Inertial Pose-Based Human Activity Recognition From Sparse On-Body IMUs

Ricarda H. Link[0009-0003-2748-4121] and Heiner Stuckenschmidt[0000-0002-0209-3859]

University of Mannheim, Mannheim, Germany
{ricarda.link,heiner.stuckenschmidt}@uni-mannheim.de

Abstract. Inertial measurement units (IMUs) are widely used for human activity recognition (HAR), typically by mapping raw acceleration and angular velocity signals directly to activity labels. A distinct class of models — inertial posers — uses the same sensors to reconstruct full-body 3D pose sequences via a parametric body model, yielding a structured, biomechanically interpretable representation. Yet the potential of such representations for HAR remains largely unexplored. In this work, we investigate how inertial posers can be integrated into a HAR pipeline and what benefits a body-model-based representation might offer for activity recognition. We (1) empirically compare three approaches for a HAR task on the TotalCapture dataset: a purely IMU-based model, a purely pose-based model and a hybrid model combining both representations. Poses are expressed via the SMPL (Skinned Multi-Person Linear Model) body model, estimated by a sparse inertial poser that uses only IMU data. On TotalCapture, pose-based HAR surpasses the IMU-only baseline, although pose is estimated from the same IMU signals only (macro-F1: 0.562 vs. 0.546), and combining both representations further improves performance to 0.574. With ground-truth pose, performance rises further to 0.587 for the best pose-only model, confirming that gains in inertial pose estimation quality translate directly into HAR improvements. An oracle ceiling of 0.660 confirms that the two modalities carry substantial complementary information, suggesting that inertial-poser-based representations are a promising direction for activity recognition. We (2) discuss conceptual advantages of incorporating a structured pose representation.

Keywords: Human activity recognition · Inertial sensors · Body model · Inertial poser · SMPL

1 Introduction

Human activity recognition (HAR) based on inertial sensors has been extensively studied [3]. Inertial measurement units (IMUs) are typically placed at one or several on-body locations to capture motion dynamics and may be complemented by other modalities to improve recognition accuracy [18]. Such systems aim

to infer human activities from sensor data, ranging from everyday actions and predefined movement sequences to fine-grained movements.

Recently, inertial posers have emerged as models that reconstruct full-body poses from sparse on-body IMU recordings by mapping them onto the SMPL body model [16]. Instead of recognizing activities, they aim to recover the underlying body configuration from inertial signals alone. These models provide a structured representation of human body motion and have found applications in motion analysis, healthcare, sports, and virtual reality, where camera-free and occlusion-robust motion capture is particularly beneficial [21].

While the pose estimation capabilities of these models have been shown to improve, the potential of inertial posers for HAR has not yet been explored. By translating IMU signals into interpretable pose representations, they may provide a structured intermediate space that improves recognition accuracy, robustness, and generalizability. This motivates our investigation of how projecting IMU data from five on-body locations onto a body model can benefit the HAR task.

Our goal is to evaluate empirically how incorporating a body model representation can enhance inertial-based HAR performance and to identify what other conceptual advantages this approach may offer. We conduct experiments on the TotalCapture dataset [22], which comprises four activity classes. We hypothesize that the structured prior introduced by the body model complements raw IMU signals in several ways. By projecting sparse sensor readings onto a parametric body model, anatomical constraints are enforced and the kinematic relationships between body segments are made explicit. Furthermore, whereas individual IMU readings are locally anchored to a single sensor location, the body model aggregates information across all sensor locations into a globally consistent, full-body representation, contextualizing each sensor's contribution relative to the others. The resulting pose sequence, expressed in terms of joint rotations and positions, may thereby expose activity-discriminative patterns more explicitly than raw acceleration and orientation signals alone.

The remainder of this paper is organized as follows. Section 2 reviews the relevant background and related work, while Section 3 presents the proposed methodology and experimental setup. Section 4 reports and discusses the experimental results. Section 5 explores additional conceptual advantages of inertial-poser-based HAR. Section 6 highlights the limitations of the study and suggests directions for future research, and Section 7 concludes the paper.

2 Background and Related Work

In this section, we review common practices in inertial-based HAR and recent developments in inertial posers. We further discuss pose-based HAR methods from the vision domain, which become directly applicable to IMU data once an inertial poser is introduced as an intermediate representation.

2.1 Inertial-Based HAR

Two main approach families are exist in inertial-based HAR. The first group consists of feature-based methods, which derive statistical and frequency-domain features from the inertial signals and feed them into classical machine learning models to recognize predefined activities [4]. The second group comprises deep learning methods, where inertial signals are processed directly by sequential architectures such as LSTMs or CNNs, removing the need for handcrafted feature extraction [19, 29].

While these approaches perform well for activity classification, they offer no insight into the spatial structure of the underlying human motion in 3D space. Furthermore, since raw IMU signals are locally anchored to individual sensor locations, the spatial relationships between body segments remain implicit and unrepresented, potentially limiting recognition performance for activities that are best distinguished by their global body configuration rather than local motion patterns alone.

2.2 SMPL Model and Inertial Poser

Combining IMU recordings from multiple on-body locations with a parametric body model such as SMPL enables the reconstruction and analysis of human movement in 3D space.

The SMPL body model [16] represents the human body as a parametric mesh of 6,890 vertices built around a 24-joint skeleton, where every three neighboring vertices form a triangular face, together approximating a realistic body surface, referred to as a mesh. The model is governed by two sets of parameters. The pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{72}$ encode the orientation of each of the 24 joints in three-dimensional axis-angle form, arranged in a kinematic tree, yielding $24 \times 3 = 72$ pose parameters in total. Concretely, $\boldsymbol{\theta}_{0:2}$ encodes the root joint orientation, and $\boldsymbol{\theta}_{3:5}$ encodes a first child joint to the root joint in the SMPL kinematic tree. When $\boldsymbol{\theta} = \mathbf{0}$, the body assumes the canonical T-pose, an upright standing position with arms extended horizontally and palms facing downward. Since joints are arranged in a kinematic tree, a non-zero rotation at any joint propagates to all descendant joints, while ancestor joints remain unaffected. The shape parameters $\boldsymbol{\beta}$ capture subject-specific body characteristics such as height and body composition. The model is available in male, female, and gender-neutral variants.

The idea of estimating full-body pose from sparse on-body IMUs was first realized by von Marcard et al. [23], who introduced six inertial sensors attached to fixed body locations to reconstruct full-body pose sequences. Unlike dense motion capture suits such as the 17-sensor Xsens Awinda system [20], inertial posers rely on a sparse sensor configuration to estimate the full-body pose sequence. Global sensor orientation is first derived from IMU data, typically comprising accelerometer, gyroscope, and magnetometer readings, using fusion algorithms such as Extended Kalman Filters [6], which are often implemented directly within the sensor hardware. These orientations are then aligned with

their corresponding body segments and, combined with gravity-corrected accelerometer data, integrated over time to produce a pose sequence.

Since von Marcard et al.'s initial work [23], numerous studies have extended and improved inertial posers [7, 27, 8, 24], with some of the models also being able to estimate the subject's global translation (i.e., the subject's movement through space) in addition to the local pose. Ground truth poses for the inertial posers training and evaluation are obtained from optical motion capture systems or full-body suits with dense sensor configurations.

2.3 Inertial-Poser-Based HAR

HAR based on pose sequences obtained from inertial posers remains generally largely unexplored despite some related work [12, 28]. While pose-based recognition has been widely studied in vision-based HAR, its application to inertial-based settings remains largely unexplored, despite the fact that vision-based pose recognition pipelines transfer directly to the inertial poser setting, as both operate on the same skeletal representation.

Early works have begun to explore this direction, such as Reiss et al. [30], who combined signal-oriented and body model-based features derived from joint angles for activity recognition, and more recent approaches like Konak et al. [12], who employ an inertial poser to construct 2D heatmaps from SMPL pose sequences for activity recognition. Zhou et al. [28] use skeletal data derived from an inertial poser as input to a downstream skeleton-based recognition model but provide no insights into the underlying motivation for this choice. These initial studies motivate our investigation into the benefits of incorporating body model representations in inertial-based HAR. Once a pose sequence is obtained, one can directly apply vision-based HAR models that operate on 2D or 3D skeletal data [15, 5], which have shown strong performance and benefited from advances in deep learning-based pose estimation from video data. Crucially, since inertial posers produce skeletal pose sequences in the same format as vision-based pose estimators, these recognition models transfer directly to the inertial setting, while inheriting none of the limitations of their visual counterpart, such as occlusion sensitivity, the need for controlled environments, and privacy concerns associated with camera-based recording.

3 Method and Approach

This section introduces an Inertial-Poser-Based HAR pipeline and demonstrates it on the TotalCapture dataset. The approach first estimates a full-body pose sequence from IMU data recorded at five on-body locations, which is then passed to a pose-based HAR model. While such a five sensor setup is not yet suitable for everyday use, it is realistic for controlled scenarios such as workflow monitoring, fitness training, or virtual reality applications. The overall process is illustrated in Fig. 1.

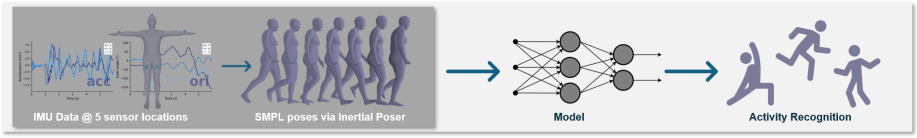

Fig. 1. Visualization of the proposed pipeline, illustrating how IMU data are mapped to a body model representation and subsequently used for HAR.

The goal of our experiments is to investigate whether projecting IMU data onto a body model improves activity recognition performance. To this end, we consider three experimental conditions:

(A) **IMU-based HAR**: activity recognition directly from IMU signals.
(B) **Pose-based HAR**: activity recognition from pose sequences estimated by an inertial poser.
(C) **Combined HAR**: activity recognition from a fusion of IMU signals and inertial poser-derived pose sequences.

3.1 Experimental Setup for the Pose Estimation

Sensor Data. For our experiments, we use the TotalCapture dataset by Trumble et al. [22], which was originally collected for pose estimation using synchronized IMU and optical motion capture recordings. The dataset provides multi-sensor on-body IMU recordings alongside ground truth body poses obtained via optical motion capture.

From the 13 available sensor locations, we select five: the left and right thighs, left and right wrists, and the head. Each IMU provides three-axis gravity-corrected acceleration and a 3×3 rotation matrix encoding the sensor's orientation in the global reference frame. Gyroscope readings, magnetometer data, and raw (non-gravity-corrected) acceleration are also available.

Inertial Poser. Our inertial poser is based on MobilePoser by Xu et al. [24], which was originally designed for sparse consumer-device sensor configurations. We adapt their model to the five sensor locations described above and retrain it. Inertial posers are commonly trained on the AMASS dataset [17], which consolidates several motion capture datasets, including TotalCapture, into a unified SMPL parameterization. Synthetic IMU signals are generated from the SMPL pose sequences to serve as training data. Since TotalCapture is part of AMASS, we explicitly exclude it from the inertial poser's training data to prevent data leakage.

The pose model takes as input gravity-corrected accelerometer data and orientation from all five sensor locations, resulting in a 60-dimensional input vector per time step.

The pose model outputs a full-body pose sequence at 30 Hz, parameterized as 24 SMPL joint rotations, along with the root translation in world space, as

illustrated in Figs. 2 and 3. Note that in the experiments presented here, the root translation, representing the subject's movement trajectory through space, is not used as input to the HAR models.

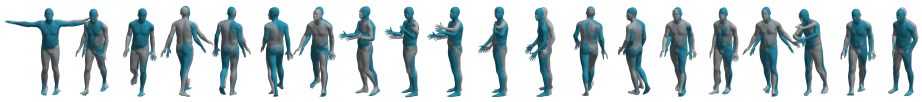

Fig. 2. Snapshots from an exemplary 19-second pose sequence, showing one frame per second (20 snapshots total). Each snapshot displays two overlapping poses: the grey pose represents the ground truth, taken from the SMPL parameterization of the TotalCapture sequence in AMASS, while the blue pose is the inertial poser's prediction, estimated from sensor orientation and gravity-corrected accelerometer data recorded at five on-body locations. The predicted root orientation is also shown, reflected in the overall body orientation of the blue pose.

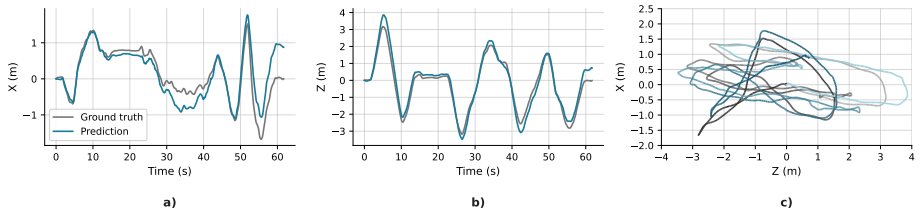

Fig. 3. Exemplary root translation trajectory for the full sequence of 62 seconds, corresponding to the same pose sequence as in Fig. 2. **a)** World-space X coordinate over time. **b)** World-space Z coordinate over time. **c)** Full 2D floor trajectory in the X–Z plane, viewed from above, where color encodes temporal progression from early (light) to late (dark). Ground truth is shown in grey, prediction in blue, both aligned to the same origin. While pose estimation quality remains stable (cf. Fig. 2), the global translation error accumulates over time, visible as increasing divergence towards the end in a) and b) and as spatial drift in c).

The inertial poser architecture is borrowed from MobilePoser and consists of four sequential RNN modules: (1) a Joints module that predicts 3D joint positions; (2) a Poser module that takes the raw IMU signals and the joint positions as input to predict 16 joint rotations; (3) a FootContact module that estimates the probability of foot-ground contact as a binary classification; and (4) a Velocity module that predicts joint velocities for root translation estimation. Each module shares the same backbone: a linear projection layer, followed by a two-layer bidirectional LSTM (hidden size 256, dropout 0.4), and a final linear output layer. The pose model was trained in under 24 hours on a NVIDIA A40 GPU. Root translation is estimated by fusing two signals: the velocity predicted

Table 1. Pose estimation results of our MobilePoser adaptation on TotalCapture using 5 sensor locations, averaged over 37 test sequences. **SIP Err.**: mean geodesic rotation error over 4 selected joints (right hip, right knee, left shoulder, left wrist). **Ang. Err.**: mean geodesic rotation error over all 24 SMPL joints. **MPJPE**: mean euclidean joint position error after root alignment. **Mesh Err. (MPVPE)**: mean euclidean vertex position error over all 6,890 SMPL mesh vertices after root alignment. **Jitter**: mean magnitude of the third-order finite difference of predicted joint positions, measuring temporal smoothness. **Dist. Err.**: mean ℓ_2 root displacement error over 1-second windows. All metrics: lower is better. See Appendix A for formal definitions of the metrics.

Inertial Poser	SIP Err. (°)	Ang. Err. (°)	MPJPE (cm)	Mesh Err. (cm)	Jitter ($100\,\mathrm{m/s^3}$)	Dist. Err. (cm)
	18.56	17.30	9.17	10.42	0.91	20.04

by the network and a physics-based foot-contact constraint, which enforces a stationary foot assumption when contact is detected. The model is applied in offline mode, with a temporal context of 20 past and 5 future frames. The inertial poser is not fine-tuned on TotalCapture.

To assess pose estimation quality, we evaluate the predicted poses against the ground-truth SMPL sequences from AMASS using standard pose estimation metrics, more thoroughly described in Appendix A. The results are shown in Table 1.

Working with ground-truth poses also allows us to isolate the contribution of body model representations to inertial-based HAR, decoupling it from any confounding errors introduced by imperfect pose estimation. For this reason, our pose-based HAR experiments are conducted using both predicted and ground-truth poses. Ground-truth SMPL sequences are taken from the AMASS dataset [17]. Each frame of the SMPL pose data is represented by the rotations of $J = 24$ joints in axis-angle form, yielding a pose vector $\boldsymbol{\theta} \in \mathbb{R}^{24\times 3}$.

TotalCapture Data. The TotalCapture dataset contains four activity labels: *acting*, *walking*, *freestyle*, and *ROM*, where *ROM* includes structured isolated movements of upper and lower limbs and head rolls. These classes exhibit high intra-class variability, making classification challenging, particularly for *acting* and *freestyle*.

The IMU and pose data are temporally aligned and resampled to a common frame rate of 30 Hz. In total, the dataset comprises approximately 97 minutes of recordings across five subjects, with each recording corresponding to a single activity class.

3.2 Experimental Design for the HAR Task

We now define the three experimental conditions, which are distinguished by the input representation fed to the HAR models.

(A) **IMU-based HAR**: activity recognition directly from IMU signals, comprising gravity-corrected accelerometer data (g-acc), sensor orientation (ori), and gyroscope (gyr) readings from all 5 sensor locations, yielding a $5 \times 15 = 75$-dimensional (3 (g-acc) +9 (ori) +3 (gyr) = 15) input vector per frame. This sensor combination was found to yield the best recognition performance among the configurations evaluated (see Appendix B). It also ensures a fair comparison with condition (B), which requires gravity-corrected acceleration and sensor orientation as input. Using raw acceleration without orientation would introduce an artificial disadvantage, as orientation carries significant information about body movement.

(B) **Pose-based HAR**: activity recognition from pose sequences represented as SMPL joint rotation parameters $\boldsymbol{\theta} \in \mathbb{R}^{72}$ per frame. Note that while the inertial poser internally requires gravity-corrected acceleration and sensor orientation as input to estimate the pose, these raw IMU signals are not passed to the HAR model, only the resulting pose parameters $\boldsymbol{\theta}$ are used for recognition. We evaluate on both inertial-poser-estimated and ground-truth pose sequences, allowing us to disentangle recognition performance from pose estimation error.

(C) **Combined HAR**: activity recognition from a fusion of the IMU signals described in (A) and the pose sequences described in (B). This condition examines whether the two representations carry complementary information, the raw IMU signals capturing local sensor dynamics and the pose parameters providing a globally consistent, anatomically structured body representation.

All experiments are conducted with a window length of 50 frames and a 50% overlap, resulting in a temporal window length of 1.67 seconds. All models predict the activity class from a single time window of fixed length, without access to predictions from adjacent windows. This reflects a realistic deployment scenario where the current activity window must be classified immediately, and the next may belong to a different activity class. To evaluate model generalization, we employ Leave-One-Subject-Out (LOSO) cross-validation and report results averaged across all five folds. All models are trained and evaluated under this same protocol with fixed, benchmark-level hyperparameters rather than extensive model-specific hyperparameter optimization, to keep the comparison across models fair.

The HAR models we report on are[1]:

(A) **IMU-based HAR:**
  - (A1) LightGBM [10]: hand-crafted statistical and frequency-based features extracted from the gravity-corrected accelerometer, orientation, and gyroscope signals of five body-worn IMUs, classified with gradient boosting.
  - (A2) DeepConvLSTM [19]: raw IMU time series (30 Hz) processed end-to-end using stacked convolutional and recurrent layers.

[1] Implementation details are available at https://github.com/rilink/harpose.

(A3)–(A5) CNNHAR [26], Attend & Discriminate [1], and TinyHAR [29]: more widely adopted wearable HAR baselines operating on raw IMU time series.

(B) **Pose-based HAR:**
- (B1) SMPL-CNN: the pose sequence is treated as a 2D feature map of shape $(J\times3) \times T$ (J=24 joints, T=50 frames within one window) and processed by stacked 2D convolutional layers that jointly capture temporal patterns and cross-joint correlations.
- (B2) vel-CNN [11]: extends (B1) by concatenating per-joint angular velocities via finite differences of the axis-angle pose, providing an explicit motion signal alongside the static pose representation.
- (B3) TCN-FK [2]: a Temporal Convolutional Network (TCN) operating on 3D joint positions obtained via SMPL forward kinematics, using dilated causal convolutions for long-range temporal modeling.
- (B4) ST-GCN [25]: models the SMPL skeleton as a graph and propagates features along both the spatial (kinematic chain) and temporal dimensions.

All (B)-category models are evaluated under ground-truth pose (GT) and inertial-poser-estimated pose (PRED).

(C) **Fusion of IMU and Pose:**
- (C1) Threshold fusion of an IMU model and a pose model: the pose models output is used by default. The IMU model overrides it when the pose model's maximum class probability falls below a threshold τ, selected per LOSO fold on held-out subjects. We evaluate two IMU models: (A2) DeepConvLSTM and (A1) LightGBM against (B3) TCN-FK.
- (C2) Oracle: selects whichever of the IMU model or the pose model predicts the correct label at each window, representing the performance ceiling of any fusion of these two modalities, while introducing a data leak. Evaluated with both (A2) and (A1) as the IMU models and (B3) as the pose model.

Both (C1) and (C2) are evaluated under the same GT/PRED conditions as category (B).

4 Results and Discussion

Table 2 summarizes the results of our experiments.

(A) Using only inertial data, the feature-based approach ((A1), macro-F1 = 0.545) and the DeepConvLSTM ((A2), macro-F1 = 0.546) achieve comparable performance. The three additional wearable HAR baselines (A3)–(A5) score between 0.522 and 0.527, confirming that (A1) and (A2) represent strong IMU-only references for this dataset.

(B) Pose-based models almost consistently outperform the IMU-only baselines when ground-truth pose is available (GT macro-F1 ranging from 0.552 to 0.587). This improvement is not unexpected: the pose data captures the

full-body configuration and limb coordination, offering a more comprehensive view of the performed activity than local inertial measurements alone. Among the pose-based models, the velocity-augmented CNN (B2) achieves the highest GT performance (0.587), while the TCN operating on forward-kinematics positions (B3) proves the most robust under estimated pose (PRED macro-F1 = 0.562 vs. GT macro-F1 = 0.573), showing only a small degradation when the inertial poser's output replaces ground-truth SMPL parameters. Notably, even under estimated pose, the best pose-based model ($(B3)_{\mathrm{PRED}}$, 0.562) still exceeds the IMU-only ceiling ((A2), 0.546), indicating that the inertial poser captures sufficient body-configuration information to benefit downstream activity recognition.

(C) Both (C1) variants fuse an IMU model with (B3) via a confidence-threshold rule, deferring to the IMU model only when the pose model's confidence falls below a per-fold threshold τ. Fusing (A2) with (B3) yields macro-F1 = 0.588 (GT) and 0.574 (PRED); fusing (A1) with (B3) gives virtually identical results (0.589 GT, 0.572 PRED). Both configurations improve over either individual model, and the negligible difference between (A1) and (A2) as fusion partners suggests that the pose model dominates the combined prediction. The oracle (C2) selects whichever of the two models is correct at each window, representing the ceiling achievable by any fusion of these modalities. Again, the choice of IMU model has little effect: the (A2) + (B3) oracle reaches macro-F1 = 0.660 (GT) and 0.651 (PRED), while (A1) + (B3) yields 0.659 (GT) and 0.655 (PRED). The near-identical oracle scores confirm that the performance ceiling is set by (B3), not by the IMU model, and that the two modalities carry substantial complementary information, leaving considerable room for more sophisticated fusion strategies. Overall, these results support our hypothesis that incorporating a body model representation within the HAR pipeline improves recognition performance, particularly when combined with inertial measurements.

5 Exploration of Conceptual Advantages of Inertial-Poser-Based HAR

Comparing the raw inertial data-based HAR approaches discussed in Section 2.1 several key advantages of the proposed inertial-poser-based approach from Section 2.3 become apparent. The main advantages we identified for using inertial posers for HAR are summarized below.

Methodology

1) Through the projection of IMU data into the motion space, established computer vision techniques, such as skeleton-based HAR methods [15], can now be applied to inertial data. This opens possibilities for cross-domain transfer between vision- and sensor-based approaches, a direction already demonstrated empirically in Section 3.

Table 2. Activity recognition LOSO results (mean ± std over 5 subjects, 50-frame windows at 30 Hz). Category A uses IMU only; B uses SMPL pose only; C fuses both. "GT" denotes ground-truth SMPL pose, "PRED" denotes inertial-poser-estimated pose.

Cat.	Model	Pose	macro-F1	Accuracy
A	A1 LightGBM	–	0.545 ± 0.061	0.681 ± 0.095
	A2 DeepConvLSTM	–	$\mathbf{0.546 \pm 0.044}$	$\mathbf{0.710 \pm 0.040}$
	A3 CNNHAR	–	0.522 ± 0.049	0.674 ± 0.075
	A4 AttendDiscriminate	–	0.527 ± 0.017	0.683 ± 0.063
	A5 Tinyhar	–	0.522 ± 0.028	0.665 ± 0.064
B	B1 SMPL-CNN	GT	0.564 ± 0.081	0.708 ± 0.122
		PRED	0.537 ± 0.085	0.672 ± 0.076
	B2 vel-CNN	GT	$\mathbf{0.587 \pm 0.061}$	$\mathbf{0.754 \pm 0.068}$
		PRED	0.529 ± 0.090	0.675 ± 0.061
	B3 TCN-FK	GT	0.573 ± 0.040	0.737 ± 0.068
		PRED	$\mathbf{0.562 \pm 0.045}$	$\mathbf{0.703 \pm 0.019}$
	B4 ST-GCN	GT	0.552 ± 0.053	0.728 ± 0.077
		PRED	0.529 ± 0.058	0.702 ± 0.090
C	C1 Threshold Fusion (A2+B3)	GT	0.588 ± 0.037	$\mathbf{0.749 \pm 0.059}$
		PRED	$\mathbf{0.574 \pm 0.043}$	$\mathbf{0.730 \pm 0.021}$
	C1 Threshold Fusion (A1+B3)	GT	$\mathbf{0.589 \pm 0.044}$	0.744 ± 0.058
		PRED	0.572 ± 0.044	0.719 ± 0.024
	C2 Oracle (A2+B3)	GT	0.660 ± 0.051	0.820 ± 0.040
		PRED	0.651 ± 0.047	0.817 ± 0.035
	C2 Oracle (A1+B3)	GT	0.659 ± 0.072	0.807 ± 0.054
		PRED	0.655 ± 0.060	0.808 ± 0.036

2) Some inertial posers also estimate the subject's translation within the environment, as illustrated in Fig. 3. The resulting spatial information, such as the position or trajectory within a room, can be used as additional information for activity recognition, which is especially relevant for applications like virtual reality or spatially constrained tasks.

Interpretability

3) Raw inertial data cannot be easily projected back into the 3D movement space. In contrast, inertial posers enable visual reconstruction of pose sequences, allowing ambiguous or uncertain predictions to be reviewed visually. This improves interpretability, supports interactive system feedback, and enables post-hoc data annotation, which was previously not possible without ground truth motion capture or video data.

4) The 3D trajectory in the movement space of individual IMUs can only be theoretically estimated through dead-reckoning and sensor fusion. Practically, this is infeasible due to drift accumulation [9]. Via the inertial poser, we have recovered the possibility of obtaining this spatial trajectory information.

Robustness and Generalizability

5) Incorporating a body model into the HAR pipeline introduces strong priors about feasible poses and sensor locations in 3D space, allowing sensors to compensate for each other's drift, thereby increasing robustness. Moreover, body parts without direct sensor coverage can still be inferred through the pose model.
6) The orientation of the root joint can be used to correct for apparent sensor movements induced by the global body rotation, improving the overall consistency of the reconstructed motion sequence, as illustrated in Fig. 4.
7) Using a parametric body model such as SMPL also enhances generalizability. Individual body characteristics are represented by the β parameter, allowing for adaptation across subjects with different body shapes. This offers opportunities for data augmentation and personalized HAR models that account for user-specific biomechanics.

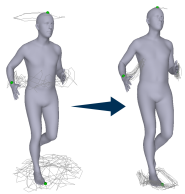

Fig. 4. Movement trajectories of five virtual sensor locations derived from an AMASS jogging sequence. The subject's circular motion causes each sensor to trace a circular path, which could mislead a raw IMU-based model into confusing global locomotion with local limb movement. Since joints in the SMPL kinematic tree are defined by local rotations relative to their parent, the root orientation accounted for without affecting any child joint's local motion, effectively decoupling global body rotation from activity-relevant movement patterns.

6 Limitations and Future Work

Although the presented results confirm the potential of incorporating body model representations into inertial-based HAR, several limitations remain. First, the TotalCapture dataset contains only a limited range of broad activities and a small number of subjects, which restricts generalizability. Extending evaluations to more diverse datasets and will be essential to assess robustness across

users, contexts, and activity types. Recent advances in synthetic IMU data generation [13, 14] have shown promise in alleviating the data scarcity problem in inertial-based HAR, potentially enabling the training of more generalizable models without the need for large-scale real-world sensor recordings. In parallel, developments in soft-tissue and wearable research may enable more flexible and comfortable sensor configurations that do not require rigid attachment to the body. In future work, we also aim to evaluate the conceptual advantages outlined in Section 5 to assess their impact on recognition performance and model interpretability. In particular, the global translation estimated by the body model was excluded from our experiments, as the activities in the TotalCapture dataset are not tied to specific locations within the capture space. In scenarios where activities are spatially distributed, for example, cooking confined to a kitchen area or desk work at a fixed workstation, global translation would provide a strong complementary cue for activity recognition and is a natural extension of the proposed pipeline. Beyond translation, future work could explore more expressive fusion architectures that learn to combine IMU and pose representations, enabling more robust and interpretable HAR systems. Future studies should also investigate how reduced sensor setups perform when only partial body coverage, e.g. upper body only, is available and whether the projection onto a body model remains beneficial in such cases.

7 Conclusion

In this work, we explored the integration of body model representations into inertial-based human activity recognition. By comparing IMU-based, pose-based, and fused approaches, we demonstrated that incorporating a body model, here through the SMPL model, can enhance recognition performance. Despite the IMU-only models and the pose-based models both being based on the same IMU data, the pose-based model outperforms IMU-only baselines, suggesting that the intermediate body model representation extracts structure from the inertial data that raw IMU models do not fully capture. Furthermore, our oracle analysis revealed that the IMU and pose branches carry complementary information, indicating substantial room for more sophisticated fusion strategies. In addition, we argued that such body model representations offer greater interpretability and robustness by embedding strong priors about human kinematics and spatial structure. These findings highlight the potential of inertial-poser-based pipelines as a promising direction for future HAR systems, particularly as inertial posers continue to improve.

Acknowledgments. The authors acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG.

References

1. Abedin, A., Ehsanpour, M., Shi, Q., Rezatofighi, H., Ranasinghe, D.C.: Attend and discriminate: Beyond the state-of-the-art for human activity recognition using wearable sensors. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **5**(1), 1–22 (2021)
2. Bai, S., Kolter, J.Z., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271 (2018)
3. Bian, S., Liu, M., Zhou, B., Lukowicz, P.: The state-of-the-art sensing techniques in human activity recognition: A survey. Sensors **22**(12), 4596 (2022)
4. Dehzangi, O., Sahu, V.: Imu-based robust human activity recognition using feature analysis, extraction, and reduction. In: 2018 24th international conference on pattern recognition (ICPR). pp. 1402–1407. IEEE (2018)
5. Duan, H., Zhao, Y., Chen, K., Lin, D., Dai, B.: Revisiting skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2969–2978 (2022)
6. Farahan, S.B., Machado, J.J., de Almeida, F.G., Tavares, J.M.R.: 9-dof imu-based attitude and heading estimation using an extended kalman filter with bias consideration. Sensors **22**(9), 3416 (2022)
7. Huang, Y., Kaufmann, M., Aksan, E., Black, M.J., Hilliges, O., Pons-Moll, G.: Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. ACM Transactions on Graphics (TOG) **37**(6), 1–15 (2018)
8. Jiang, Y., Ye, Y., Gopinath, D., Won, J., Winkler, A.W., Liu, C.K.: Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022)
9. Jimenez, A.R., Seco, F., Prieto, C., Guevara, J.: A comparison of pedestrian dead-reckoning algorithms using a low-cost mems imu. In: 2009 IEEE International Symposium on Intelligent Signal Processing. pp. 37–42. IEEE (2009)
10. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems **30** (2017)
11. Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F.: A new representation of skeleton sequences for 3d action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3288–3297 (2017)
12. Konak, O., Wegner, P., Arnrich, B.: Imu-based movement trajectory heatmaps for human activity recognition. Sensors **20**(24), 7179 (2020)
13. Kwon, H., Tong, C., Haresamudram, H., Gao, Y., Abowd, G.D., Lane, N.D., Ploetz, T.: Imutube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **4**(3), 1–29 (2020)
14. Leng, Z., Bhattacharjee, A., Rajasekhar, H., Zhang, L., Bruda, E., Kwon, H., Plötz, T.: Imugpt 2.0: Language-based cross modality transfer for sensor-based human activity recognition. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **8**(3), 1–32 (2024)
15. Liu, M., Liu, H., Hu, Q., Ren, B., Yuan, J., Lin, J., Wen, J.: 3d skeleton-based action recognition: A review. arXiv preprint arXiv:2506.00915 (2025)
16. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015)

17. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: International Conference on Computer Vision. pp. 5442–5451 (Oct 2019)

18. Ni, J., Tang, H., Haque, S.T., Yan, Y., Ngu, A.H.: A survey on multimodal wearable sensor-based human action recognition. arXiv preprint arXiv:2404.15349 (2024)

19. Ordóñez, F.J., Roggen, D.: Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors **16**(1), 115 (2016)

20. Roetenberg, D., Luinge, H., Slycke, P., et al.: Xsens mvn: Full 6dof human motion tracking using miniature inertial sensors. Xsens Motion Technologies BV, Tech. Rep **1**(2009), 1–7 (2009)

21. Stenum, J., Cherry-Allen, K.M., Pyles, C.O., Reetzke, R.D., Vignos, M.F., Roemmich, R.T.: Applications of pose estimation in human health and performance across the lifespan. Sensors **21**(21), 7315 (2021)

22. Trumble, M., Gilbert, A., Malleson, C., Hilton, A., Collomosse, J.: Total capture: 3d human pose estimation fusing video and inertial sensors. In: Proceedings of 28th British Machine Vision Conference. pp. 1–13 (2017)

23. Von Marcard, T., Rosenhahn, B., Black, M.J., Pons-Moll, G.: Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In: Computer graphics forum. vol. 36, pp. 349–360. Wiley Online Library (2017)

24. Xu, V., Gao, C., Hoffmann, H., Ahuja, K.: Mobileposer: Real-time full-body pose estimation and 3d human translation from imus in mobile consumer devices. In: Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology. pp. 1–11 (2024)

25. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)

26. Yang, J., Nguyen, M.N., San, P.P., Li, X., Krishnaswamy, S., et al.: Deep convolutional neural networks on multichannel time series for human activity recognition. In: Ijcai. vol. 15, pp. 3995–4001. Buenos Aires, Argentina (2015)

27. Yi, X., Zhou, Y., Habermann, M., Shimada, S., Golyanik, V., Theobalt, C., Xu, F.: Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13167–13178 (2022)

28. Zhou, H., Arakawa, R., Agarwal, Y., Goel, M.: Imucoco: Enabling flexible on-body imu placement for human pose estimation and activity recognition. In: Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology. pp. 1–16 (2025)

29. Zhou, Y., Zhao, H., Huang, Y., Riedel, T., Hefenbrock, M., Beigl, M.: Tinyhar: A lightweight deep learning model designed for human activity recognition. In: Proceedings of the 2022 ACM International Symposium on Wearable Computers. pp. 89–93 (2022)

30. Zinnen, A., Blanke, U., Schiele, B.: An analysis of sensor-oriented vs. model-based activity recognition. In: 2009 International Symposium on Wearable Computers. pp. 93–100. IEEE (2009)

A Pose Estimation Metrics

SIP Error: mean geodesic rotation error over a predefined set of 4 joints $\mathcal{S} = \{\text{right hip, right knee, left shoulder, left wrist}\}$,

$$\text{SIP Error} := \frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} d_g(\hat{\mathbf{R}}_j, \mathbf{R}_j), \quad \text{where} \quad d_g(\hat{\mathbf{R}}_j, \mathbf{R}_j) = \arccos\left(\frac{\text{tr}(\hat{\mathbf{R}}_j^\top \mathbf{R}_j) - 1}{2}\right)$$

and $\hat{\mathbf{R}}_j, \mathbf{R}_j \in SO(3)$ are the predicted and ground-truth rotation matrices of joint $j \in \mathcal{S}$, respectively.
Angular Error: mean geodesic rotation error over all $J = 24$ SMPL joints,

$$\text{Angular Error} := \frac{1}{J} \sum_{j=1}^{J} d_g(\hat{\mathbf{R}}_j, \mathbf{R}_j).$$

MPJPE (Mean Per Joint Position Error): mean euclidean distance between predicted and ground-truth joint positions across all 24 SMPL joints, computed after aligning the predicted root joint to ground truth,

$$\text{MPJPE} := \frac{1}{J} \sum_{j=1}^{J} \|\hat{\mathbf{p}}_j - \mathbf{p}_j\|_2,$$

where $J = 24$ is the number of joints, and $\hat{\mathbf{p}}_j, \mathbf{p}_j \in \mathbb{R}^3$ are the predicted and ground-truth positions of joint j, respectively.
Mesh Error: mean euclidean error over all 6,890 SMPL mesh vertices after root alignment,

$$\text{Mesh Error} := \frac{1}{V} \sum_{v=1}^{V} \|\hat{\mathbf{v}}_v - \mathbf{v}_v\|_2,$$

where $V = 6{,}890$ is the number of mesh vertices, and $\hat{\mathbf{v}}_v, \mathbf{v}_v \in \mathbb{R}^3$ are the predicted and ground-truth positions of vertex $v \in \{1, \ldots, V\}$, respectively.
Jitter: temporal smoothness measured as the mean magnitude of the third-order finite difference of predicted joint positions across all $J = 24$ SMPL joints,

$$\text{Jitter} := \frac{1}{J} \sum_{j=1}^{J} \left\| \frac{\Delta^3 \hat{\mathbf{p}}_j}{\Delta t^3} \right\|_2,$$

where $\hat{\mathbf{p}}_j \in \mathbb{R}^3$ is the predicted position of joint j.
Distance/Translation Error: root translation error per second, computed as the mean ℓ_2 distance between predicted and ground-truth root displacement over all 1-second windows,

$$\text{Dist. Error} := \frac{1}{W} \sum_{w=1}^{W} \|\hat{\mathbf{r}}_w - \mathbf{r}_w\|_2,$$

where W is the number of 1-second windows and $\hat{\mathbf{r}}_w, \mathbf{r}_w \in \mathbb{R}^3$ are the predicted and ground-truth root displacements in window w, respectively.

B IMU Sensor Combination Study

To determine the most informative input representation for the IMU-based HAR baseline (A), we conducted a preliminary study comparing different combinations of IMU signal types on the TotalCapture dataset using a DeepConvLSTM [19] and a feature-based LightGBM classifier. The evaluated combinations vary in whether gravity-corrected acceleration (g-acc), raw acceleration (acc), sensor orientation (ori), gyroscope (gyr) and magnetometer (mag) readings are included. Table 3 reports the macro-F1 scores and accuracies averaged over all LOSO folds, and confirms that the combination of gravity-corrected acceleration, sensor orientation, and gyroscope data yields the best overall performance, motivating its use as the input representation in condition (A).

Table 3. Sensor-combination for IMU-only approach (LOSO mean ± std over 5 subjects).

	A1 LightGBM		A2 DeepConvLSTM	
Sensors	F1	Acc	F1	Acc
raw acc, gyr, mag	0.492 ± 0.030	0.657 ± 0.038	0.448 ± 0.093	0.611 ± 0.119
g-acc, ori, gyr, mag	0.519 ± 0.054	0.676 ± 0.089	0.523 ± 0.031	0.695 ± 0.046
g-acc, ori, gyr	$\mathbf{0.545 \pm 0.061}$	$\mathbf{0.681 \pm 0.095}$	0.546 ± 0.044	$\mathbf{0.710 \pm 0.040}$
g-acc, ori	0.541 ± 0.077	0.679 ± 0.106	$\mathbf{0.547 \pm 0.050}$	0.694 ± 0.037