our hybrid setup allowed us to discard a portion of the data for faster training and processing without compromising accuracy, benefiting from the stability introduced by the late-fusion stage. During downsampling, we ensured that Nyquist constraints were not violated. This approach was therefore adopted for all experiments.

After synchronization, we segmented the time-series data into fixed-size windows for model training and inference. Care was taken to avoid aliasing and to ensure that transitions between dynamic gestures were adequately captured.

4.3 Orientation Calibration

Donning a glove introduces orientation inconsistencies across glove-mounted IMUs. We therefore perform an *extrinsic* orientation calibration that aligns each sensor's local frame to a consistent anatomical hand frame, complementing (but distinct from) intrinsic calibration that compensates gyro bias and axis misalignment. As in Fig. 4(a), uncalibrated sensors may deviate from bone directions, yielding large rotational errors; after extrinsic calibration, sensors align anatomically, improving tracking fidelity and downstream processing.

Calibration operates in two modes: (i) *Train-time calibration*, applied after data collection when ground-truth orientations (e.g., from optical tracking systems such as OptiTrack) are available. This is performed offline during data preprocessing for training; and (ii) *Test-time calibration*, required for real-time deployments; performed online immediately before tracking.

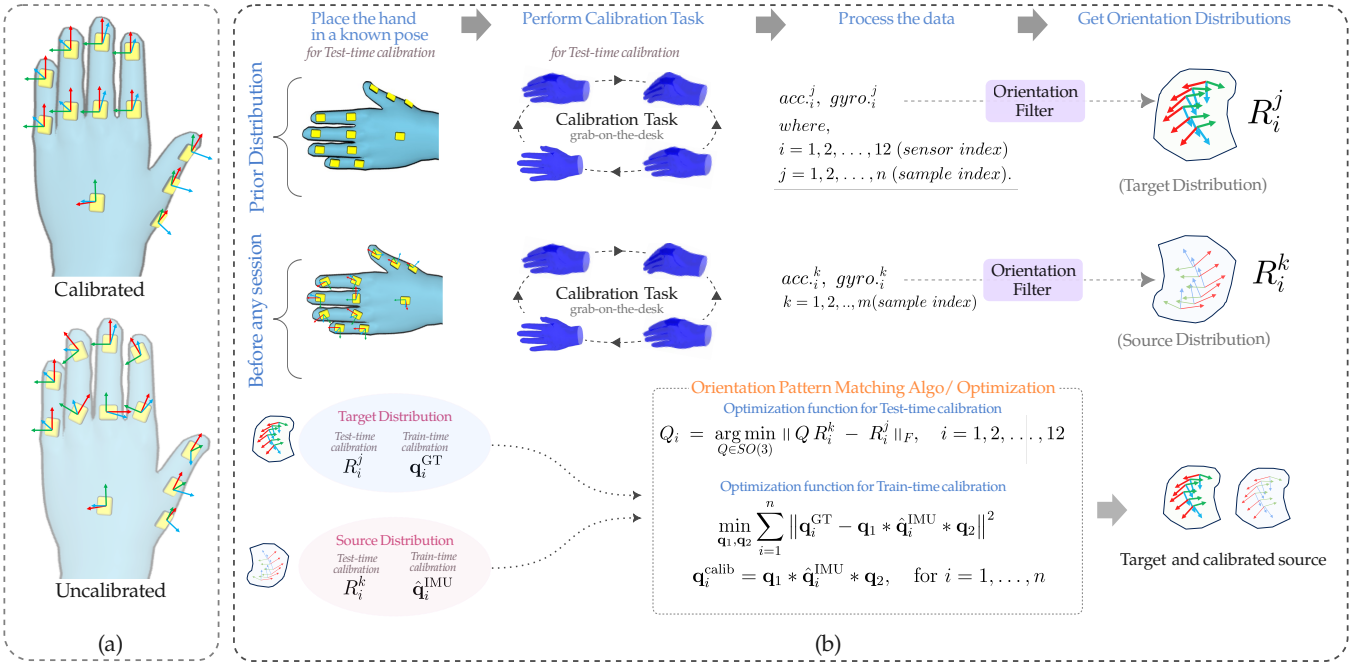

Figure 4: Orientation calibration. (a) Misalignment in uncalibrated vs. anatomically aligned gloves. (b) Calibration pipeline with pre- and train-time calibration procedures.

Test-time calibration. Before tracking, the extrinsic IMU calibration is kept short to minimize user friction, and is performed only once. We align sensor orientation distributions on $SO(3)$ by solving, for each IMU $i \in \{1, \ldots, 12\}$,

$$Q_i = \arg\min_{Q \in SO(3)} \left\| Q R_i^k - R_i^j \right\|_F, \tag{1}$$

where $R_i^k, R_i^j \in SO(3)$ are rotations sampled from the *source* and *target* distributions with indices $k = 1, \ldots, N_s$ and $j = 1, \ldots, N_t$, and $\|\cdot\|_F$ is the Frobenius norm. The target distribution is recorded once via a structured 3-minute "grab-on-the-desk" task that promotes (1) motion repeatability and (2) sensor-trajectory diversity; at deployment, users perform a 10-second version to form the source distribution. The optimal Q_i is then pre-multiplied to all IMU orientations, aligning source to target in a consistent reference frame. While Eq. 1 admits classical solvers (e.g., BFGS), we adopt the dedicated, outlier-robust method of [31, 32], which does not require one-to-one correspondences ($N_s \neq N_t$).

Train-time calibration. When ground-truth is available, we estimate a global alignment from a short window (typically 10-13 seconds) at the start of each session. Let $\mathbf{q}_i^{\mathrm{GT}}$ denote ground-truth quaternions and $\hat{\mathbf{q}}_i^{\mathrm{IMU}}$ the IMU estimates. We solve for unit quaternions $\mathbf{q}_1, \mathbf{q}_2 \in \mathbb{H}$, where $\mathbb{H}$ denotes the quaternion algebra, and the operator $\otimes$ represents quaternion multiplication

$$\min_{\mathbf{q}_1, \mathbf{q}_2} \sum_{i=1}^{n} \left\| \mathbf{q}_i^{\mathrm{GT}} - \mathbf{q}_1 \otimes \hat{\mathbf{q}}_i^{\mathrm{IMU}} \otimes \mathbf{q}_2 \right\|^2, \tag{2}$$

and apply

$$\mathbf{q}_i^{\mathrm{calib}} = \mathbf{q}_1 \otimes \hat{\mathbf{q}}_i^{\mathrm{IMU}} \otimes \mathbf{q}_2, \quad i = 1, \ldots, n. \tag{3}$$

This ensures that each session begins from a globally aligned sensor frame (Figure 4(a), "Calibrated"), thereby enabling accurate inter-session comparisons and enhancing model generalizability.

4.4 Model Architecture

After data synchronization, downsampling, and orientation calibration, the processed signals are prepared for input into our machine learning model. Designing an effective ML-based or hybrid real-time hand tracking system using IMUs involves two key considerations: (1) selecting an appropriate model architecture, and (2) choosing an effective representation of the input signals.

We only focused on transformer architecture ablation, instead of analyzing different ML model options (e.g. LSTM). We prioritized three core criteria: (1) *accuracy*, to minimize mean joint angle error, (2) *robustness*, to ensure stable performance over time, and (3) *speed*, to ensure real-time inference.

The ML component is designed to predict 20 joint angles from a short temporal window of IMU data and has $\approx 1M$ parameters. We experimented with various input features including raw accelerometer/gyroscope signals, quaternions, rotation matrices, and gravity vectors. Our ablation studies (detailed in the Results section) revealed that gravity vectors and wrist-relative gyroscope signals offered the best trade-off among accuracy, robustness, and speed.

Gravity vectors are estimated using a complementary filter, which is efficient and well-suited for real-time deployment. Gyroscope signals are bias-corrected using a static average baseline computed at the beginning of each session. To reduce dimensionality and introduce invariance, gyroscope values from each IMU are made relative to the root (wrist) IMU.

Each input frame consists of 69 values: 12×3 gravity features and 11 × 3 root-relative gyroscope features. These are fed into a Transformer encoder comprising two attention blocks and a feedforward layer (hidden dimension = 512). We evaluated input window lengths of 5, 10, 15, and 30 frames at 60 Hz. A 5-frame window (∼ 83.3 ms) offered a good balance between responsiveness and accuracy, and was selected for the final model.

Sinusoidal positional encoding is used to preserve temporal information, and dropout ($p = 0.1$) promotes generalization. The model outputs 20 joint angles corresponding to the last frame of the input window, thus maintaining minimal processing time for real-time applications. We do not impose explicit joint-limit constraints during training; the model learns feasible ranges from data. At inference, predictions can be post-processed by clipping to predefined anatomical limits when they fall outside those ranges.

4.5 EKF Late Fusion

Although the ML model yields accurate per-frame predictions, it lacks temporal regularization and can produce jitter during transitions or when exposed to noisy input features. To address this, we employ a late-stage Extended Kalman Filter (EKF) that treats ML predictions as noisy observations.

The EKF maintains a state vector of joint angles and uses a hand motion model derived from human kinematics to propagate state over time. Observations include both ML-predicted joint angles and raw IMU signals, allowing the EKF to refine outputs based on physical plausibility and continuity.

This fusion significantly smooths prediction trajectories, eliminates outlier spikes caused by ML overfitting, and maintains robust performance across varied hand motions and tasks. As shown in our evaluation, the hybrid model consistently outperforms both standalone ML and EKF baselines in both accuracy and stability.

5 Results and Discussion

We trained our learning-based model to minimize the mean absolute joint angle error (MAE), which serves as our primary performance metric. Multiple input configurations were evaluated to determine the optimal representation for the ML component of our hybrid algorithm.

In addition to MAE, we report the Mean Key Point Error Transform (MKPET) as a secondary metric to assess the accuracy of reconstructed hand joint positions in 3D space. MKPET is computed by converting predicted joint angles into 3D hand landmarks using a forward kinematics solver adapted from the EMG2Pose framework [30]. This metric measures the wrist-relative Euclidean distance between predicted and ground-truth positions for 21 hand landmarks, expressed in millimeters.

5.1 Ablation Experiments

We systematically ablate IMU feature representations to measure accuracy, drift, and computational performance of our Transformer-based hand pose estimation model. Hidden widths are scaled with input dimensionality for a fair compute budget. The reported latency is only for model inference; real-time deployment adds windowing, preprocessing, buffering overhead.

Feature configurations (nomenclature). *Raw accel/gyro*: unprocessed 3-axis accelerometer and gyroscope; *GT_quat*: ground-truth joint quaternions (ideal); *Sensor_quat*: quaternions from 6-axis IMU integration; *Wrist_Rel_*: signals expressed relative to the wrist IMU orientation; *Calib_*: offline calibration for initial misalignment and sensor-to-segment transform; *Sensor_Gravity* and *Sensor_calib_gyro*: gravity vectors and calibrated gyros. For each we report *Input Dim* (per-frame length), *Train/Val/Test MAE* (°), *Drift/Jitter status* (qualitative), and *Mean/STD inference time* (ms).

Upper bound and raw signals. Rows 2–3 of Table 1 (GT_quat) give an idealized lower bound with the lowest test error (2.92° MAE). Training on raw accelerometer+gyroscope (72D) yields 12.92° MAE with pronounced jitter, indicating raw signals are insufficient for learning pose.

Estimated quaternions and variants. `Sensor_quat` improves to 11.13° MAE but drifts due to unobservable yaw in magnetometer-free systems. Wrist-relative and calibrated variants reduce error marginally yet remain vulnerable to cumulative orientation bias. Mixing clean training inputs (e.g., `GT_quat`) with noisy test signals (e.g., `Sensor_quat`) produces large generalization gaps, underscoring the need to align train and deployment domains.

Drift-insensitive hybrids. Hybrid inputs combining gravity vectors with wrist-relative angular velocities improve temporal stability. Notably, `Sensor_Gravity` + `Sensor_Wrist_relative_gyro` achieves 10.56° MAE with fast inference (2 ms/frame), suitable for real-time interaction. Higher-dimensional inputs (e.g., 84D) increase processing time to ∼4 ms/frame, reflecting a robustness–throughput trade-off on embedded devices.

Takeaways. Quaternion features offer superior *static* accuracy but typically degrade after 8–10 s without recalibration. Gravity + gyro features trade a small loss in precision for stronger *temporal* robustness. Increasing model complexity or sequence length yields diminishing accuracy gains while significantly increasing inference time. Given these trade-offs, we adopt calibrated, wrist-relative angular velocities and gravity vectors as the default input for fast, robust hand-pose estimation in real-world settings. All subsequent ML results use this configuration.

5.2 Quantitative and Visual Results

We summarize (Fig. 5 top) the comparative results across three systems: Traditional EKF-based tracking, our proposed learning-based (ML) model, and our hybrid ML+EKF pipeline. The ML model alone achieves a substantial improvement over the EKF baseline, reducing MAE from 15.91° to 10.56°. When integrating our ML predictions into an EKF post-processing stage (ML+EKF), the performance improves further to 9.77°, representing a 38.6% improvement over the baseline.

Similarly, MKPET drops from 17.08 mm in the EKF-only configuration to 13.47 mm with ML, and further to 11.67 mm in the ML+EKF variant—a total reduction of 31.7%.

Figure 6 presents a qualitative evaluation of selected hand poses using different methods, alongside the ground truth. The ML+EKF approach better preserves overall pose accuracy compared to EKF-only and ML-only implementations.

Table 1: Ablation study evaluating different input signal combinations for ML-only hand pose estimation using a 12-IMU glove.

Training & Validation Inputs	Test Inputs	Input Dim	Train MAE (°) ↓	Val MAE (°) ↓	Test MAE (°) ↓	Mean Inf. Time (ms) ↓	STD Inf. Time (ms) ↓	Drift / Jitter
Raw Acce, Raw gyro	Raw Acce, Raw gyro	72	9.84	10.62	12.92	2	1	High Jitter
GT_quat	GT_quat	48	3.03	4.75	3.56	2	1	–
Wrist_Rel_GT_quat	Wrist_Rel_GT_quat	44	1.92	3.20	2.92	2	1	–
Sensor_quat	Sensor_quat	48	6.41	10.25	11.13	2	1	High Drift
Wrist_Rel_Sensor_quat	Wrist_Rel_Sensor_quat	44	6.47	10.82	13.13	2	1	High Drift
Calib_Sensor_quat	Calib_Sensor_quat	48	7.44	10.19	10.05	2	1	High Drift
Calib_Wrist_Rel_Sensor_quat	Calib_Wrist_Rel_Sensor_quat	44	6.87	8.85	7.30	2	1	High Drift
Wrist_Rel_GT_quat	Calib_Wrist_Rel_Sensor_quat	44	1.92	3.32	8.84	2	1	High Drift
GT_quat	Calib_Sensor_quat	48	3.83	4.87	9.52	2	1	High Drift
GT_quat	Sensor_quat	48	3.83	4.87	18.98	2	1	High Drift
GT_quat, Sensor_Acc	Calib_Sensor_quat, Sensor_Acc	84	3.03	4.72	9.10	4	2	High Drift
Wrist_Rel_GT_quat, Wrist_Rel_Sensor_Acc	Calib_Wrist_Rel_Sensor_quat, Wrist_Rel_Sensor_Acc	77	1.54	3.20	10.07	3	2	High Drift
Calib_Wrist_Rel_Sensor_quat, Wrist_Rel_Sensor_Acc	Calib_Wrist_Rel_Sensor_quat, Sensor_Wrist_Rel_Acc	77	6.30	9.22	9.73	3	2	High Drift
Sensor_Gravity	Sensor_Gravity	36	10.81	10.75	12.13	1	0	High Jitter, Low Drift
Sensor_Gravity + Sensor_calib_gyro	Sensor_Gravity + Sensor_calib_gyro	72	9.52	10.77	10.82	2	1	Medium Jitter, Low Drift
Sensor_Gravity + Sensor_Wrist_relative_gyro	Sensor_Gravity + Sensor_Wrist_relative_gyro	69	9.43	10.31	10.56	2	1	Medium Jitter, Low Drift

5.3 Joint-Wise Error Trends

Figure 5 shows per-joint absolute error distributions for EKF (baseline), ML, and ML+EKF. Across most joints, ML substantially reduces error relative to EKF; late fusion (ML+EKF) yields further gains by combining learned priors with sequential filtering. The largest improvements occur on the thumb—`thumb_CMC_flex`, `thumb_CMC_abad`, `thumb_IP`—with more than 50% reduction vs. EKF; these joints are difficult for purely kinematic filters due to non-rigid motion and muscle cross-talk, favoring learned inference. A single exception is `thumb_MCP`, where ML+EKF slightly underperforms ML, likely because the EKF state diverged from the ML estimate and late fusion biased toward the EKF trajectory under high inter-model variance. Beyond the thumb, MCP/PIP/DIP joints across index–pinky fingers show consistent gains: ML > EKF, and ML+EKF further suppresses variance and tightens the median error.

6 Real-time Tracking

We demonstrate the real-time capabilities of our algorithm in a live tracking setup. Although the average session length in the training dataset was approximately 12 seconds, our algorithm maintained robust performance and never drifted off track in real-time evaluation of 15+ minutes, without re-initialization or re-calibration.

For real-time tracking, IMU data was processed using our trained model and visualized using the ReRun[3] visualizer. One critical implementation detail is that while the model was trained with IMU signals sampled at 60 Hz, the live tracking system operated at 200 Hz. As shown in Figure 3, we performed frame rate adjustment before passing the input to the ML model to ensure temporal compatibility and consistent inference behavior.

We demonstrate multiple short clips of diverse object manipulation tasks, highlighting the robustness and practical use cases of our algorithm.

Overall, the real-time tracking performance closely aligns with our quantitative evaluation. The system exhibits accurate, stable pose estimation in dynamic interactions, confirming the practical utility of the proposed framework.

7 Limitations

While the proposed ML+EKF framework demonstrates a viable pathway for real-time hand tracking using only 6-axis IMUs, and overcomes several inherent limitations of such sensor configurations, it still presents certain constraints.

The system (ML+EKF) achieves a mean joint angle error of under 10°—comparable to several vision-based methods such as [34,

[3] https://rerun.io/

Overall Performance

		EKF (Baseline)	ML	ML+EKF
Mean Absolute Joint Angle Error (MAE):	[Degrees]	15.91	10.56	9.77
Mean Key Point Error Transform (MKPET):	[mm]	17.08	13.47	11.67

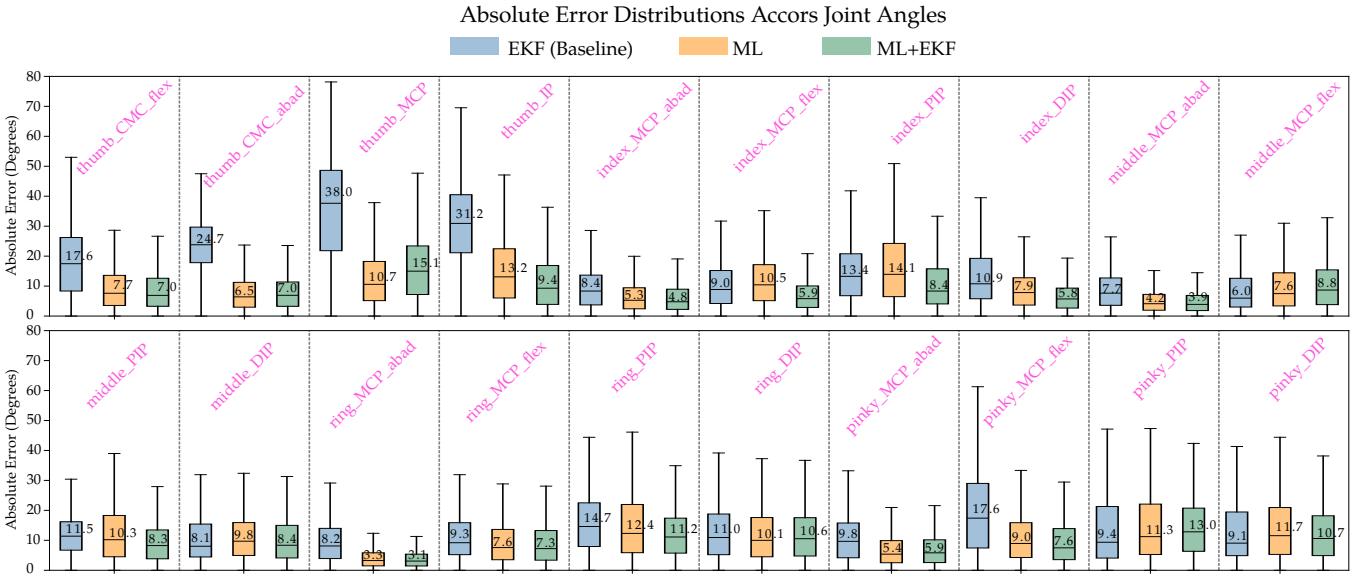

Figure 5: Box plots showing absolute angular error (in degrees) across 20 joint angles for three tracking models: EKF baseline, ML-only, and the proposed ML+EKF hybrid.

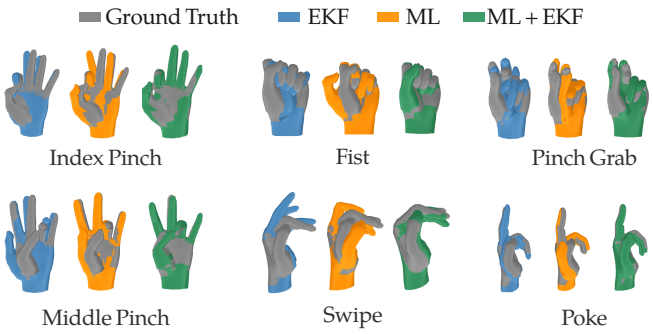

Figure 6: Qualitative pose results. Best viewed in color.

35]—yet its precision is currently insufficient for ultra fine-grained manipulation or precision grasping tasks. Specifically, we observed a fingertip-level Mean Key Point Error Transform (MKPET) of approximately 16 mm. Potential avenues to close the gap include scaling the training dataset, increasing model capacity, or fusing RGB/RGB-D cues with our inertial pipeline (CV+IMU).

Furthermore, the current method only tracks the skeletal pose of the hand (joint angles relative to the wrist). It does not estimate the global hand root position. This limitation can be mitigated by adopting inside-out tracking modules such as the HTC VIVE Ultimate Tracker.

8 Conclusion

We presented the first real-time hand pose tracking system that operates exclusively on 6-axis IMU data, without relying on magnetometers or other external sensors. Our 6-axis method maintains low user friction during live tracking by requiring a brief one-time calibration procedure. Our hybrid framework combines a lightweight Transformer model with an Extended Kalman Filter (EKF) for temporal smoothing, enabling robust and stable finger joint angle estimation even during lengthy, dynamic, real-world human-object interaction (HOI) scenarios.

Through extensive experiments including ablation studies, long live tracking sessions, and activity-wise evaluations, we validated the system's accuracy, temporal consistency, and real-time viability. Our ablation analysis highlights how the choice of input representations, particularly drift-insensitive features like gravity vectors and wrist-relative gyroscope signals, plays a critical role in achieving both precision and robustness.

The system achieves sub-10° mean joint angle error, comparable to existing egocentric vision-based methods, while avoiding their limitations related to occlusion, lighting, and environmental constraints. Overall, our method offers a compelling balance of performance, scalability, and portability for a wide range of robotics, HRI, AR/VR, and HCI applications.

References

[1] Gianluca Amprimo, Giulia Masi, Giuseppe Pettiti, Gabriella Olmo, Lorenzo Priano, and Claudia Ferraris. 2024. Hand tracking for clinical applications: Validation of the Google MediaPipe Hand (GMH) and the depth-enhanced GMH-D frameworks. *Biomedical Signal Processing and Control* 96 (2024), 106508. doi:10.1016/j.bspc.2024.106508

[2] Riku Arakawa, Bing Zhou, Gurunandan Krishnan, Mayank Goel, and Shree K Nayar. 2023. MI-Poser: Human Body Pose Tracking Using Magnetic and Inertial Sensor Fusion with Metal Interference Mitigation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–24. doi:10.1145/3610891

[3] Alireza Bilesan, Shunsuke Komizunai, Teppei Tsujita, and Atsushi Konno. 2021. Improved 3D Human Motion Capture Using Kinect Skeleton and Depth Sensor. *Journal of Robotics and Mechatronics* 33, 6 (2021), 1408–1422. doi:10.20965/jrm.2021.p1408

[4] A. Cannavo, F. G. Prattico, A. Bruno, and F. Lamberti. 2023. AR-MoCap: Using Augmented Reality to Support Motion Capture Acting. *Proceedings of the 2023 IEEE Conference on Virtual Reality and 3D User Interfaces* 2023 (2023), 318–327. doi:10.1109/VR55154.2023.00047

[5] Alberto Cannavò, Filippo Gabriele Pratticò, Alberto Bruno, and Fabrizio Lamberti. 2023. AR-MoCap: Using Augmented Reality to Support Motion Capture Acting. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. 318–327. doi:10.1109/VR55154.2023.00047

[6] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence* 43, 1 (2019), 172–186. doi:10.1109/TPAMI.2019.2929257

[7] B Castillo, C Riascos, JM Franco, J Marulanda, and P Thomson. 2024. Assessing Spatiotemporal Behavior of Human Gait: A Comparative Study Between Low-Cost Smartphone-Based Mocap and OptiTrack Systems. *Experimental Techniques* (2024), 1–11. doi:10.1007/s40799-024-00716-x

[8] J. Connolly, J. Condell, B. O'Flynn, J. T. Sanchez, and P. Gardiner. 2018. IMU Sensor-Based Electronic Goniometric Glove for Clinical Finger Movement Analysis. *IEEE Sensors Journal* 18, 3 (2018), 1273–1281. doi:10.1109/JSEN.2017.2776262

[9] Nathan DeVrio, Vimal Mollyn, and Chris Harrison. 2023. Smartposer: Arm pose estimation with a smartphone and smartwatch using uwb and imu data. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–11. doi:10.1145/3586183.3606821

[10] Shengshun Duan, Fangzhi Zhao, Huiying Yang, Jianlong Hong, Qiongfeng Shi, Wei Lei, and Jun Wu. 2023. A pathway into metaverse: Gesture recognition enabled by wearable resistive sensors. *Advanced Sensor Research* 2, 8 (2023), 2200054. doi:10.1002/adsr.202200054

[11] Joshua Savio Furtado, Hugh Hong-Tao Liu, Gilbert Lai, Hervé Lacheray, and Jason Desouza-Coelho. 2018. Comparative Analysis of OptiTrack Motion Capture Systems. *Lecture Notes in Mechanical Engineering* (2018). doi:10.1007/978-3-030-17369-2_2

[12] Jinuk Heo, Hyelim Choi, Yongseok Lee, Hyunsu Kim, Harim Ji, Hyunreal Park, Youngseon Lee, Cheongkee Jung, Hai-Nguyen Nguyen, and Dongjun Lee. 2024. Hand Tracking: Survey. *International Journal of Control, Automation and Systems* 22, 6 (2024), 1761–1778. doi:10.1007/s12555-024-0298-1

[13] Pei-Chi Hsiao, Shu-Yu Yang, Bor-Shing Lin, I-Jung Lee, and W. Chou. 2015. Data glove embedded with 9-axis IMU and force sensing sensors for evaluation of hand function. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 4631–4634. doi:10.1109/EMBC.2015.7319426

[14] Xinxin Huang, Yunan Xue, Shuyun Ren, and Fei Wang. 2023. Sensor-based wearable systems for monitoring human motion and posture: A review. *Sensors* 23, 22 (2023), 9047. doi:10.3390/s23229047

[15] C. K. Jha, K. Gajapure, and A. L. Chakraborty. 2021. Design and Evaluation of an FBG Sensor-Based Glove to Simultaneously Monitor Flexure of Ten Finger Joints. *IEEE Sensors Journal* 21, 6 (2021), 7620–7630. doi:10.1109/JSEN.2020.3046521

[16] Yongseok Lee, Wonkyung Do, Hanbyeol Yoon, Jinuk Heo, WonHa Lee, and Dongjun Lee. 2021. Visual-inertial hand motion tracking with robustness against occlusion, interference, and contact. *Science Robotics* 6, 58 (2021), eabe1315. doi:10.1126/scirobotics.abe1315

[17] Tong Li and Haoyong Yu. 2023. Visual–inertial fusion-based human pose estimation: A review. *IEEE Transactions on Instrumentation and Measurement* 72 (2023), 1–16. doi:10.1109/TIM.2023.3286000

[18] Yutong Li, Wenqiang Xu, Jieyi Zhang, Tutian Tang, and Cewu Lu. 2025. FSGlove: An Inertial-Based Hand Tracking System with Shape-Aware Calibration. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. doi:10.48550/arXiv.2509.21242

[19] Bor-Shing Lin, Pei-Chi Hsiao, Shu-Yu Yang, Che-Shih Su, and I-Jung Lee. 2017. Data glove system embedded with inertial measurement units for hand function evaluation in stroke patients. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25, 11 (2017), 2204–2213. doi:10.1109/tnsre.2017.2720727

[20] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019). doi:10.48550/arXiv.1906.08172

[21] Roberto Meattini, Raúl Suárez, Gianluca Palli, and Claudio Melchiorri. 2023. Human to Robot Hand Motion Mapping Methods: Review and Classification. *IEEE Transactions on Robotics* 39, 2 (2023), 842–861. doi:10.1109/TRO.2022.3205510

[22] Ali Nikkhah Bahrami, Mohammad Reza Nayeri, Reza Almasi Ghaleh, and Behzad Moshiri. 2025. ASTRA Glove: A Wearable Tracking Device for "Accurate Sensing and Tracking of Realtime Articulations". *IEEE Sensors Journal* 25, 5 (2025), 8631–8644. doi:10.1109/JSEN.2025.3528308

[23] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. 2023. Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12999–13008. doi:10.48550/arXiv.2304.12301

[24] OptiTrack. 2023. OptiTrack IR active camera network tracking system. https://optitrack.com/.

[25] Mingzhang Pan, Yingzhe Tang, and Hongqi Li. 2023. State-of-the-art in data gloves: A review of hardware, algorithms, and applications. *IEEE Transactions on Instrumentation and Measurement* 72 (2023), 1–15. doi:10.1109/TIM.2023.3243614

[26] Yeongyu Park, Jeongsoo Lee, and Joonbum Bae. 2014. Development of a finger motion measurement system using linear potentiometers. In *2014 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*. 125–130. doi:10.1109/AIM.2014.6878066

[27] Kyung Rok Pyun, Kangkyu Kwon, Myung Jin Yoo, Kyun Kyu Kim, Dohyeon Gong, Woon-Hong Yeo, Seungyong Han, and Seung Hwan Ko. 2024. Machine-learned wearable sensors for real-time hand-motion recognition: toward practical applications. *National Science Review* 11, 2 (2024), nwad298. doi:10.1093/nsr/nwad298

[28] Jing Qi, Li Ma, Zhenchao Cui, and Yushu Yu. 2024. Computer vision-based hand gesture recognition for human-robot interaction: a review. *Complex & Intelligent Systems* 10, 1 (2024), 1581–1606. doi:10.1007/s40747-023-01173-6

[29] Marco Rosa-Clot and Giuseppe Marco Tina. 2020. Chapter 7 - Tracking Systems. In *Floating PV Plants*, Marco Rosa-Clot and Giuseppe Marco Tina (Eds.). Academic Press, 79–87. doi:10.1016/B978-0-12-817061-8.00007-5

[30] Sasha Salter, Richard Warren, Collin Schlager, Adrian Spurr, Shangchen Han, Rohin Bhasin, Yujun Cai, Peter Walkington, Anuoluwapo Bolarinwa, Robert Wang, et al. [n. d.]. emg2pose: A Large and Diverse Benchmark for Surface Electromyographic Hand Pose Estimation. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. doi:10.48550/arXiv.2412.02725

[31] Anik Sarker and Alan T Asbeck. 2025. Correspondence-Free Fast and Robust Spherical Point Pattern Registration. (2025), 28156–28166. doi:10.48550/arXiv.2508.02339

[32] Anik Sarker and Alan T Asbeck. 2025. Fast, Robust, Permutation-and-Sign Invariant SO (3) Pattern Alignment. *arXiv preprint arXiv:2512.00659* (2025). doi:10.48550/arXiv.2512.00659

[33] Xiyuan Shen, Chun Yu, Xutong Wang, Chen Liang, Haozhan Chen, and Yuanchun Shi. 2024. MouseRing: Always-available Touchpad Interaction with IMU Rings. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19. doi:10.1145/3613904.3642225

[34] Ayan Sinha, Chiho Choi, and Karthik Ramani. 2016. Deephand: Robust hand pose estimation by completing a matrix imputed with deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4150–4158. doi:10.1109/CVPR.2016.450

[35] Srinath Sridhar, Franziska Mueller, Antti Oulasvirta, and Christian Theobalt. 2015. Fast and Robust Hand Tracking Using Detection-Guided Optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3213–3221. doi:10.1109/CVPR.2015.7298941

[36] Rayane Tchantchane, Hao Zhou, Shen Zhang, and Gursel Alici. 2023. A review of hand gesture recognition systems based on noninvasive wearable sensors. *Advanced intelligent systems* 5, 10 (2023), 2300207. doi:10.1002/aisy.202300207

[37] Yiwei Wu, Kuan Tao, Qi Chen, Yinsheng Tian, and Lixin Sun. 2022. A Comprehensive Analysis of the Validity and Reliability of the Perception Neuron Studio for Upper-Body Motion Capture. *Sensors* 22, 18 (2022). doi:10.3390/s22186954

[38] H. Yamaura, K. Matsushita, R. Kato, and H. Yokoi. 2009. Development of Hand Rehabilitation System for Paralysis Patient – Universal Design Using Wire-Driven Mechanism. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Minneapolis, MN, USA, 7122–7125. doi:10.1109/IEMBS.2009.5332885

[39] Qiang Zhang, Yuanqiao Lin, Yubin Lin, and Szymon Rusinkiewicz. 2023. Hand Pose Estimation with Mems-Ultrasonic Sensors. , Article 79 (2023), 11 pages. doi:10.1145/3610548.3618202

[40] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. 2023. Deep learning-based human pose estimation: A survey. *Comput. Surveys* 56, 1 (2023), 1–37. doi:10.48550/arXiv.2012.13392

[41] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. 2023. Deep learning-based human pose estimation: A survey. *Comput. Surveys* 56, 1 (2023), 1–37. doi:10.48550/arXiv.2012.13392

Received 2025-09-30; accepted 2025-12-01