This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

UMotion: Uncertainty-driven Human Motion Estimation from Inertial and Ultra-wideband Units

Huakun Liu Hiroki Ota Xin Wei Yutaro Hirao
Monica Perusquía-Hernández Hideaki Uchiyama Kiyoshi Kiyokawa
Nara Institute of Science and Technology, Japan
{liu.huakun.li0, ota.hiroki.oc6, wei.xin.wy0, yutaro.hirao}@is.naist.jp
{m.perusquia, hideaki.uchiyama, kiyo}@is.naist.jp

Abstract

Sparse wearable inertial measurement units (IMUs) have gained popularity for estimating 3D human motion. However, challenges such as pose ambiguity, data drift, and limited adaptability to diverse bodies persist. To address these issues, we propose UMotion, an uncertainty-driven, online fusing-all state estimation framework for 3D human shape and pose estimation, supported by six integrated, body-worn ultra-wideband (UWB) distance sensors with IMUs. UWB sensors measure inter-node distances to infer spatial relationships, aiding in resolving pose ambiguities and body shape variations when combined with anthropometric data. Unfortunately, IMUs are prone to drift, and UWB sensors are affected by body occlusions. Consequently, we develop a tightly coupled Unscented Kalman Filter (UKF) framework that fuses uncertainties from sensor data and estimated human motion based on individual body shape. The UKF iteratively refines IMU and UWB measurements by aligning them with uncertain human motion constraints in real-time, producing optimal estimates for each. Experiments on both synthetic and real-world datasets demonstrate the effectiveness of UMotion in stabilizing sensor data and the improvement over state of the art in pose accuracy. Code is available at: https://github.com/kk9six/umotion.

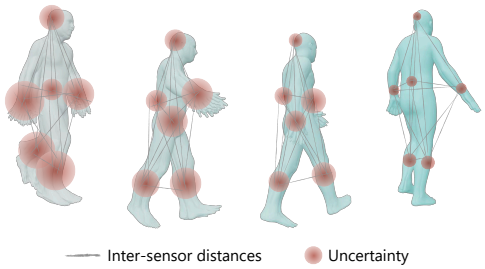


Figure 1. UMotion integrates IMU-UWB data inputs and pose outputs uniformly under uncertainty, constrained by individual body structure. The online state estimation framework iteratively refines sensor data confidence and stabilizes pose estimation, reducing ambiguities and improving robustness.

1. Introduction

Estimating 3D human motion from wearable sensors has become increasingly popular due to their portability, accessibility, and versatility. Wearable sensors, such as inertial measurement units (IMUs), enable continuous monitoring of body motion measurements across unrestricted spaces. These advances shift motion capture from controlled laboratory environments to everyday settings [13, 19, 32, 41]. This transition benefits fields such as healthcare, sports performance, ergonomics, and emerging areas in human-computer interaction [3, 8, 14, 31].

One of the widely chosen wearable sensors for 3D human motion estimation is IMU. Commercial systems, such as Xsens [27], utilize 17 or more IMUs for comprehensive pose coverage. While highly accurate, these densely placed IMUs are inconvenient and intrusive. Recent studies have reduced the required number of IMUs to just six—placed on the forearms, lower legs, pelvis, and head—while still achieving promising performance through data-driven methods [1, 7, 10, 31, 32, 34, 35, 37, 39]. With fewer IMUs, pose estimation becomes under-constrained and prone to ambiguity. Recent work has attempted to disambiguate poses by incorporating temporal consistency, physics-based constraints, or additional, easy-to-integrate sensors [1, 10, 35]. However, these methods still face significant challenges, including noisy sensor data, body shape variations, and ambiguities arising from under-constrained setups with sparse sensors.

In this work, we propose UMotion, an uncertainty-driven human motion estimation framework that combines online state estimation with an integrated system of six inertial and

ultra-wideband (UWB) distance sensors. Prior work often infers poses from sensor data without considering the reverse influence. UMotion uniformly treats input and output under uncertainty and body-specific constraints, blending data to optimize estimates (see Fig. 1). Our approach maintains inter-sensor distances as a core system state, capturing spatial relationships between sensor nodes. These distances, combined with basic anthropometric measurements (height and weight) and IMU data, serve as inputs for our learning-based shape and pose estimators. Since sensors are inherently noisy, we propose a tightly coupled online state estimation system that associates IMU and UWB sensor measurements and pose estimates using an Unscented Kalman Filter (UKF) and an uncertainty propagation method, iteratively correcting errors and stabilizing pose estimation in real-time. Our main contributions are:

- An ensemble learning-based human shape estimation approach using distance constraints and anthropometric measurements from six integrated IMU-UWB sensors.
- A learning-based method for human pose estimation from inertial and distance constraints, incorporating mesh distribution estimation.
- A filtering-based state estimation system that couples sensor measurements, pose estimates, and body variations in real-time to improve sensor stability and pose accuracy.

2. Related Work

Pose Estimation from Sparse Inertial Sensors In contrast to the predominant use of vision-based methods, wearable sensors, specifically IMUs, offer greater freedom and flexibility. Von Marcard et al. [32] present SIP that makes inertial pose estimation practical by using only six IMUs attached to the body, combined with an offline iterative SMPL body model [17] pose optimization. Huang et al. [7] propose a real-time pose estimation method, DIP, which uses a bidirectional recurrent neural network (biRNN) [28] to learn the mapping from a sequence of six IMU measurements to SMPL body poses. To obtain sufficient training data, they synthesize IMU measurements from the AMASS dataset [18], further advancing the development of learning-based methods for sparse sensor motion capture.

Following previous studies, TransPose [34] refines pose estimation by decomposing the end-to-end framework into a multi-stage process with intermediate joint position estimation. To disambiguate the poses, PIP [35] proposes a physics-based motion optimization, while TIP [10] incorporates a conditional Transformer model for plausible terrain generation. Training data is crucial for learning-based methods. Unlike prior studies relying on synthetic IMU data, DynaIIP [39] adapts real IMU data from diverse human skeleton formats to the target SMPL model and shows superior performance when training with real sensor data. Similarly focusing on the data, PNP [37] recently addresses limitations in existing IMU synthesis by incorporating non-inertial effects and fictitious forces, enhancing estimation robustness through a physics-informed neural network and realistic IMU synthesis.

In contrast to methods focusing solely on pose estimation, our approach treats human shape and pose as equally important. By integrating shape information, we establish a tight connection between sensor data and estimated motions, forming a positive feedback loop that enhances the entire process. Additionally, most previous studies fail to fully utilize IMU accelerations due to high noise and drift issues. Within our framework, accelerations serve as control inputs for state estimation, undergoing continuous correction and acting as a crucial component in tracking spatial relationships among sensors.

Pose Estimation from Hybrid Sensors In addition to IMUs, various wearable and hybrid sensor systems have been explored to overcome inherent IMU limitations such as restricted positional awareness [9, 12, 13, 15, 16, 21–23, 25, 36]. The closest works to ours are SmartPoser [4] and Ultra Inertial Poser (UIP) [1], both of which integrate UWB and IMU sensors for pose estimation. UWB sensors complement IMUs with additional distance information between sensors while preserving the portability and flexibility of tracking devices. SmartPoser [4] focuses on wearable arm pose estimation, combining UWB measurements with IMU data using an off-the-shelf smartwatch and smartphone. UIP [1] integrates six UWB sensors with IMUs for full-body pose estimation. Inter-sensor distances, estimated using an EKF that fuses IMU and UWB data, help reduce global translation drift and minimize position jitter compared to inertial-only tracking. However, body occlusion frequently affects certain node pairs, e.g., head-knee, causing the EKF to fail in estimating distances due to rapid IMU drift and unreliable UWB measurements, resulting in unstable distance data. Our approach addresses this limitation by using output pose uncertainties and body-specific constraints as observations to refine input sensor measurements, improving the stability of both IMU and UWB data.

3. Method

3.1. Preliminaries

SMPL body model We use the SMPL model [17] to represent human motion. SMPL decomposes the human body into pose parameters, $\boldsymbol{\theta} \in \mathbb{R}^{23\times3+3}$, which define relative rotations of 23 joints and the global root joint orientation in axis-angle form, and shape parameters, $\boldsymbol{\beta} \in \mathbb{R}^{10}$, capturing body shape variations across individuals. The model is defined by a linear blend-skinning (LBS) function:

$$M(\boldsymbol{\beta},\boldsymbol{\theta}) = W\left(\bar{\mathbf{T}} + B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\theta}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathcal{W}\right), \quad (1)$$

where $B_s(\boldsymbol{\beta})$ and $B_p(\boldsymbol{\theta})$ are shape and pose blend shapes that deform the template mesh $\bar{\mathbf{T}}$ in the zero pose. The mesh is then reposed according to specified poses $\boldsymbol{\theta}$, combining with joint positions $J(\boldsymbol{\beta})$ and blend-skinning weights $\mathcal{W}$.

Inertial Measurement Unit (IMU) A nine-axis MEMS IMU comprises an accelerometer, gyroscope, and magnetometer, measuring accelerations in the sensor local frame F^S, angular velocities in F^S, and magnetic field strengths relative to the Earth's magnetic field, respectively. A commonly used error model for measured accelerations of one IMU, $\boldsymbol{a} \in \mathbb{R}^3$, is given by [26]:

$$\boldsymbol{a} = \boldsymbol{a}_{\text{true}} + \boldsymbol{b}_a + \boldsymbol{n}_a, \tag{2}$$

where $\boldsymbol{a}_{\text{true}}$ is the true acceleration, $\boldsymbol{b}_a$ is the bias, modeled as a random walk process with noise $\boldsymbol{\eta}$ following a zero-mean Gaussian distribution with standard deviation $\boldsymbol{\Sigma}_b$, and $\boldsymbol{n}_a$ is high-frequency white noise. In our work, after a frame calibration process, accelerations are represented as $\boldsymbol{a}^M$ in the SMPL body-centric frame F^M, and orientations are represented as $\boldsymbol{R}^{MB}$, indicating the rotation from bone frame F^B to F^M. For simplicity, we denote $\boldsymbol{a}$ as $\boldsymbol{a}^M$ and $\boldsymbol{R}$ as $\boldsymbol{R}^{MB}$ in the following sections. Details on the calibration process are available in supplementary materials.

Ultra-wideband (UWB) UWB uses the time-of-flight technique to measure distances between two devices, with one acting as the transmitter and the other as the receiver. Unlike other RF techniques, such as WiFi and Bluetooth, UWB operates over a wide frequency range and transmits short pulses, achieving centimeter-level distance measurements with minimal interference [20]. Consequently, it is widely used in high-precision indoor localization systems with multiple fixed anchors and movable tags [38]. In our work, all six integrated UWB sensors are movable and alternately function as both transmitters and receivers, capturing inter-distances between each sensor pair.

3.2. Framework Design

As shown in Fig. 2, our method uses six IMU-UWB sensors and consists of three main modules: a shape estimator, a pose estimator, and a state estimator. The shape estimator takes anthropometric measurements and inter-distances in a T-pose as input, outputting shape parameters $\hat{\boldsymbol{\beta}}$ of the SMPL model. The pose estimator uses filtered frame-aligned, root-normalized IMU measurements and inter-distances as input to estimate pose parameters, $\hat{\boldsymbol{\theta}}$, and corresponding uncertainties, $\hat{\boldsymbol{\Sigma}}$. Given our sensor placement, we do not observe the movement of the hands and feet. Therefore, similar to previous studies, we estimate poses for only 16 joints, including the root joint, while excluding the hand and foot joints. For simplicity, we denote this reduced set of pose parameters as $\boldsymbol{\theta} \in \mathbb{R}^{16\times3}$, while $\boldsymbol{\theta}^{24}$ refers to the full set of pose parameters for 24 joints. Notations with a hat symbol, e.g., $\hat{\boldsymbol{\theta}}$, indicate the corresponding estimated values. Additionally, we do not estimate global translation, as inferring it from only body-worn IMU and UWB sensors without external references would lead to unbounded error accumulation. The state estimator integrates accelerations, UWB measurements, and the estimated poses with uncertainties to track the sensor relative positions, velocities, and acceleration biases of each node at each time step. These refined estimates are then used to refine the inputs to the pose estimator.

The estimation framework is driven by inherent uncertainties present in the system. Specifically, noisy IMU data combined with tracked inter-sensor distances serve as the primary inputs for pose estimation. The generated poses, along with their associated uncertainties that reflect sensor noise, are further constrained by the human model, which in turn refines sensor measurements and distance estimates. This feedback loop propagates uncertainties through the system, continuously updating the belief in system states based on the reliability of each data source—IMU, UWB, and pose estimator with body constraints, ultimately producing robust and accurate estimation.

3.3. Shape and Pose Estimator

3.3.1. Shape Estimator

Spatial inter-distances provide data that approximates the rough skeletal structure of an individual. Previous studies, such as Virtual Caliper [24] and SHAPY [2], have demonstrated a linear relationship between body shape and body measurements. Building on these findings, we adopt an ensemble learning-based shape estimator that uses experimentally selected anthropometric measurements, defined as:

$$\hat{\boldsymbol{\beta}} = SE(H, W, \boldsymbol{D}). \tag{3}$$

Basic body measurements, height $H \in \mathbb{R}^1$ and weight $W \in \mathbb{R}^1$, are easily accessible and provide foundational information about human body proportions. Experimentally selected inter-distances between sensors, $\boldsymbol{D} \in \mathbb{R}^7$, capture relative limb lengths and body structural details that height and weight alone cannot provide.

As shown in Fig. 3, we place virtual sensors on the body mesh and conduct line-of-sight (LOS) simulation experiments with real-world tests to identify and select 7 repeatable distances out of 15 possible options. While the selection may vary based on device characteristics, the chosen distances offer a reliable foundation for adaptation. For model training, we use AutoGluon [5], an AutoML framework that automatically trains and ensembles 11 basic machine learning models, which is more stable than a single linear regression method in our validation. The estimated shape parameters $\hat{\boldsymbol{\beta}}$ are then used to reconstruct realistic human bodies and to propagate estimated poses to spatial constraints in the state estimator.

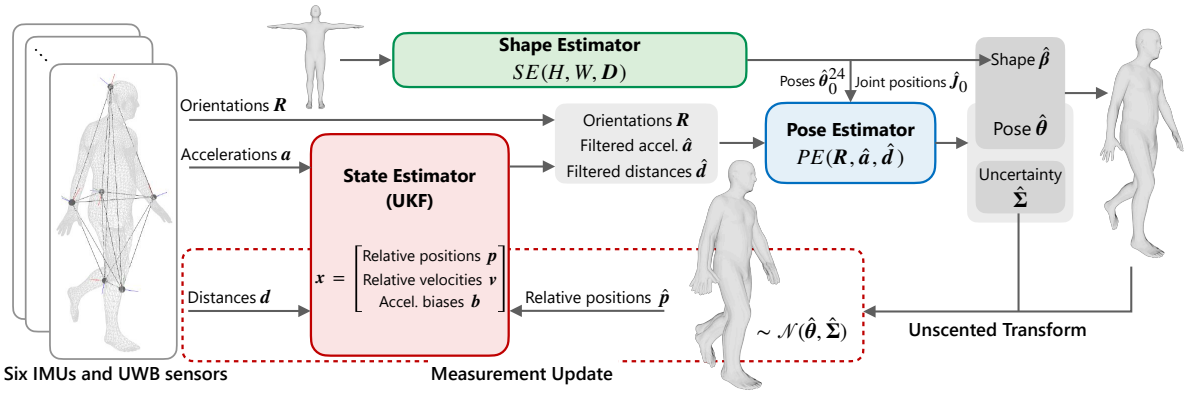


Figure 2. Overview of UMotion, consisting of three main modules: the shape estimator, pose estimator, and state estimator. The shape estimator takes anthropometric measurements and inter-distances in a T-pose as input, outputting shape parameters that reconstruct a realistic body and impose strong constraints on the system. The pose estimator receives filtered IMU data and inter-distances from the state estimator to predict poses, which are fed back to refine state estimates. The entire system integrates IMU, UWB, and estimated poses within the context of individual body structure to continuously update and improve motion estimation.

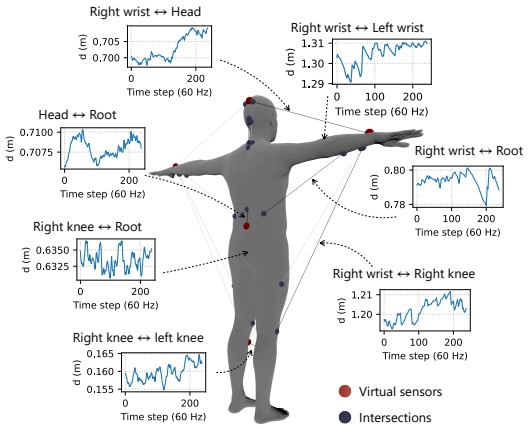


Figure 3. Visualization of selected inter-sensor distances used in the shape estimator. We place virtual sensors on the body mesh at sensor mounting points and conduct line-of-sight simulation experiments. The plots display the temporal changes in UWB measurements between various sensor pairs over time.

3.3.2. Pose Estimator

Our pose estimator is based on unidirectional multi-layer long short-term memory (LSTM) [6] recurrent neural network (RNN). The unidirectional LSTM preserves only past information, achieving real-time estimation without requiring future data. Given six IMU-UWB sensors, we use filtered accelerations, $\hat{\boldsymbol{a}} \in \mathbb{R}^{18}$, and orientations in 9D rotation matrix form, $\boldsymbol{R} \in \mathbb{R}^{54}$, along with 15 inter-distances $\hat{\boldsymbol{d}} \in \mathbb{R}^{15}$, as inputs. The estimator then outputs the poses $\hat{\boldsymbol{\theta}}_{6D} \in \mathbb{R}^{96}$ for the target joints, represented in a 6D rotation matrix form [40]. Additionally, in the final layer, the network outputs the logarithm of the estimation uncertainty, $\hat{\boldsymbol{\Sigma}} \in \mathbb{R}^{96}$, capturing prediction confidence. Consequently, the pose estimator is defined as:

$$\hat{\boldsymbol{\theta}}_{6D}, \hat{\boldsymbol{\Sigma}} = PE(\hat{\boldsymbol{a}}, \boldsymbol{R}, \hat{\boldsymbol{d}}). \tag{4}$$

To handle ambiguity in the initial frame, we adopt a learning-based initialization strategy inspired by PIP [35]. A fully-connected convolutional neural network regresses the initial hidden state of the RNN based on the initial global poses, $\hat{\boldsymbol{\theta}}_0^{24}$, and joint positions, $\hat{\boldsymbol{J}}_0 \in \mathbb{R}^{72}$. This approach provides the pose estimator with a well-informed initialization, adapting to different bodies.

We use two loss functions to train the pose estimator: Mean Squared Error (MSE) and Gaussian Negative Log Likelihood (GNLL) loss. MSE is defined as:

$$\mathcal{L}_{\text{MSE}}(\boldsymbol{\theta}_{6D}, \hat{\boldsymbol{\theta}}_{6D}) = \frac{1}{n} \sum_{i=1}^{n} \left\| \boldsymbol{\theta}_{6D} - \hat{\boldsymbol{\theta}}_{6D} \right\|_2^2, \tag{5}$$

where $\boldsymbol{\theta}_{6D}$ denotes the ground truth poses and n is the total number of frames. The GNLL, $\mathcal{L}_{\text{GNLL}}(\boldsymbol{\theta}_{6D}, \hat{\boldsymbol{\theta}}_{6D}, \hat{\boldsymbol{\Sigma}}) =$

$$\frac{1}{2} \left(\log \left(\max \left(\hat{\boldsymbol{\Sigma}}^2, \varepsilon_{\min} \right) \right) + \frac{(\hat{\boldsymbol{\theta}}_{6D} - \boldsymbol{\theta}_{6D})^2}{\max \left(\hat{\boldsymbol{\Sigma}}^2, \varepsilon_{\min} \right)} \right), \tag{6}$$

where $\varepsilon_{\min}$, set to 10^{-6} in our implementation, is used for numerical stability. First, $\mathcal{L}_{\text{MSE}}$ helps the model converge quickly during early training. Next, $\mathcal{L}_{\text{GNLL}}$ is applied to optimize for pose and uncertainty estimation.

3.4. State Estimator

We use UKF [33] for state estimation to refine the pose estimator inputs, integrating UWB measurements and pose estimator outputs with a statistically derived IMU model.

The UWB measurements provide direct inter-distance data, while they are affected by body occlusion issues. Meanwhile, pose estimator outputs offer constraints and inform distance measurements; however, they suffer from occasional inaccuracies due to challenging poses or sensor noise. The UKF combines these sources, balancing the strengths and limitations of each to make the optimal estimates.

3.4.1. State Definition

We define the state vector $\boldsymbol{x} \in \mathbb{R}^{15\times3+15\times3+6\times3}$ as:

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{p}^{12} & \dots & \boldsymbol{p}^{56} & \boldsymbol{v}^{12} & \dots & \boldsymbol{v}^{56} & \boldsymbol{b}^{1} & \dots & \boldsymbol{b}^{6} \end{bmatrix}^{\mathsf{T}}, \tag{7}$$

where $\boldsymbol{p}^{xy}$ represents the relative position between nodes x and y, with $x, y \in \{1, 2, 3, 4, 5, 6\}$ and $x < y$. The term $\boldsymbol{v}^{xy}$ denotes the relative velocity between nodes x and y, while $\boldsymbol{b}^1$ through $\boldsymbol{b}^6$ represent residual acceleration errors after transformation from F^S to F^M. These biases primarily result from orientation errors and raw accelerometer biases $\boldsymbol{b}_a$. All quantities are expressed in the SMPL body-centric coordinate frame F^M. For initialization, we set $\boldsymbol{p}^{xy}$ based on a static T-pose defined by the shape parameters $\boldsymbol{\beta}$ and the SMPL model (1). The relative velocities $\boldsymbol{v}^{xy}$ and biases are initialized to zero.

3.4.2. State Propagation

The control input $\boldsymbol{u}$ consists of the accelerations $\boldsymbol{a}$ of six nodes, defined as:

$$\boldsymbol{u} = \begin{bmatrix} \boldsymbol{a}^{1} & \dots & \boldsymbol{a}^{6} \end{bmatrix}^{\mathsf{T}}, \boldsymbol{u} \in \mathbb{R}^{18}. \tag{8}$$

During the prediction step, the current state $\boldsymbol{x}$ is propagated using the control input $\boldsymbol{u}$. The propagation model is derived from the strapdown inertial kinematic model with an acceleration error model (2), defined as follows:

$$\boldsymbol{v}_k^{xy} = \boldsymbol{v}_{k-1}^{xy} + (\boldsymbol{a}_{k-1}^{y} - \boldsymbol{a}_{k-1}^{x})\Delta t - (\boldsymbol{b}_{k-1}^{y} - \boldsymbol{b}_{k-1}^{x})\Delta t, \tag{9}$$

$$\begin{aligned} \boldsymbol{p}_k^{xy} = \boldsymbol{p}_{k-1}^{xy} + \boldsymbol{v}_{k-1}^{xy}\Delta t + \frac{1}{2}(\boldsymbol{a}_{k-1}^{y} - \boldsymbol{a}_{k-1}^{x})\Delta t^2 \\ + \frac{1}{2}(\boldsymbol{b}_{k-1}^{x} - \boldsymbol{b}_{k-1}^{y})\Delta t^2, \end{aligned} \tag{10}$$

$$\boldsymbol{b}_k^{x} = \boldsymbol{b}_{k-1}^{x} + \boldsymbol{\eta}^{x}, \boldsymbol{\eta}^{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_b^2), \tag{11}$$

where $k-1$ and k denote consecutive time steps, and Δt represents the interval between them. Together, equations (9), (10), and (11) form a state propagation model with added white noise, represented as:

$$\bar{\boldsymbol{x}}_k = f(\boldsymbol{x}_{k-1}, \boldsymbol{u}_k) + \mathbf{Q}, \tag{12}$$

where $\mathbf{Q}$ is the process noise covariance matrix, derived from the characteristics of the IMUs used in the system.

3.4.3. Measurement Update

The state is updated within measurement spaces using observed data $\boldsymbol{z}$. To accomplish this, we define a measurement model $h(\boldsymbol{x})$ that maps the current state $\boldsymbol{x}$ to measurement spaces as:

$$h(\boldsymbol{x}) = \begin{bmatrix} \|\boldsymbol{p}^{xy}\|_2 & \|\boldsymbol{v}^{xy}\|_2 & \boldsymbol{p}^{xy} & \boldsymbol{v}^{xy} \end{bmatrix}^{\mathsf{T}}. \tag{13}$$

The UWB sensor measures the distance, d^{xy}, between node x and y, which corresponds to the distance derived from relative positions in state, $\|\boldsymbol{p}^{xy}\|_2$. Additionally, from consecutive distance measurements, we derive the relative velocity norm as $(d_k^{xy} - d_{k-1}^{xy})/\Delta t$, corresponding to $\|\boldsymbol{v}\|_2^{xy}$. The covariance matrix for distance measurements at time step k, denoted as $\mathbf{R}_{1,k}$, is dynamically set based on line-of-sight conditions for each node pair, informed by $\hat{\boldsymbol{\theta}}_k$ and UWB sensor characteristics. The covariance matrix for relative velocity measurements, $\mathbf{R}_{2,k}$, is computed as:

$$\mathbf{R}_{2,k} = \frac{(\mathbf{R}_{1,k} + \mathbf{R}_{1,k-1})}{\Delta t^2}. \tag{14}$$

The pose estimator outputs the joint rotations, $\hat{\boldsymbol{\theta}}$, along with their corresponding standard deviations, $\hat{\boldsymbol{\Sigma}}$. We employ the unscented transform [11] to transform the pose distribution to the relative position distribution. Given $\boldsymbol{\Theta} \in \mathbb{R}^{96} \sim \mathcal{N}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}})$, we use Van der Merwe's scaled sigma point algorithm [30] to generate n sigma points, $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n\times96}$, along with weights $\boldsymbol{W}^m \in \mathbb{R}^n$ for the mean and $\boldsymbol{W}^c \in \mathbb{R}^n$ for the covariance. Each sigma point in $\boldsymbol{\mathcal{X}}$ is transformed through the SMPL model (1), yielding a set of sensor-relative positions denoted as:

$$\mathcal{Y} = \{\boldsymbol{p}_i^{xy} \in \mathbb{R}^{n\times15\times3} \mid i = 1, 2, \dots, n\}. \tag{15}$$

We then calculate the mean and covariance of the transformed relative positions, capturing the distribution of relative positions based on pose uncertainty:

$$\hat{\boldsymbol{p}}^{xy} = \sum_{i=1}^{n} \boldsymbol{W}_i^m \mathcal{Y}_i, \tag{16}$$

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{p}}}^2 = \sum_{i=1}^{n} \boldsymbol{W}_i^c (\mathcal{Y}_i - \hat{\boldsymbol{p}}^{xy})(\mathcal{Y}_i - \hat{\boldsymbol{p}}^{xy})^{\mathsf{T}}. \tag{17}$$

The transformed relative positions, $\hat{\boldsymbol{p}}^{xy}$, along with the measurement covariance matrix $\mathbf{R}_3 = \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{p}}}^2$, are used to update $\boldsymbol{p}^{xy}$ in the state $\boldsymbol{x}$. Similarly, we compute the relative velocity between consecutive time steps as $(\hat{\boldsymbol{p}}_k^{xy} - \hat{\boldsymbol{p}}_{k-1}^{xy})/\Delta t$, with its covariance matrix defined as $\mathbf{R}_{4,k} = (\mathbf{R}_{3,k} + \mathbf{R}_{3,k-1})/\Delta t^2$, to update $\boldsymbol{v}^{xy}$. The complete measurement vector $\boldsymbol{z}$ used for updating the state is given by:

$$\boldsymbol{z} = \begin{bmatrix} d^{xy} & \frac{(d_k^{xy} - d_{k-1}^{xy})}{\Delta t} & \hat{\boldsymbol{p}}^{xy} & \frac{(\hat{\boldsymbol{p}}_k^{xy} - \hat{\boldsymbol{p}}_{k-1}^{xy})}{\Delta t} \end{bmatrix}^{\mathsf{T}}. \tag{18}$$

4. Experiments

4.1. Datasets

We use the AMASS [18], TotalCapture [29], DIP-IMU [7], and UIP [1] datasets for training and evaluation. We synthesize inertial and distance data from the AMASS dataset, following the process in previous studies [1, 7, 34]. We select specific vertices on the body mesh to represent sensor mounting positions. Accelerations are computed as finite differences of these vertex positions over time, while orientations are obtained as the global orientations of the corresponding joints. Distance measurements are synthesized by calculating Euclidean distances between selected sensor positions, adjusted for individual body shape. Training the shape estimator requires anthropometric measurements and inter-distances with corresponding shape parameters from a large number of individuals, which is challenging to collect. Therefore, we synthesize these measurements—height H, weight W, and inter-distances $\boldsymbol{D}$—from the 3D body mesh in a T-pose, using 479 unique body shapes (273 male and 206 female) available in the AMASS dataset. We estimate H and W following the methods in SHAPY [2], and we sample $\boldsymbol{D}$ from the synthesized distances.

The training data for the pose estimator comprises synthesized data from the AMASS dataset and the training set of DIP-IMU. Unlike previous studies [10, 34, 35], we exclude the TotalCapture within AMASS from the training data to prevent overlap, ensuring that synthesized distances used in testing are unseen during training. The test data includes TotalCapture, the test split of DIP-IMU, and UIP.

4.2. Method Implementation

Our shape estimator was implemented using the official multi-label predictor in AutoGluon 1.1 with a medium-quality preset. Our pose estimator was implemented based on PyTorch 2.3. The training process used an Adam optimizer with a learning rate of 0.0001. The pose estimator model was trained for 350 epochs with a batch size of 512 on a single NVIDIA GeForce RTX 3090 GPU, completing in approximately 35 minutes. For initial 20 epochs, we adopted $\mathcal{L}_{MSE}$ to optimize the model, then switched to $\mathcal{L}_{GNLL}$ for the remaining epochs. For the UKF, we experimentally set parameters $\alpha_{UKF} = 0.2$, $\beta_{UKF} = 1.0$, and $\kappa_{UKF} = -105$ to control the distribution and weighting of sample points. In the unscented transform applied to pose estimates, we used $\alpha_{NN} = 0.09$, $\beta_{NN} = 1.0$, and $\kappa_{NN} = -93$. We applied a factor of 10 to $\mathbf{R}_3$ to compensate for overconfident predictions. The method achieves 60 Hz without online LOS inference and 30 Hz with it.

4.3. Quantitative Evaluation

We evaluate our method against two categories of methods: (1) IMU-only methods, including DIP [7], TransPose [34], TIP [10], PIP [35], and PNP [37], and (2) distance augmented methods, including TIP-D, PIP-D, and UIP [1]. TIP-D and PIP-D are modified versions of TIP and PIP with distance-augmented input data, and we use their results as reported by UIP. We follow the evaluation protocol in UIP by adding ideal synthetic inter-sensor distances without noise to TotalCapture and DIP-IMU to assess the impact of distance measurements independent of noise. For evaluations on the UIP dataset, which includes both IMU and UWB data, we perform RANSAC regression to calibrate UWB measurements, as described in UIP [1], and then use calibrated measurements as input for the state estimator. We do not apply outlier filtering and use default state estimator parameters due to the absence of ground-truth data and limited information on sensor-specific characteristics in the hardware setup.

Metrics We use the following metrics for quantitative evaluation: 1) *SIP Error (in degrees)*: the mean global angular error of the upper arms and upper legs, focusing on joints not observed by the body-worn sensors; 2) *Angular Error (in degrees)*: the mean global angular error across all body joints; 3) *Positional Error (in centimeters)*: the mean global joint position error across all joints; and 4) *Mesh Error (in centimeters)*: the mean vertex position error between the reconstructed mesh and the ground-truth mesh given the mean body shape.

Comparison with IMU-only Methods Table 1 presents a comparison of our method against IMU-only methods on the TotalCapture and DIP-IMU datasets. On the TotalCapture dataset, our approach consistently outperforms previous methods across all metrics, improving over the SOTA PNP by 32.4% in angular error and 9.4% in mesh error. A similar trend is also observed in the results on DIP-IMU dataset. This demonstrates that inter-sensor distances serve as strong constraints for refining estimated poses.

Comparison with Distance-augmented Methods Table 2 presents a fair comparison of our method with distance-augmented methods, as all approaches use the same input data. Additionally, all methods are trained on the AMASS dataset with the TotalCapture subset excluded, whereas the IMU-only methods in Table 1 include TotalCapture in their training data. Our method outperforms UIP, achieving a reduction in positional error of 21.5% on the TotalCapture dataset and 35.0% on the DIP-IMU dataset. On the UIP dataset, where sensor data quality is low and motions are more ambiguous than in the other two test datasets, our method achieves the lowest positional error, while UIP achieves the lowest SIP error. We achieve comparable results despite using less-filtered distance measurements, a simpler architecture, and non-optimal default parameters for the state estimator. This demonstrates the robustness of our approach in handling real-world sensor noise and the effectiveness of our online state estimation framework.

Method	SIP Error (deg)	Ang Error (deg)	Pos Error (cm)	Mesh Error (cm)	SIP Error (deg)	Ang Error (deg)	Pos Error (cm)	Mesh Error (cm)
	TotalCapture				DIP-IMU			
DIP [7]	18.62	17.22	9.42	11.22	17.35	15.36	7.59	9.05
TransPose [34]	16.58	12.89	6.55	7.42	17.06	8.86	6.03	7.17
TIP [10]	13.22	12.30	5.81	6.80	16.90	9.07	5.63	6.62
PIP [35]	12.93	12.04	5.61	6.51	15.33	8.78	5.12	6.02
PNP [37]	10.89	10.45	4.74	5.45	**13.71**	8.75	4.97	5.77
UMotion	**10.76**	**7.06**	**4.46**	**4.94**	14.19	**6.35**	**3.38**	**3.93**

Table 1. Comparison with state of the art IMU-only methods on TotalCapture [29] and DIP-IMU [7].

Method	SIP Error (deg)	Pos Error (cm)	SIP Error (deg)	Pos Error (cm)	SIP Error (deg)	Pos Error (cm)
	TotalCapture		DIP-IMU		UIP	
PIP [35]	15.93	7.05	15.98	6.21	30.47	13.62
TIP-D	12.18	5.51	15.91	5.26	30.34	13.96
PIP-D	13.16	6.31	13.79	5.36	30.33	13.27
UIP [1]	11.32	5.49	**13.21**	5.05	**24.12**	10.65
UMotion	**10.76**	**4.46**	14.19	**3.38**	25.69	**10.33**

Table 2. Comparison with distance-augmented methods on TotapCapture [29], DIP-IMU [7], and UIP [1] datasets.

Metric	Mean shape	Predicted	GT
Pos Error (cm)	5.15	4.34	4.31
Mesh Error (cm)	5.63	4.81	4.78
H Error (cm)	5.75	0.39	-
W Error (kg)	9.45	0.34	-
$\boldsymbol{D}$ Error (cm)	3.63	2.33	2.30
$\boldsymbol{C}$ Error (cm)	3.78	1.26	-

Table 3. Comparison of reconstructed body mesh errors on TotalCapture [29] using mean shape versus predicted shape parameters.

Method	SIP Error (deg)	Pos Error (cm)	SIP Error (deg)	Pos Error (cm)
	TotalCapture		DIP-IMU	
$\boldsymbol{J} \rightarrow \hat{\boldsymbol{\theta}}$	9.00	4.36	14.21	3.29
$\boldsymbol{S} \rightarrow \boldsymbol{J} \rightarrow \hat{\boldsymbol{\theta}}$	12.15	5.70	17.35	4.45
$\boldsymbol{P}_G \rightarrow \hat{\boldsymbol{\theta}}$	9.25	3.85	13.63	2.85
$\boldsymbol{S} \rightarrow \boldsymbol{P}_G \rightarrow \hat{\boldsymbol{\theta}}$	11.13	4.96	15.91	3.91
$\boldsymbol{P}_R \rightarrow \hat{\boldsymbol{\theta}}$	9.89	4.42	13.77	3.22
$\boldsymbol{S} \rightarrow \boldsymbol{P}_R \rightarrow \hat{\boldsymbol{\theta}}$	11.57	5.44	15.74	4.17
Ours: $\boldsymbol{S} \rightarrow \hat{\boldsymbol{\theta}}$	**10.76**	**4.46**	**14.21**	**3.38**

Table 4. Ablation study on pose estimator architecture.

4.4. Module Evaluations

Shape Estimator Table 3 compares reconstructed body mesh on the TotalCapture dataset using the mean and predicted shape, both with the same estimated poses. With the predicted shape, positional and mesh errors are nearly as low as those obtained using the ground truth shape parameters. Height and weight estimates are accurate, with errors within 1 cm and 1 kg, respectively. However, the mean error in circumferences $\boldsymbol{C}$ of the chest, waist, hips, wrists, knees, and head remains large, as relevant measurements cannot be inferred solely from inter-sensor distances. Despite this limitation, the predicted shape still reconstructs a more realistic body than the mean shape model.

Pose Estimator We conduct an ablation study by expanding our pose estimator with intermediate layers to estimate all joint positions, $\boldsymbol{J}$, sensor global positions, $\boldsymbol{P}_G$, and sensor relative positions, $\boldsymbol{P}_R$, from sensor measurements $\boldsymbol{S} = (\hat{\boldsymbol{a}}, \boldsymbol{R}, \hat{\boldsymbol{d}})$. These intermediate estimates are then combined with $\boldsymbol{S}$ to compute the final poses $\hat{\boldsymbol{\theta}}$. This layered structure follows the framework commonly used in previous methods [1, 34]. As shown in Table 4, while regressing $\hat{\boldsymbol{\theta}}$ from ideal $\boldsymbol{J}$, $\boldsymbol{P}_G$, or $\boldsymbol{P}_R$ improves accuracy, introducing an intermediate layer, that is, $\boldsymbol{S} \rightarrow \{\boldsymbol{J}, \boldsymbol{P}_G, \boldsymbol{P}_R\}$, increases errors. This error propagates through the network, ultimately degrading pose estimation performance. This suggests that integrating IMU data with distance constraints may support a simpler architecture, reducing complexity while still maintaining accuracy.

State Estimator The state estimator is designed to: 1) mitigate errors in inter-sensor distances, 2) filter and stabi-

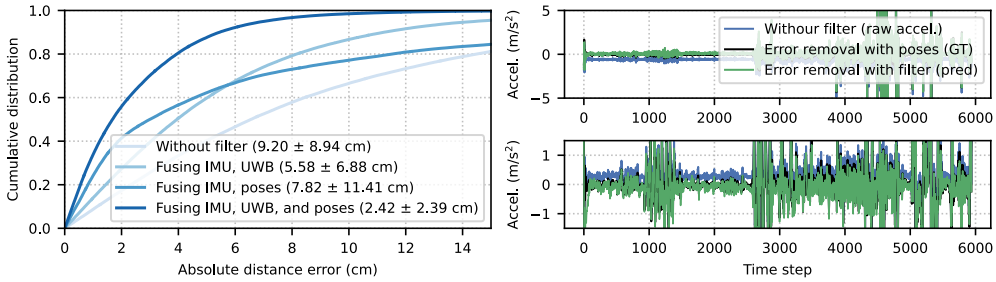


Figure 4. Cumulative distribution of distance error (left) and acceleration error reduction over time (right) for various fusion settings.

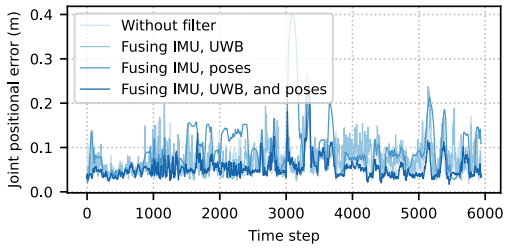


Figure 5. Joint positional error for different fusion settings.

lize accelerations, and 3) improve the accuracy of estimated poses. To examine it we conduct an ablation study by varying fusion measurements on the TotalCapture [29] with synthesized LOS-related noisy distances. Fig. 4 (left) shows cumulative distributions of absolute distance error for different fusion settings. The raw distances show an average error of 9.20 cm with a standard deviation of 8.94 cm. By fusing IMU, UWB, and pose data, this error is reduced to 2.42 cm with a standard deviation of 2.39 cm, demonstrating the effectiveness of data fusion for mitigating distance errors. Fig. 4 (right) shows the reduction in acceleration errors. The state estimator converges within a few seconds and adapts gradually, aligning with IMU characteristics. Fig. 5 presents joint positional errors over time for different fusion setups. Without filtering, positional error fluctuates with distance measurement quality. Fusing IMU with either UWB or pose data alone offers limited improvement because distance and pose outliers remain. The combined fusion of IMU, UWB, and pose data achieves the lowest positional error, validating that our state estimator refines pose accuracy by utilizing all available measurements.

4.5. Qualitative Evaluation

We developed a prototype (see supplementary material for details) that integrates BNO086 IMU and DW3000 UWB sensors to demonstrate our proposed method. Fig. 6 presents visualizations comparing pose estimates for challenging motions, illustrating the improvements of our method in disambiguating poses over IMU-only methods and enhanced pose stability compared to UIP.

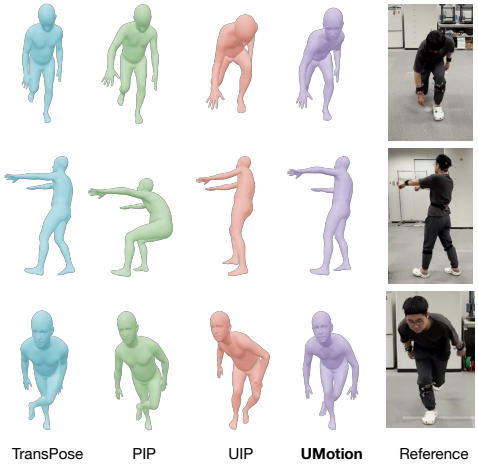


Figure 6. Qualitative comparison of pose estimates from the test data collected using our developed prototype.

5. Conclusion and Limitations

In this work, we present UMotion, an uncertainty-driven framework for 3D human shape and pose estimation that integrates six IMU and UWB sensors within a state estimation system. UMotion incorporates uncertainties from IMU, UWB, and pose estimates under individual body constraints to iteratively enhance confidence in both sensor measurements and pose accuracy. Our experiments on synthetic and real-world datasets demonstrate that UMotion outperforms existing SOTA methods in pose accuracy, and effectively stabilizes sensor measurements and pose estimation.

However, our shape estimator requires adaptation to specific sensor configurations and conditions, and, with limited training data, it may struggle with unique body variations. The state estimator also requires careful parameter tuning to reflect sensor characteristics accurately. Additionally, our pose estimator simplifies constraints, which may limit its performance. As future work, adding a physics-aware module could potentially enhance our method's robustness.

References

[1] Rayan Armani, Changlin Qian, Jiaxi Jiang, and Christian Holz. Ultra inertial poser: Scalable motion capture and tracking from sparse inertial sensors and ultra-wideband ranging. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 1, 2, 6, 7

[2] Vasileios Choutas, Lea Müller, Chun-Hao P Huang, Siyu Tang, Dimitrios Tzionas, and Michael J Black. Accurate 3d body shape regression using metric and semantic attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2718–2728, 2022. 3, 6

[3] Peng Dai, Yang Zhang, Tao Liu, Zhen Fan, Tianyuan Du, Zhuo Su, Xiaozheng Zheng, and Zeming Li. Hmd-poser: On-device real-time human motion tracking from scalable sparse observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 874–884, 2024. 1

[4] Nathan DeVrio, Vimal Mollyn, and Chris Harrison. 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*, pages 1–11, 2023. 2

[5] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*, 2020. 3

[6] S Hochreiter. Long short-term memory. *Neural Computation MIT-Press*, 1997. 4

[7] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J Black, Otmar Hilliges, and Gerard Pons-Moll. 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. 1, 2, 6, 7

[8] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *European conference on computer vision*, pages 443–460. Springer, 2022. 1

[9] Jiaxi Jiang, Paul Streli, Manuel Meier, and Christian Holz. Egoposer: Robust real-time ego-body pose estimation in large scenes. *arXiv preprint arXiv:2308.06493*, 2023. 2

[10] Yifeng Jiang, Yuting Ye, Deepak Gopinath, Jungdam Won, Alexander W Winkler, and C Karen Liu. Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 1, 2, 6, 7

[11] Simon J Julier, Jeffrey K Uhlmann, and Hugh F Durrant-Whyte. A new approach for filtering nonlinear systems. In *Proceedings of 1995 American Control Conference-ACC'95*, pages 1628–1632. IEEE, 1995. 5

[12] Manuel Kaufmann, Yi Zhao, Chengcheng Tang, Lingling Tao, Christopher Twigg, Jie Song, Robert Wang, and Otmar Hilliges. Em-pose: 3d human pose estimation from sparse electromagnetic trackers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11510–11520, 2021. 2

[13] Jiye Lee and Hanbyul Joo. Mocap everyone everywhere: Lightweight motion capture with smartwatches and a head-mounted camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1091–1100, 2024. 1, 2

[14] Sunmin Lee, Sebastian Starke, Yuting Ye, Jungdam Won, and Alexander Winkler. Questenvsim: Environment-aware simulated motion tracking from sparse sensors. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–9, 2023. 1

[15] Han Liang, Yannan He, Chengfeng Zhao, Mutian Li, Jingya Wang, Jingyi Yu, and Lan Xu. Hybridcap: Inertia-aid monocular capture of challenging human motions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1539–1548, 2023. 2

[16] Huajun Liu, Xiaolin Wei, Jinxiang Chai, Inwoo Ha, and Taehyun Rhee. Realtime human motion control with a small number of inertial sensors. In *Symposium on interactive 3D graphics and games*, pages 133–140, 2011. 2

[17] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 2

[18] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 2, 6

[19] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. Imuposer: Full-body pose estimation using imus in phones, watches, and earbuds. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2023. 1

[20] Ian Oppermann, Matti Hämäläinen, and Jari Iinatti. *UWB: theory and applications*. John Wiley & Sons, 2004. 3

[21] Shaohua Pan, Qi Ma, Xinyu Yi, Weifeng Hu, Xiong Wang, Xingkang Zhou, Jijunnan Li, and Feng Xu. Fusing monocular images and sparse imu signals for real-time human motion capture. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 2

[22] Gerard Pons-Moll, Andreas Baak, Thomas Helten, Meinard Müller, Hans-Peter Seidel, and Bodo Rosenhahn. Multisensor-fusion for 3d full-body human motion capture. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 663–670. IEEE, 2010.

[23] Jose Luis Ponton, Haoran Yun, Andreas Aristidou, Carlos Andujar, and Nuria Pelechano. Sparseposer: Real-time full-body motion reconstruction from sparse data. *ACM Transactions on Graphics*, 43(1):1–14, 2023. 2

[24] Sergi Pujades, Betty Mohler, Anne Thaler, Joachim Tesch, Naureen Mahmood, Nikolas Hesse, Heinrich H Bülthoff, and Michael J Black. The virtual caliper: rapid creation of metrically accurate avatars from 3d measurements. *IEEE transactions on visualization and computer graphics*, 25(5): 1887–1897, 2019. 3

[25] Yiming Ren, Chengfeng Zhao, Yannan He, Peishan Cong, Han Liang, Jingyi Yu, Lan Xu, and Yuexin Ma. Lidar-aid

inertial poser: Large-scale human motion capture by sparse inertial and lidar sensors. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2337–2347, 2023. 2

[26] Xu Ru, Nian Gu, Hang Shang, and Heng Zhang. Mems inertial sensor calibration technology: Current status and future trends. *Micromachines*, 13(6):879, 2022. 3

[27] Martin Schepers, Matteo Giuberti, Giovanni Bellusci, et al. Xsens mvn: Consistent tracking of human motion using inertial sensing. *Xsens Technol*, 1(8):1–8, 2018. 1

[28] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45 (11):2673–2681, 1997. 2

[29] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John P Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *BMVC*, pages 1–13. London, UK, 2017. 6, 7, 8

[30] Rudolph Van Der Merwe. *Sigma-point Kalman filters for probabilistic inference in dynamic state-space models*. Oregon Health & Science University, 2004. 5

[31] Tom Van Wouwe, Seunghwan Lee, Antoine Falisse, Scott Delp, and C Karen Liu. Diffusion inertial poser: Human motion reconstruction from arbitrary sparse imu configurations. *arXiv preprint arXiv:2308.16682*, 2023. 1

[32] Timo Von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer graphics forum*, pages 349–360. Wiley Online Library, 2017. 1, 2

[33] Eric A Wan and Rudolph Van Der Merwe. The unscented kalman filter for nonlinear estimation. In *Proceedings of the IEEE 2000 adaptive systems for signal processing, communications, and control symposium (Cat. No. 00EX373)*, pages 153–158. Ieee, 2000. 4

[34] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions On Graphics (TOG)*, 40(4):1–13, 2021. 1, 2, 6, 7

[35] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. 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*, pages 13167–13178, 2022. 1, 2, 4, 6, 7

[36] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Vladislav Golyanik, Shaohua Pan, Christian Theobalt, and Feng Xu. Egolocate: Real-time motion capture, localization, and mapping with sparse body-mounted sensors. *ACM Transactions on Graphics (TOG)*, 42(4):1–17, 2023. 2

[37] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Physical non-inertial poser (pnp): Modeling non-inertial effects in sparse-inertial human motion capture. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 1, 2, 6, 7

[38] Faheem Zafari, Athanasios Gkelias, and Kin K Leung. A survey of indoor localization systems and technologies. *IEEE Communications Surveys & Tutorials*, 21(3):2568–2599, 2019. 3

[39] Yu Zhang, Songpengcheng Xia, Lei Chu, Jiarui Yang, Qi Wu, and Ling Pei. Dynamic inertial poser (dynaip): Part-based motion dynamics learning for enhanced human pose estimation with sparse inertial sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1889–1899, 2024. 1, 2

[40] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019. 4

[41] Chengxu Zuo, Yiming Wang, Lishuang Zhan, Shihui Guo, Xinyu Yi, Feng Xu, and Yipeng Qin. Loose inertial poser: Motion capture with imu-attached loose-wear jacket. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2209–2219, 2024. 1