Seeing Around the Corner: Fusing Visual Flow and Inertial Sensors for Indoor Pedestrian Navigation

Noah Meißner*, Tim Sieber*, and Bernd Ludwig*

*Chair of Information Science, University of Regensburg

Abstract

Outdoors, a quick look at a smartphone is enough to find your way around; indoors, this convenience disappears. Pedestrian dead reckoning based on inertial sensors fails precisely where pedestrians need it most — at directional changes — and progress is slowed down as there is too little training data in the real world to train complex models. In our indoor navigation system, URWalking (described in (Ludwig et al. 2023)), we provide routing instructions to many users per day for routes of 500 metres or more. An analysis of the implemented tracking module in (Jackermeier and Ludwig 2018) revealed a 50% drop in PDR accuracy at turns in tight corridors. To improve performance, we collected new multi-modal training data and applied a standard optical flow algorithm to improve turn detection. Our experiments demonstrate that we can predict turn-taking with 90% accuracy. We consider this result to be a significant step forward in improving the long-distance tracking capabilities of indoor navigation systems.

Keywords: Indoor Positioning, Pedestrian Dead Reckoning, Spatial Grounding, Optical Flow, Multimodal Dataset, Ground Truth Generation

Introduction

A quick look at your smartphone is usually enough to determine your location outdoors. However, this convenience ends the moment you step inside a building. With satellite-based methods and standard GPS unavailable, accurately locating pedestrians indoors over distances of 500m or more remains problematic to this day. This limitation is reflected in recent research: (Khedr and El-Sheimy 2021) report maximum distances of 293 m for tracking persons with smartphones and no other sensors available. For scaling tracking to real-world routes in huge indoor complexes solutions to break these limitations have to be developed.

In previous research (Jackermeier and Ludwig 2018), we identified turn-taking as a major cause of tracking errors. While on straight segments of long-distance tracks, IMU sensor drift could be

Published in "Proceedings of the 1st International Conference on Geospatial Artificial Intelligence (GeoAI 2026) – Oral Presentation Papers", edited by Haosheng Huang and Nico Van de Weghe, GeoAI 2026, 3-6 June 2026, Ghent, Belgium.

This contribution underwent single-blind peer review based on the extended abstract.

compensated for using data from other sensors. However, the localisation score drops by around 50% in turn segments compared to straight sections. Therefore, current methods break down on turns, not straight segments, on an 182 m route.

To mitigate these limitations, the field has increasingly relied on sensor fusion, combining Wi-Fi, IMU, magnetic field, and visual signals (Ashraf et al. 2021). However, their effectiveness depends critically on the availability of suitable training data: robust fusion approaches require extensive, precisely annotated datasets (De-La-Llana-Calvo et al. 2019). While publicly available datasets do exist (Shu et al. 2021; Ashraf et al. 2021), high-precision ground truth (GT) data has predominantly been collected using dedicated acquisition pipelines, such as permanently installed motion capture systems (Schubert et al. 2018), post-processing registration against laser scans (Sarlin et al. 2022), or specialised wearable sensor configurations such as foot-mounted IMUs with zero-velocity updates (Foxlin 2005). While these methods provide precise references, using them for data recording is complex to organise. Furthermore, the hardware requirements prevent the approaches from being usable in everyday situations. Finally, they are scalable to a limited extent only (Li et al. 2024; Badawy and Ziedan 2025; Sarlin et al. 2022).

To address the issue of available realistic data for indoor tracking, recent work has explored AR hardware for data collection. LaMAR (Sarlin et al. 2022) pairs HoloLens 2 with smartphones, but its ground truth is established offline through registration against dedicated laser scans. InCrowd-VI (Bamdad, Hutter, and Darvishy 2024), in contrast, uses Meta Aria Glasses as the sole recording platform, relying on the headset's internal SLAM service as ground truth. What is missing is a setup for off-the-shelf hardware that combines smartphone sensing with an external 6DoF reference without offline registration. To address this gap, we recorded a multimodal dataset of 27 trajectories from 10 participants and will continue to do so in the future. We used this data to investigate whether optical flow could be used as a signal for turn-taking when following an indoor route. Our results are encouraging as we achieve 90% accuracy in post-hoc turn classification and significantly improve the IMU approach in Jackermeier and Ludwig 2018. So, our contribution is twofold:

- We make data from our dataset described above accessible to the research community.
- We provide empirical evidence that optical flow can effectively improve tracking persons in indoor environments when they are equipped with standard smartphones only.

Methodology

Apparatus

For data recording, we rely on off-the-shelf hardware: a Microsoft HoloLens 2 paired with a standard smartphone yields synchronised video, IMU, and 6DoF pose without dedicated infrastructure, validated by a closed-loop error of 0.13% over 399 m. By anchoring every smartphone sample to the HoloLens pose and the building floor plan, otherwise ambiguous sensor signals become interpretable against architectural structure.

Data Recording and Validation

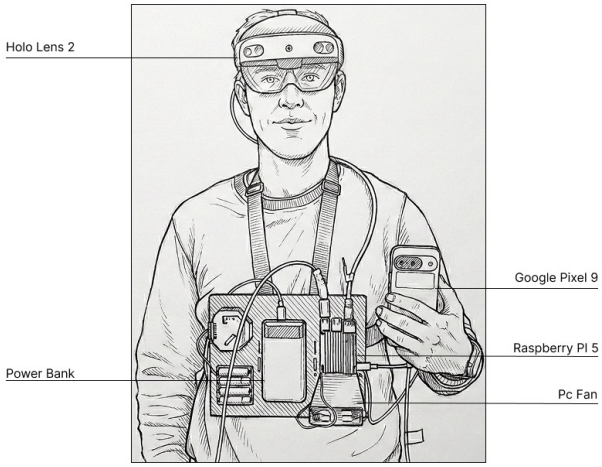

Figure 1: Recording system: Microsoft HoloLens 2 (worn on head) and Google Pixel 9 (held upright at chest height, rear camera facing forward).

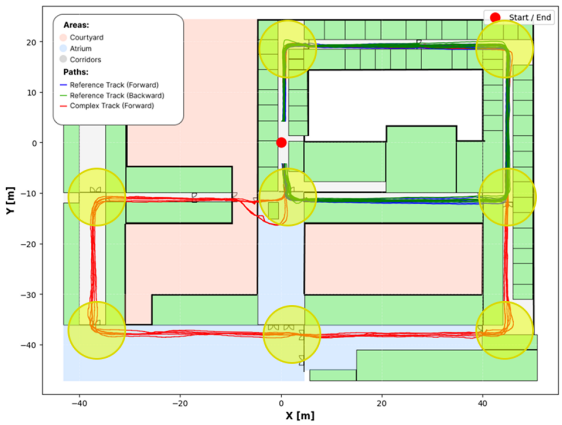

Figure 2: Recorded Trajectories

The recording system (shown in Fig. 1) combines a Microsoft HoloLens 2 with a Google Pixel 9 as the target navigation device, which is held upright at chest height (rear camera facing the direction of view, approx. 30–40 cm away from the HoloLens). The Raspberry Pi 5, power bank and PC fan are mounted on a chest plate; the Pi serves as a central time server, which stamps all incoming sensor data with a shared timestamp via UDP and stores it (latency: 7.95 ± 0.24 ms at 30 Hz). The HoloLens 2 provides continuous 6DoF pose estimates via its World-Locking SLAM system, which serve directly as ground truth without any environmental modifications.

While the on-device tracking of the HoloLens 2 is known to accumulate drift in larger scenes and requires laser-scanner corrections to achieve absolute centimeter-level accuracy for AR applications (Sarlin et al. 2022), we validate the GT precision on our tracks using the closed-loop method: the mean loop closure error (LCE) was $0.05\,\%$ (≈ 8.8 cm) on the reference track (176 m; 182 m nominal corridor length) and $0.13\,\%$ (≈ 52 cm) on the complex track (399 m). Unlike absolute pose accuracy required for holographic AR rendering, PDR-relevant metrics such as step length and directional consistency do not require sub-centimetre ground truth (Elyasi and Manduchi 2023). When tracking pedestrians, a step is the basic unit of movement, with an expected length of around 80 cm. Therefore, the measured LCE is sufficient for PDR-relevant trajectory annotation.

Dataset

The recordings cover three structurally distinct zone types within a building at the University of Regensburg (Fig. 2): narrow office corridors with uniform artificial lighting (*Confined Space*), a low-texture atrium with diffuse daylight (*Open Space*) and a section of the inner courtyard with abrupt changes in light and texture (*Transition Zone*). This combination deliberately exposes the system to conditions that challenge both inertial and vision-based methods.

Data collection
Ten participants ($M_{\text{Age}} = 23.7$ years) completed a bidirectional reference route (176 m; identical to (Jackermeier and Ludwig 2018)) as well as a complex route (399 m) across all three zone types. The dataset comprises 27 trajectories (10 forward, 9 backward, 8 complex) with a frame validity of more than 99 %. As an initial release, we are making a complete recording session available[1]. An expanded dataset covering various environments is under preparation.

Curve classification

A pedestrian turning off the path creates a systematic lateral shift in the camera's field of view, which manifests as a directional trend in the horizontal optical flow. In our experiments, decisions are made exclusively within spatially defined turning zones (*Spatial Gating*, here derived from ground-truth position) in order to exclude noise from straight sections.

Generation of Turn-Taking Data
We use the HoloLens location data to assign ground truth positions to the video data recorded with the smartphone (see Fig. 1). So at each turning zone z along the route (see Fig. 2), we filter a set F_z of video frames from the recorded stream. A frame is added to F_z if its position is within a fixed radius r around the center of the turning zone. For F_z we can determine the ground truth for the walking direction from the routing instruction for z and assign the same label to all frames in F_z. Valid labels are (*Left*, *Right*, *Straight ahead*).

The task now is to predict these labels from optical flow calculated from the frames in F_z.

Optical flow signal

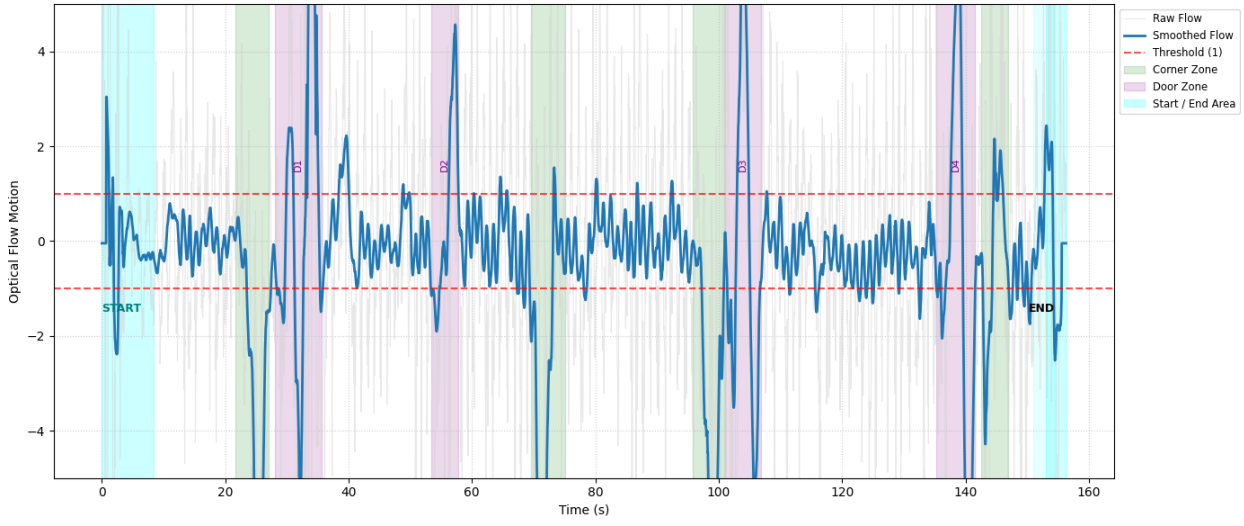

Figure 3: Optical flow signal along a trajectory. Distinct peaks in the curved sections (coloured areas) stand out clearly from the noise in the straight sections.

For each frame f, Lucas-Kanade optical flow (LK) is computed (Lucas and Kanade 1981): Shi-Tomasi keypoints (Shi et al. 1994) (max. 50) are tracked into the subsequent frame using pyramid-

1. https://github.com/URWalking/Optical-Flow-Analysis

based LK; the median of the horizontal displacements yields a robust scalar flow value Δx_f. The vertical component is discarded, as it primarily encodes the step pattern. Δx_f is smoothed using a centred moving average over 45 frames (≈ 1.5 s) and corrected by a global bias b (mean optical flow for walking straight ahead). In this way, for each $f \in F_z$ we obtain a value o_f for the smoothed optical flow (Fig. 3) as the only source of information.

Region-based scoring and classification
To predict the turn taking that actually took place, we calculate the zone average of o_f:

$$\text{signal} = \frac{1}{|F_z|} \sum_{f \in F_z} o_f$$

For the final prediction, we apply a simple rule based on a learnable threshold t:

$$\text{turn}(z) = \begin{cases} \texttt{Straight ahead} & -t < \text{signal} < t \\ \texttt{Left} & \text{signal} > t \\ \texttt{Right} & \text{signal} < -t \end{cases}$$

Positive values for signal correspond to left-hand bends, as the camera then pans to the right in the world view.

Hyperparameter optimisation and evaluation
r and t are optimised using grid search $r \in \{2.5, \ldots, 6.5\}$ m; $t \in \{0.5, \ldots, 2.0\}$ by means of Leave-One-Route-Type-Out-CV (LORTO-CV): One route type (*Complex*, *Forward*, *Backward*) serves as the test set, the remaining two for hyperparameter selection. Optimal configuration: $r = 4.5$ m, $t = 1.0$. Since the approach does not learn specific parameters, LORTO-CV is the methodologically appropriate strategy: it tests generalisation to unseen spatial structures; data leakage between participants is structurally precluded.

Results and Discussion

With the optimal configuration (r = 4.5 m, t = 1.0), the system achieves a test accuracy of 90% (Table 1) on the track described in (Jackermeier and Ludwig 2018), making this our primary point of comparison with that work. However, since the forward (4 right turns), backward (4 left turns), and complex tracks (5 right, 1 left, 2 straight) differ substantially in geometry, the resulting class distribution is inherently unbalanced, and aggregate metrics should be interpreted with care. The

Fold	Route	N	Acc	F1 Left	F1 Straight	F1 Right
1	Complex	56	0.93	0.77	0.87	0.99
2	Ref (Fwd)	40	0.9	-	-	0.95
3	Ref (Back)	36	**1.000**	**1.0**	-	-

Table 1: LORTO-CV results ($r = 4.5$ m, $t = 1.0$). Folds correspond to the route types shown in Fig. 2.

high accuracies we obtained on the reference routes (folds 2 and 3) demonstrate the reliability of the proposed method in corridors. As such architectural structures occur frequently in indoor buildings, the approach can contribute to improve tracking of smartphones over long distances.

Comparison with the inertial reference frame method

While Jackermeier and Ludwig 2018 report area match scores between 0.45 and 0.68 in turning zones, implying a drop in accuracy of around 50% compared to the score for preceding areas, our results (overall accuracy of 90%) indicate that this drop can be eliminated almost completely using optical flow.

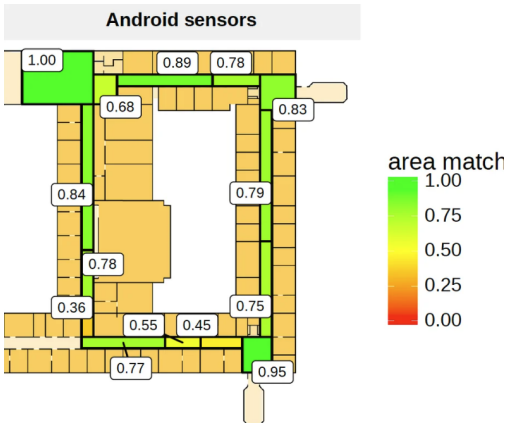

Figure 4: Prediction accuracies for turn taking as reported in Jackermeier and Ludwig 2018

Conclusion

Our experiments indicate that analysing the optical flow of smartphone video data at turn-taking points significantly improves the tracking of people carrying smartphones, even without other sensors. As our approach does not require specific training, prior mapping or other prerequisites, it can be applied to environments other than the one in which we collected our data. This result could contribute to stabilising indoor tracking in smartphone-only setups over very long distances. Indeed, we conclude that similar specialised detectors for human walking behaviour (e.g. walking up or down stairs, taking lifts, moving between indoor and outdoor areas) could further improve indoor tracking under realistic conditions. Integrating such detectors into a live PDR pipeline poses the challenge of automatically determining whether a person is approaching a corner, door, elevator, or stairs. To address this challenge and to train reliable models, the important next step in future work is to construct a large corpus of multimodal sensor data annotated with ground truth information about the true location and the context of the location, in order to align the predictions from the sensor data with the environment, for example using visual (semantic) landmarks (such as doorways) or information taken from the map of the underlying navigation system.

References

Ashraf, Imran, Sadia Din, Muhammad Usman Ali, Soojung Hur, Yousaf Bin Zikria, and Yongwan Park. 2021. "MagWi: Benchmark dataset for long term magnetic field and Wi-Fi data involving heterogeneous smartphones, multiple orientations, spatial diversity and multi-floor buildings." *IEEE Access* 9:77976–77996.

Badawy, Tamer Ramadan, and Nesreen I Ziedan. 2025. "Optimized intelligent localization through mathematical modeling and crow search algorithms." *Sensors* 25 (15): 4804.

Bamdad, Marziyeh, Hans-Peter Hutter, and Alireza Darvishy. 2024. "InCrowd-VI: A Realistic Visual–Inertial Dataset for Evaluating Simultaneous Localization and Mapping in Indoor Pedestrian-Rich Spaces for Human Navigation." *Sensors* 24 (24): 8164.

Elyasi, Fatemeh, and Roberto Manduchi. 2023. “Step length is a more reliable measurement than walking speed for pedestrian dead-reckoning.” In *2023 13th International Conference on Indoor Positioning and Indoor Navigation (IPIN),* 1–6. IEEE.

Foxlin, Eric. 2005. “Pedestrian tracking with shoe-mounted inertial sensors.” *IEEE Computer graphics and applications* 25 (6): 38–46.

Jackermeier, Robert, and Bernd Ludwig. 2018. “Exploring the limits of PDR-based indoor localisation systems under realistic conditions.” *Journal of Location Based Services* 12 (3-4): 231–272.

Khedr, Maan, and Naser El-Sheimy. 2021. “S-PDR: SBAUPT-Based Pedestrian Dead Reckoning Algorithm for Free-Moving Handheld Devices.” *Geomatics* 1 (2): 148–176. ISSN: 2673-7418. https://doi.org/10.3390/geomatics1020010. https://www.mdpi.com/2673-7418/1/2/10.

De-La-Llana-Calvo, Álvaro, José-Luis Lázaro-Galilea, Alfredo Gardel-Vicente, David Rodríguez-Navarro, Ignacio Bravo-Muñoz, and Felipe Espinosa-Zapata. 2019. “Characterization of multipath effects in indoor positioning systems by AoA and PoA based on optical signals.” *Sensors* 19 (4): 917.

Li, Jiayi, Yinhao Song, Zhiliang Ma, Yu Liu, and Cheng Chen. 2024. “A Review of Indoor Localization Methods Leveraging Smartphone Sensors and Spatial Context.” *Sensors* 24 (21): 6956.

Lucas, Bruce D, and Takeo Kanade. 1981. “An iterative image registration technique with an application to stereo vision.” In *IJCAI’81: 7th international joint conference on Artificial intelligence,* 2:674–679.

Ludwig, Bernd, Gregor Donabauer, Dominik Ramsauer, and Karema al Subari. 2023. “URWalking: Indoor Navigation for Research and Daily Use.” *KI - Künstliche Intelligenz* 37 (1): 83–90. https://doi.org/10.1007/s13218-022-00795-1.

Sarlin, Paul-Edouard, Mihai Dusmanu, Johannes L Schönberger, Pablo Speciale, Lukas Gruber, Viktor Larsson, Ondrej Miksik, and Marc Pollefeys. 2022. “Lamar: Benchmarking localization and mapping for augmented reality.” In *European Conference on Computer Vision,* 686–704. Springer.

Schubert, David, Thore Goll, Nikolaus Demmel, Vladyslav Usenko, Jörg Stückler, and Daniel Cremers. 2018. “The TUM VI benchmark for evaluating visual-inertial odometry.” In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),* 1680–1687. IEEE.

Shi, Jianbo, et al. 1994. “Good features to track.” In *1994 Proceedings of IEEE conference on computer vision and pattern recognition,* 593–600. IEEE.

Shu, Yuanchao, Qiang Xu, Jie Liu, Romit Roy Choudhury, Niki Trigoni, and Victor Bahl. 2021. *Indoor Location Competition 2.0 Dataset,* January. https://www.microsoft.com/en-us/research/publication/indoor-location-competition-2-0-dataset/.