

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

Supplementary Material

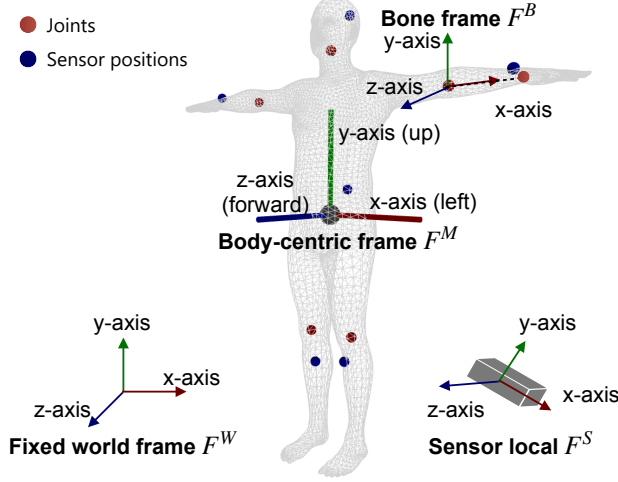


Figure 1. Overview of coordinate frames.

A. IMU-UWB Prototype

We developed a prototype integrating the off-the-shelf CEVA BNO086 9-axis IMU and Qorvo DW3000 UWB sensors on a customized board. An ESP32 microcontroller handles on-board data preprocessing and wireless transmission. The BNO086 operates at 100 Hz, using an on-board sensor fusion algorithm to output linear acceleration (gravity-removed) in the sensor’s local coordinate frame F^S and orientation relative to the initial frame. The DW3000 sensors measure 15 inter-sensor distances at an average rate of 80 Hz, with a customized asymmetric double-sided two-way ranging protocol. A time synchronization step is applied, followed by downsampling to align all measurements to 60 Hz.

B. From IMU Readings to Input Measurements

We follow the calibration procedures described in DIP [2] and TransPose [6], adapting them to suit the specific characteristics of the sensors used in our system.

Frame Definition IMU reading coordinate frame transformation is essential for aligning IMU data with the model input requirements. As shown in Fig. 1, the system operates with four types of coordinate frames:

- Sensor local coordinate frame F^S : Each sensor has its own local frame, resulting in six frames in total.

- Fixed world frame F^W : For the BNO086, the fixed world frame corresponds to the first sensor frame upon power-up. Each sensor thus has its own F^W , totaling six frames.
- SMPL Body-centric frame F^M : A single frame per person, defined as Left-Up-Forward in this work. Motions are described relative to this fixed frame, which is initialized in the T-pose at the start of the motion sequence.
- Respective bone coordinate frame F^B : Each bone with a mounted IMU has its own coordinate frame, giving six frames in total.

In total, the system consists of 19 coordinate frames: one body-centric frame, F^M , and six groups of three frames each, comprising $F^{S,i}$, $F^{W,i}$, and $F^{B,i}$, where $i \in \{1, 2, \dots, 6\}$.

Problem Statement The IMU measures linear acceleration \mathbf{a}^S in the sensor local frame F^S and orientation \mathbf{R}^{WS} , which represents the rotation matrix that transforms vectors from the sensor frame F^S to the fixed world frame F^W . When applied to an acceleration in F^S , $\mathbf{a}^W = \mathbf{R}^{WS} \mathbf{a}^S$ describes the acceleration’s representation in F^W . The inputs to the network are bone orientations relative to the body-centric frame, \mathbf{R}^{MB} , and linear accelerations in the body-centric frame, \mathbf{a}^M . \mathbf{R}^{MB} describes the rotation of each bone around the axes of the body-centric frame. These orientations also represent the global poses of the adjacent joints. To align the IMU readings with the model input, we need to transform the sensor-local accelerations \mathbf{a}^S into the body-centric frame \mathbf{a}^M , and the sensor-to-world orientation \mathbf{R}^{WS} into the bone-to-body orientation \mathbf{R}^{MB} . These transformations are expressed as:

$$\mathbf{R}^{MB} = \mathbf{R}^{MW} \mathbf{R}^{WS} \mathbf{R}^{SB}, \quad (1)$$

$$\begin{aligned} \mathbf{a}^M &= \mathbf{R}^{MS} \mathbf{a}^S \\ &= \mathbf{R}^{MW} \mathbf{R}^{WS} \mathbf{a}^S. \end{aligned} \quad (2)$$

The calibration process aims to determine \mathbf{R}^{MW} and \mathbf{R}^{SB} to enable these transformations.

Calculation of \mathbf{R}^{MW} As shown in Fig. 1, the body-centric frame F^M is established as the Left-Up-Forward orientation of the initial T-pose at the start of the motion. The fixed world frame of the BNO086 is defined as the first sensor frame after power-up. To ensure consistency, we position all IMUs in the same initial orientation, aligning their initial sensor frames such that $F_{\text{init}}^{S,1} = \dots = F_{\text{init}}^{S,6} = F^{W,1} = F^{W,2} = \dots = F^{W,6}$. To simplify computation,

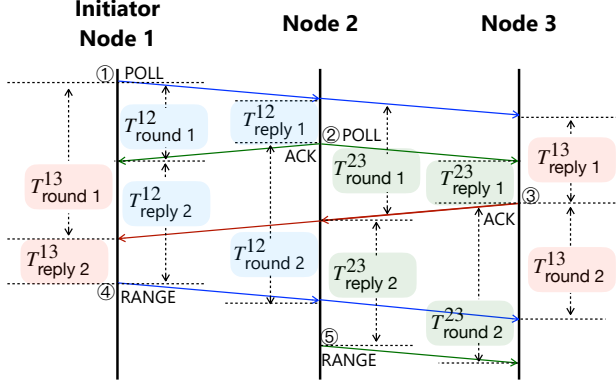


Figure 2. Ranging transaction with three devices. Timestamps to resolve time-of-flight are included in the UWB message payload and thus broadcast to all network participants.

we align the axes of the F_{init}^S and F^W with the corresponding axes of F^M or use a known transformation. For example, we position the IMU with its x-axis pointing left, y-axis pointing up, and z-axis pointing forward in the real world. This alignment defines R^{MW} . In cases where F^W is aligned with F^M , $R^{MW} = I$.

Calculation of R^{SB} Next, we mount IMUs onto the corresponding body part in arbitrary orientations. The subject is then instructed to remain still in a T-pose for several seconds. In this pose, the orientation of bone frame relative to the SMPL body-centric frame is zero, meaning $R_{\text{T-pose}}^{MB} = I$. Thus, given the measured average orientation of the IMU in T-pose, $\bar{R}_{\text{T-pose}}^{WS}$, we have

$$R_{\text{T-pose}}^{MB} = R^{MW} \bar{R}_{\text{T-pose}}^{WS} R^{SB}, \quad (3)$$

$$R^{SB} = \text{inv}(R^{MW} \bar{R}_{\text{T-pose}}^{WS}) R_{\text{T-pose}}^{MB}, \quad (4)$$

$$R^{SB} = \text{inv}(\bar{R}_{\text{T-pose}}^{WS}). \quad (5)$$

C. Ranging Protocol

We implemented an efficient distance matrix ranging method based on asymmetric double-sided two-way ranging (ADS-TWR) protocol [4]. Compared to the standard two-way ranging protocol, ADS-TWR minimizes the impact of clock drift and synchronization errors. Fig. 2 illustrates an example with three sensors. One sensor is designated as the initiator and transmits a POLL signal. Subsequently, other sensors sequentially act as transmitters, sending POLL signals to the remaining sensors after receiving POLL signals from all preceding sensors in order. These POLL signals simultaneously serve as ACK signals for the previous sensors, streamlining communication. This efficient broadcasting strategy reduces the number of transmitted signals from 45 (calculated as 15 pairs, each requiring

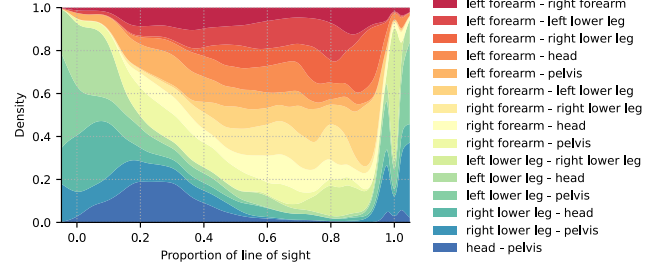


Figure 3. Stacked density plot showing the proportion relative to the total distribution of LOS availability for inter-sensor distances across different sensor pairs.

3 transmissions) to just 11. A sequence of timestamps is recorded during this process to measure the time-of-flight (ToF), T , between sensor pairs. T is determined using the formula:

$$T = \frac{T_{\text{round 1}} \times T_{\text{round 2}} - T_{\text{reply 1}} \times T_{\text{reply 2}}}{T_{\text{round 1}} + T_{\text{round 2}} + T_{\text{reply 1}} + T_{\text{reply 2}}}. \quad (6)$$

The corresponding distance, d , between the sensor pairs is then calculated as:

$$d = cT, \quad (7)$$

where c represents the speed of light in vacuum.

D. Line of Sight Simulation

One challenge in using body-worn UWB sensors for tracking inter-sensor distances is body occlusion, which degrades measurement accuracy [1]. To address this, we simulate line-of-sight (LOS) conditions to learn the distribution of the occlusion on TotalCapture dataset [5]. The simulation utilizes the SMPL body model [3] to calculate LOS and non-line-of-sight (NLOS) conditions based on different poses. The visibility of each sensor pair is determined by tracing straight-line paths between them and checking for intersections with the body mesh. We employ the Möller-Trumbore intersection algorithm to identify these intersections. The LOS proportion is then calculated as the total length of unobstructed (LOS) segments divided by the entire distance.

Fig. 3 shows a stacked density plot of LOS proportions across 15 sensor pairs, representing the relative contribution of each sensor pair to the total distribution of LOS proportions. For a given LOS proportion, the stacked regions indicate how frequently different sensor pairs contribute to that proportion. It reveals that pairs such as “lower leg - pelvis” and “lower leg - head” exhibit consistently low LOS availability due to frequent occlusion caused by body movement and overlapping limbs. Accordingly, the corresponding distance measurements are unreliable and could not be effectively used for pose estimation or measurement filtering. This analysis highlights the varying reliability of UWB

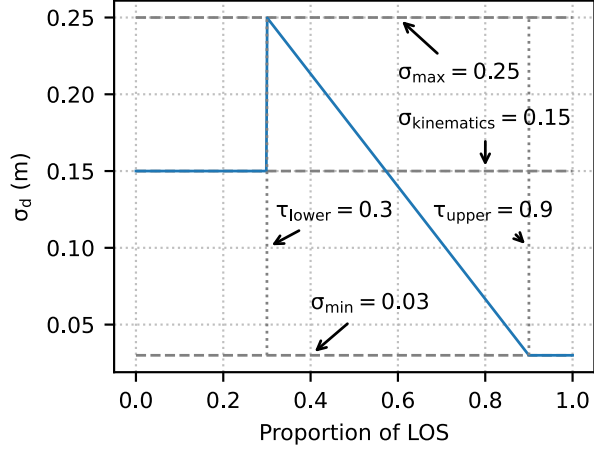


Figure 4. Example of the distance error model based on the LOS proportions for the sensors used in our system.

measurements across sensor pairs, offering guidelines for weighting measurement uncertainties in our state estimation framework.

Distance Error Model In this work, we simplify the standard deviation of distance measurements, σ_d , as a function of the LOS proportion, l , as follows:

$$\sigma_d = \begin{cases} \sigma_{\min}, & \text{if } l \geq \tau_{\text{upper}}, \\ \sigma_{\text{kinematics}}, & \text{if } l < \tau_{\text{lower}}, \\ (\sigma_{\max} - \sigma_{\min}) \frac{(\tau_{\text{upper}} - l)}{\tau_{\text{upper}} - \tau_{\text{lower}}} + \sigma_{\min}, & \text{otherwise,} \end{cases} \quad (8)$$

where τ_{upper} and τ_{lower} are LOS proportion thresholds, and σ_{\min} and σ_{\max} represent the minimum and maximum noise parameters for the distance standard deviation. When the LOS proportion falls below τ_{lower} , the distance measurement is replaced with one derived from kinematics, with an associated standard deviation of $\sigma_{\text{kinematics}}$. Fig. 4 provides an example of this model based on our selected sensors. The parameters may vary depending on the specific sensors used.

E. Discussions on Predicted Uncertainty

To assess the correctness of the predicted uncertainty, we analyze the transformed axis-wise relative position error distributions. We calculate distance errors given predicted poses and compare them with the distance uncertainty into which the predicted pose uncertainty is converted. Fig. 5 shows the proportion of frame counts within different confidence intervals. The results indicate that the predicted uncertainty aligns well with actual errors for smaller deviations, with 85% of predictions falling within 3σ . However, for larger errors, the predicted uncertainty tends to be underestimated.

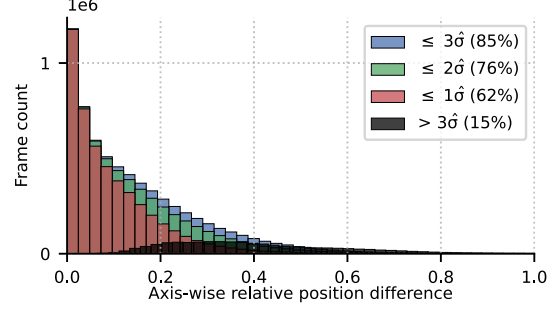


Figure 5. Histogram of axis-wise relative position differences, illustrating the alignment between predicted uncertainty and true errors.

F. Implementation Details

We train the pose estimator using synthesized data from the AMASS dataset without integrating the state estimator. We apply noise only to the synthesized distances, while synthesized IMU data remains noise-free. In the state estimator, the process noise covariance \mathbf{Q} is determined using Allan variance analysis with a noise propagation model. The observation noise covariance \mathbf{R}_1 follows our distance error model, while \mathbf{R}_3 is derived from predicted poses via the unscented transformation. To mitigate overconfidence in high-error scenarios, we scale \mathbf{R}_3 by a factor of 10 for improved stability.

G. Ablation on Shape Estimator

To evaluate the impact of different anthropometric data on shape estimation, we conduct an ablation study using the TotalCapture dataset. Table 1 presents the mean absolute error of the reconstructed T-pose mesh under different subsets of anthropometric inputs. Since circumferences are not directly observed, their errors remain the highest across all conditions. Using only height (H) or weight (W) results in relatively large distance and mesh errors, demonstrating that these individual measurements alone do not sufficiently

	Mesh (mm)	Mean absolute error			
		H (mm)	W (kg)	D (mm)	C (mm)
H	12.10	1.11	3.77	10.62	21.08
W	23.45	58.70	0.28	31.69	16.11
D	6.14	2.67	4.47	0.9	22.62
HW	10.40	1.2	0.19	11.30	13.34
HD	6.30	1.83	4.10	1.34	21.08
WD	4.31	3.37	1.03	1.14	13.26
HWD	4.72	3.89	0.35	2.09	12.76

Table 1. Comparison of reconstructed T-pose mesh errors on TotalCapture [5] using different sources of anthropometric data.

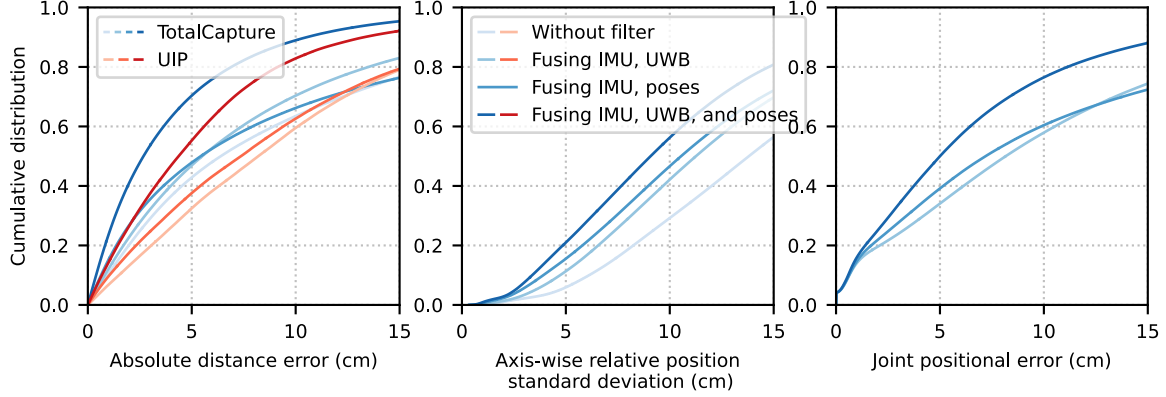


Figure 6. Cumulative distribution of distance error (left), predicted relative position standard deviation (middle), and joint positional error (right) for various fusion settings.

constrain body shape. Combining height and weight (HW) improves shape estimation, leading to slight reductions in mesh errors. Incorporating inter-sensor distances (D) provides better constraints on body proportions, further reducing mesh and distance errors.

H. Ablation on State Estimator

We compare absolute distance error, predicted uncertainty, and joint positional error across various configurations on TotalCapture and UIP datasets to evaluate the impact of different fusion strategies. Fig. 6 (left) illustrates the cumulative distribution of absolute distance errors. Incorporating IMU and UWB fusion reduces distance errors, and the addition of pose information further improves accuracy. This demonstrates that integrating multiple sensing modalities enhances distance estimation by leveraging complementary information. Fig. 6 (middle) shows the axis-wise relative position standard deviations, evaluating the effect of different information on the predicted uncertainty. The results indicate that the full fusion model, i.e., IMU, UWB, and poses, improves the consistency of uncertainty estimation, resulting in the most confident predictions. Fig. 6 (right) evaluates the cumulative distribution of joint positional errors. Compared to the unfiltered case, fusing IMU and UWB data reduces error, while incorporating pose constraints further improves tracking performance. These results demonstrate that jointly fusing IMU, UWB, and pose constraints improves distance accuracy, refines uncertainty estimation, and reduces joint positional errors.

References

[1] Rayan Armani, Changlin Qian, Jiayi 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. 2

[2] 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

[3] 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

[4] Michael McLaughlin and Billy Verso. Asymmetric double-sided two-way ranging in an ultrawideband communication system, 2019. US Patent 10,488,509. 2

[5] 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. 2, 3

[6] 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