Fine-grained Stop-Move Detection with UWB: Quality Metrics and Real-world Evaluation

FATIMA HACHEM, University of Milano, Italy
DAVIDE VECCHIA, University of Trento, Italy
MARIA LUISA DAMIANI, University of Milano, Italy
GIAN PIETRO PICCO, University of Trento, Italy

The synergy between the accurate trajectories offered by ultra-wideband (UWB) systems and techniques to extract higher-level mobility patterns is largely unexplored. We study whether staple techniques designed for systems with coarser resolution apply to UWB, investigating *quantitatively* the quality of the fine-grained analyses enabled by the latter. To this end, we contribute a *novel family of metrics* suited to the high UWB spatio-temporal resolution and use them to configure and ascertain the quality of representative techniques along several dimensions. We focus on the well-known stop-move pattern and derive our findings from a real museum setting with the use case of capturing visits to exhibits. We acquire UWB trajectories in both controlled (*in vitro*) and uncontrolled (*in vivo*) conditions, along with ground truth. Despite exhibits being very close to each other, our results show that stops near them can be correctly identified and associated in the vast majority of cases and with very small spatio-temporal error. These positive results from real-world experiments, along with our technical contributions, open new opportunities in exploiting UWB for mobility analyses.

CCS Concepts: • **Networks → Location based services**; • **Human-centered computing → Ubiquitous and mobile computing**.

Additional Key Words and Phrases: Ultra-wideband (UWB), trajectory, mobility pattern, stop-move detection

ACM Reference Format:
Fatima Hachem, Davide Vecchia, Maria Luisa Damiani, and Gian Pietro Picco. 2025. Fine-grained Stop-Move Detection with UWB: Quality Metrics and Real-world Evaluation. *ACM Trans. Sensor Netw.* 1, 1 (May 2025), 33 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The pervasiveness of positioning technologies facilitates the collection of *spatial trajectories* [41, 49] sampling the movement of an entity (e.g., a person, animal, or vehicle) in a target space. These trajectories are sequences $T = \langle (\mathbf{p}_1, t_1) \dots (\mathbf{p}_n, t_n) \rangle$ of *units* $(\mathbf{p}_i, t_i)$ associating a position $\mathbf{p}_i = (x_i, y_i)$ of the entity with the corresponding timestamp t_i. Once efficiently processed [14, 15, 51], spatial trajectories enable the extraction of *mobility patterns* describing the behavior of individuals at a higher level of abstraction [34, 47]. In turn, these form the basis of *semantic trajectories* [27, 34, 41, 43], further enriching mobility patterns with contextual information (e.g., points of interest, POIs) directly providing actionable information to domain experts.

Authors' addresses: Fatme Hachem and Maria Luisa Damiani: {fatme.hachem, maria.damiani}@unimi.it, University of Milano, Dept. of Computer Science, v. Celoria 18, Milano, Italy, 20133. Davide Vecchia and Gian Pietro Picco: {davide.vecchia, gianpietro.picco}@unitn.it, University of Trento, Dept. of Information Engineering and Computer Science (DISI), v. Sommarive 9, Trento, Italy, 38122.

1550-4859/2025/5-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Stops and their detection. In this respect, the stop-move [36] pattern is one of the most popular, key to many applications. Broadly, we define a *stop event* as the abstraction capturing the permanence of an entity in a representative *position* for a time interval determining the stop *duration*; a *move event* captures instead a transition between stops. For instance, a stop event may capture the dwelling of a person in front of an item (POI) in a store, and a move event the travel to reach the item from the previous stop. Stop-move detection techniques extract each stop and move event in a trajectory T as a sub-sequence $U = \langle (\mathbf{p}_i, t_i) \dots (\mathbf{p}_{i+k}, t_{i+k}) \rangle \subseteq T$, called *segment*. As these techniques partition a trajectory into alternating, temporally disjoint stop and move segments, they are often referred to as stop-move *segmentation* techniques [4, 47].

In this paper, we focus our attention only on stop events, extracted from the corresponding segments. Given a stop segment U identified by the stop detection technique (§2), the corresponding stop event is captured as $s = \langle \mathbf{p}, [t_{start}, t_{end}] \rangle$; the position $\mathbf{p}$ is computed as the centroid of the set of positions $\{\mathbf{p}_i \dots \mathbf{p}_{i+k}\}$ belonging to the units in U, while its interval is the one extending from the first unit timestamp in U until the last one, i.e., $[t_{start}, t_{end}] = [t_i, t_{i+k}]$ in our example.

Moreover, in many applications, a set $O = \{o_1, \dots, o_n\}$ of *spatial objects* representing POIs is known a priori and must be associated to stops, e.g., to capture which museum exhibit a visitor has stopped in front of. This application-dependent, contextual information is captured by augmenting stops with a *semantic annotation* expressing the association of a stop s to a spatial object o. A stop event becomes then a triple $s = \langle \mathbf{p}, [t_{start}, t_{end}], o \rangle$, where the spatial object $o = \langle l, \mathbf{q} \rangle$ is denoted by a label l and position $\mathbf{q}$. A sequence of these stop events constitutes a *semantic trajectory*, enabling domain experts to analyze data directly at their level of abstraction.

Existing techniques for stop-move detection focus on large-scale, outdoor settings germane to Global Navigation Satellite Systems (GNSS) where their application is widespread but yields trajectories with coarse spatio-temporal resolution. Stops typically represent the home range of migrating animals, lasting months over a large area [9], or capture the visits of people to relevant places (e.g., home, workplace, or tourist locations), lasting hours [34].

Ultra-wideband. Still, GNSS are not the only option. A recent technological wave targets sub-meter position accuracy, a powerful enabler in several applications notably including indoor ones [45, 46]. Leading this wave, ultra-wideband (UWB) radios enable communication *and* accurate localization. UWB signals are characterized by a bandwidth ≥500 MHz or a fractional bandwidth ≥20% during transmission. In modern impulse radio (IR-UWB) systems, this large bandwidth is achieved via narrow pulses (≤2 ns) that better separate the signal from multipath components and, crucially, provides excellent time resolution (Fig. 1). This enables UWB devices to precisely estimate the signal time of flight τ and, after multiplying by the speed of light c, estimate their distance $d = c\tau$ with decimeter-level accuracy, significantly better than the meter-level one typically offered by narrowband radios like WiFi and Bluetooth [45]. The increasing role of UWB is witnessed by the many Real-Time Location Systems (RTLS) based on it and its recent inclusion in smartphones where, however, public APIs currently focus on device-to-device proximity rather than localization.

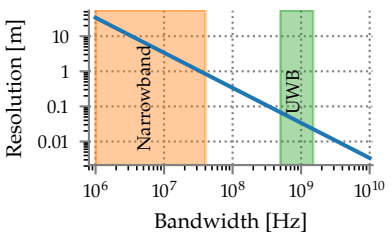


Fig. 1. Distance resolution vs. bandwidth.

Motivation and research questions. We observe an obvious synergy between mobility pattern analyses and UWB-based positioning. Compared to GNSS, UWB offers trajectories with much higher spatial resolution and temporal frequency, in principle enabling *fine-grained* analyses capturing stop events a few decimeters apart and lasting only a few seconds—a far cry from the large areas (m to km) and durations (hours to months) targeted by current approaches.

Unfortunately, mobility analyses exploiting UWB trajectories are largely unexplored and stop-move detection is no exception. This leaves unanswered the two crucial and intertwined questions we address in this work:

i) Are existing stop-move techniques applicable to UWB trajectories, and
ii) to what quantitative *extent does this synergy enable accurate detection with* fine-grained *spatio-temporal resolution?*

A positive answer to the first question would enable the use of the ample literature using GNSS positioning to exploit the new opportunities offered by UWB. Although we focus on representative techniques for stop-move detection (§2), our findings could inspire advancements for other techniques or mobility patterns. Still, this question cannot be separated from the second one about *quantifying* the benefits of UWB and exploring its limits. Indeed, the increased spatio-temporal density of UWB positions, along with errors induced by indoor environments, could clash with the mechanics of existing techniques, ultimately severely limiting the expected fine-grained quality of their output.

Real-world use case and deployment. These questions cannot be investigated in a vacuum. They must be grounded in *realistic requirements* and investigated *experimentally in a real-world setting* against the yardstick of ground truth. We satisfy these goals via a collaboration with the MUSE science museum (Trento, Italy), interested in analyzing the fruition of their exhibits by visitors. This is hardly a novel topic [31]; however, state-of-the-art works [5, 23, 30, 44] rely on Bluetooth and are therefore limited to coarse, room-level spatial resolution. Instead, our target area contains exhibits placed only few decimeters apart from each other, and curators focus on stops as short as tens of seconds. This demands a fine-grained spatio-temporal resolution, hard if not impossible to achieve with Bluetooth, but in principle achievable with UWB. To this end, we track users wearing UWB tags via a time-difference-of-arrival (TDoA) localization system [40] we deployed in the area.

1.1 Methodological challenges

Crucial to the practical use of stop-move detection techniques is a clear understanding of their performance, both to compare them and inform their configuration. However, for reliable results, their output must be *quantitatively* compared against ground truth. This goal poses several challenges. Before discussing them, we first introduce some basic notation, used throughout the paper, to capture precisely the concepts under consideration.

Notation and definitions. Table 1 offers a summary of the notation used throughout the paper. Given a stop event $s = \langle \mathbf{p}, [t_{start}, t_{end}], o \rangle$, we access its components via the functions $\mathsf{pos}(s)$, $\mathsf{int}(s)$, and $\mathsf{obj}(s)$; the stop *duration* is computed as $|\mathsf{int}(s)|$. Similarly, we use $\mathsf{label}(o)$ and $\mathsf{pos}(o)$ to access the components of a spatial object $o = \langle l, \mathbf{q} \rangle$.

When the stop event is computed from a segment $U = \langle (\mathbf{p}_i, t_i) \dots (\mathbf{p}_{i+k}, t_{i+k}) \rangle$ determined by the stop detection technique, then $\mathsf{pos}(s) = \mathcal{P}(U)$ and $\mathsf{int}(s) = \mathcal{I}(U)$, where $\mathcal{P}(U)$ and $\mathcal{I}(U)$ are the functions computing the centroid and interval from the units in U, as mentioned earlier. As for the association of spatial objects to stop events, it is computed as $\mathsf{obj}(s) = \mathcal{O}(s)$, where $\mathcal{O}(s)$ encodes a notion of proximity and yields the object $o_i \in O$ whose position $\mathsf{pos}(o_i)$ is closest to the position of the stop event, $\mathsf{pos}(s)$. The specific definition of $\mathcal{O}$ and the geometrical modeling of the objects in O (e.g., points, lines, polygons) are application-dependent; we later provide instantiations for our use case. Finally, a stop event can occur anywhere and possibly far from all spatial objects (e.g., a visitor moving away from museum exhibits to answer a phone call), a case we denote with $\mathsf{pos}(s) = \emptyset$.

Finally, hereafter we refer to stop events simply as *stops*, for readability.

How to establish a reliable and accurate ground truth? The answer is non-trivial. Ideally, the ground truth should be constituted by the *exact* position $\mathsf{pos}(s)$ and interval $\mathsf{int}(s)$ for each stop s performed by the entity being tracked. In practice, when these entities are mobile this is possible

Table 1. A summary of notation.

$T = \langle (\mathbf{p}_1, t_1) \dots (\mathbf{p}_n, t_n) \rangle$	generic trajectory as a sequence of units (*position, timestamp*)
$U = \langle (\mathbf{p}_i, t_i) \dots (\mathbf{p}_{i+k}, t_{i+k}) \rangle$	generic segment of k units in T, representing either a move or a stop
$\mathcal{P}(U)$, $\mathcal{I}(U)$	functions yielding the position and interval of a stop segment
$o = \langle l, \mathbf{q} \rangle, o \in O$	spatial object (POI) as a tuple $\langle$*label, position*$\rangle$, belonging to a predefined set of objects
$s = \langle \mathbf{p}, [t_{start}, t_{end}], o \rangle$	generic stop event s as a tuple $\langle$*position, [start timestamp, end timestamp], object (POI)*$\rangle$
pos(s), int(s), |int(s)|, obj(s)	functions to access the position, interval, duration, and object (POI) associated to the stop s
$O(\hat{s})$	function mapping an estimated stop $\hat{s}$ to an object o (POI), based on proximity
$s_{real} \in S_{real}, s \in S, \hat{s} \in \hat{S}$	real, true, estimated stops (Fig. 2) and their respective sets
ρ	minimum duration of relevant stops (for all techniques)
$\mathcal{D}(T, \Pi, \rho)$	stop detection (segmentation) operation, whose parameters Π (below) depend on the technique
ϵ, N	SeqScan parameters (max. distance of units from core point, min. number of units in a cluster)
δ	SPD parameter (max. distance of units from the first unit of the stop)
θ	KBV parameter (speed threshold)
$(S, \hat{S}, E, \mathcal{W})$	bipartite graph for computing the optimal match
$e_{ij} = (s_i, \hat{s}_j) \in E$	edge between true and estimated stops, denoting a potential matching
$\mathcal{W}(e_{ij})$	weight function combining temporal and spatial similarities into the overall stop similarity
$\mathcal{W}_t(e_{ij})$, $\mathcal{W}_s(e_{ij})$	temporal-only and spatial-only weight functions
$\alpha \in [0, 1]$	relative weight of spatial and temporal components, e.g., $\alpha = 1 \Rightarrow \mathcal{W}(e_{ij}) = \mathcal{W}_t(e_{ij})$
$\mathcal{W}_o(e_{ij})$	weight function enforcing matching POIs, i.e., obj(s_i) = obj($\hat{s}_j$)
M	optimal matching set maximizing the sum of $\mathcal{W}(e_{ij})$
$\hat{S}_{TP}, \hat{S}_{FP}, S_{FN}$	set of true positive (TP), false positive (FP), and false negative (FN) stops
F-score, S-score	correctness and spatio-temporal similarity of matched stops in M
$\Delta t, \Delta \mathbf{p}$	temporal and spatial error between matched stops
fake, split, short, missing, merged, mislabeled	nature of false detections

only with, e.g., a motion capture system [32] offering mm-level accuracy. However, these systems are very expensive, suffer from occlusions, and require subjects to wear special markers in different points of the body; consequently, they are not applicable to our context.

Therefore, we resort to the following methodology and setup. First, we perform experiments where the subjects being tracked stop near a museum exhibit and remain *immobile* for some time before moving to another exhibit along a path of their choosing. The stop positions are predefined and known, while the stop durations are recorded by the subject via a custom smartphone application as well as derived from the video footage from cameras we temporarily deploy in the target area (§4). This highly-controlled *"in-vitro"* setup trades realism for very accurate ground truth, and is key in verifying the *feasibility* of stop-move detection with UWB and accurately quantify its quality in the target while minimizing behavioral biases. However, this stylized subject behavior is hardly representative of actual visitors, and is therefore insufficient to confirm the *applicability* of our techniques in a real-world, practical setting. To this end, we perform a second set of *"in-vivo"* experiments where we track volunteer visitors *moving arbitrarily* inside the target area. This setup allows us to ascertain whether the results derived in the controlled, limited in-vitro setting still hold in this real-world, uncontrolled setting; a positive answer would provide solid grounds for the exploitation of our results in operational systems targeting large-scale, longitudinal campaigns. The price to pay for realism, however, is a slight reduction in ground-truth information because, unlike in vitro, the precise position of stops is not known. However, the exhibit (POI) visited and the stop interval can still be reliably derived from the video footage.

How and what to compare against ground truth? Fig. 2 illustrates the methodological challenge. Subjects are tracked by the UWB localization system deployed in the target area, yielding a trajectory for each of them. In the simplified view of Fig. 2, we show a portion of a UWB trajectory in which the subject stops near a spatial object o (POI). In principle, the position $\mathbf{p}$ and interval $[t_{start}, t_{end}]$ of this *real* stop s_{real} and its association to the object o are all accurately and *directly* captured by ground-truth information, and *not* computed via segmentation. However, as discussed above, this is the case only in vitro; in vivo, the stop position is not available. Still, the fact that the *exact* stop

interval $\text{int}(s_{real})$ and associated object $\text{obj}(s_{real})$ are known from ground truth allows us to define a *true* stop as $s = \langle \mathcal{P}(U), \text{int}(s_{real}), \text{obj}(s_{real}) \rangle$ where $\mathcal{P}(U)$ is computed as the centroid on the subsequence $U = \langle (\mathbf{p}_{start}, t_{start}) \dots (\mathbf{p}_{end}, t_{end}) \rangle \subseteq T$, i.e., the one delimited by the ground-truth interval. The stop s is "true" because $\text{int}(s) = \text{int}(s_{real})$ and $\text{obj}(s) = \text{obj}(s_{real})$ hold by construction, yet it does not coincide with the "real" stop as their positions are slightly different, i.e., $\text{pos}(s) \neq \text{pos}(s_{real})$. The reason is that the UWB positioning error ①, albeit significantly smaller w.r.t. GNSS, cannot be neglected as its magnitude is comparable to the fine-grained stop resolution we target.

True stops are the result of the best segmentation that can be extracted from UWB trajectories alone, given ground-truth intervals. In practice, however, segmentation induces an error ② between the true stop s and the *estimated* stop, $\hat{s}$; a real technique may incorrectly estimate a duration $\text{int}(\hat{s}) \neq \text{int}(s)$, consequently misplacing its position $\text{pos}(\hat{s}) \neq \text{pos}(s)$.

Finally, the combination of the spatial error induced by UWB positioning ①, $\|\text{pos}(s_{real}) - \text{pos}(s)\|$, with the one induced by segmentation ②, $\|\text{pos}(s) - \text{pos}(\hat{s})\|$, yields the overall spatial error ③, $\|\text{pos}(s_{real}) - \text{pos}(\hat{s})\|$. However, as already mentioned, this cannot be quantified in general, as the exact, ground-truth position $\text{pos}(s_{real})$ is available only in-vitro. For this reason in our analyses (§7–§9) we always compare estimated stops $\hat{s}$ with true stops s, and analyze separately (§6) the UWB positioning error present in true stops vs. real stops s_{real}. This allows us to retain a common analysis framework across the in-vitro and in-vivo campaigns and, it is also better aligned with real setups where, like in vivo, the real position is not available and the concern is to choose/configure the techniques to minimize the additional error they introduce w.r.t. the pure UWB trajectories. On the other hand, the overall spatial error ③ affects also semantic annotations by potentially yielding

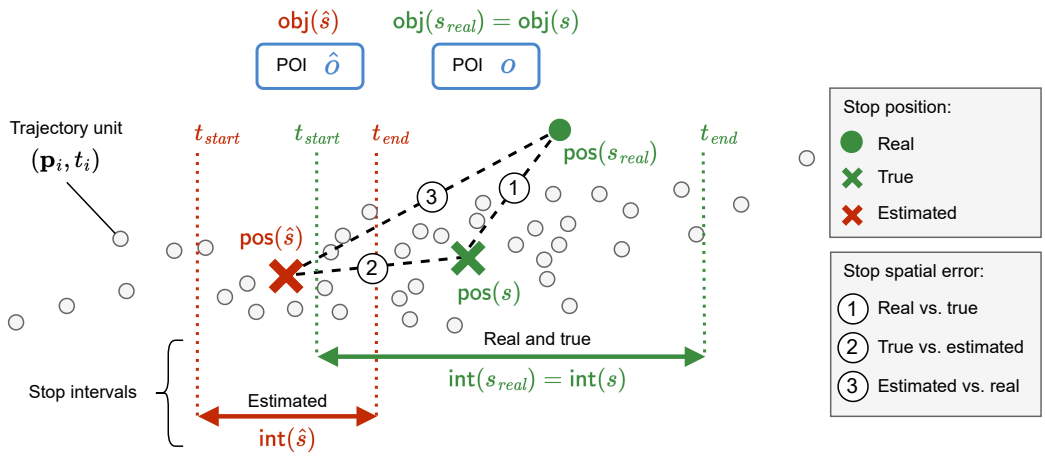


Fig. 2. Estimated (red) vs. ground-truth (green) stops. The tracked subject stops near a spatial object (POI) o. The accurate position, time interval, and associated object of this *real* stop are captured as $s_{real} = \langle \mathbf{p}, [t_{start}, t_{end}], o \rangle$. However, the precise values for these components can be determined only in vitro. In vivo, a *true* stop s approximates the real one by computing the position from UWB traces, inducing a spatial error ①; still, the time interval $[t_{start}, t_{end}]$, determining the stop segment U used to compute the centroid $\mathcal{P}(U)$, and the associated object o are the same as in the real one. A true stop represents the best possible output for segmentation techniques. In practice, however, they output an *estimated* stop $\hat{s}$ inducing a spatio-temporal error ② w.r.t. the true stop s and, along with the error from UWB positioning ①, an overall spatial error ③ w.r.t. the real stop s_{real}, potentially yielding an incorrect association with the object $\hat{o}$.

an association $\text{obj}(\hat{s}) \neq \text{obj}(s_{real})$ to an incorrect object, e.g., $\hat{o}$ in the figure. We do consider this aspect, key to domain experts, in our analyses.

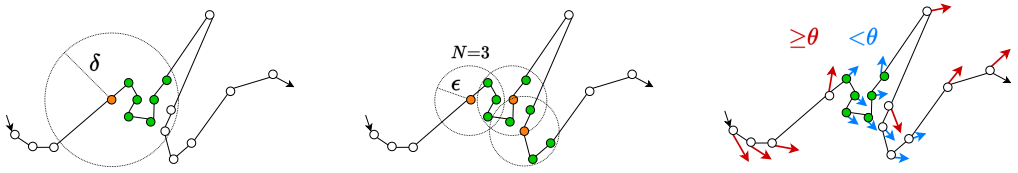
1.2 Contributions

These concerns received scarce attention in literature, where they are addressed only qualitatively or by using *unit-centric* metrics [2, 27], whose expressiveness is insufficient to capture the complexity above. Indeed, these metrics quantify the mismatch induced by stop-move detection at the level of individual units, e.g., computing the fraction correctly labeled as belonging to a stop among all units in a trajectory. Nevertheless, this low level of abstraction hampers analyses capturing the multi-unit nature of stops *as a whole*, e.g., to identify cases where a true stop is estimated as several ones or, dually, when several true ones are estimated as a single one. Similarly, unit-centric techniques are oblivious to the association of POI to stops, since these are not directly captured. In contrast, we offer a **novel *family* of metrics** (§3) that *i)* adopts a *stop-centric* perspective directly capturing the mismatch between estimated and true *ii)* quantifies separately spatial and temporal errors, whose relative importance can be tailored to the requirements at hand, and *iii)* connects *directly* segmentation quality to the semantic application context relevant to domain experts (e.g., POIs), by reuniting the two into a single methodological framework. This higher-level stop-centric perspective is more expressive and intuitive, and characterizes quality directly via the first-class notion of stop utilized by domain experts, rather than low-level units composing them.

Crucial to our study (§1.1) is the collection of **datasets in controlled conditions (in vitro) and with real visitors (in vivo)**, consisting of 70,090 positions (209 stops) and 219,937 positions (550 stops), respectively. These datasets (§5) are a contribution per se, as indoor mobility traces are scarce in the literature, especially for UWB. We release them publicly [17], enabling others to reproduce and build upon our results.

We input both datasets to representative segmentation techniques (§2) and **quantitatively analyze** their output along several dimensions. This serves simultaneously as a **validation of our metrics** as well as **concrete evidence of the stop detection accuracy and resolution enabled by UWB.** We begin with the in-vitro dataset and an accurate characterization (§6) of the UWB positioning error (Fig. 2, ①) before distilling our findings (§7). We confirm the higher expressiveness and accuracy of our stop-centric approach vs. unit-centric ones, then compare techniques after selecting their best configuration. We quantify the impact of raw UWB trajectories as input vs. those "smoothed" via Kalman filters, exploit our metrics to analyze the *nature* of false detections, and quantify spatio-temporal errors and the ability to correctly associate estimated stops to ground-truth POIs. Despite the challenges of our environment, the quality of the best technique is very high (F-score=0.947). Stop durations are estimated with an average error of only 3.7 s. Regarding positions, UWB localization is the main source of spatial error (46 cm), increased only to a small extent (3.2 cm) by segmentation. Finally, their combination affects only marginally the association of estimated stops to ground-truth POIs. These results are confirmed in vivo (§8); despite the unpredictable visitor movement and the more complex exhibit setup, we observe only a small degradation in quality (F-score=0.911). In these uncontrolled experiments, we also showcase the expressive power and versatility of our metrics by applying them to quantify the quality of the higher-level notion of *visit* obtained by aggregating consecutive stops around the same exhibit (§9).

Our survey of related efforts (§10) shows that our work is the first studying stop-move detection on UWB-based trajectories and, importantly, offering an evaluation against systematically-acquired ground truth in a real-world environment. Our earlier work [16] first provided evidence of the feasibility of our approach in a controlled, in-vitro environment. This paper builds upon these early results but also goes beyond them by:



(a) SPD. The stop starts with the orange unit and is spatially delimited by a circle of radius δ. The stop segment ends at the first unit (possibly an outlier) with distance $>\delta$ from the orange unit.

(b) SeqScan. A core point (orange) has at least N units within an ϵ radius. A cluster includes all the neighbors of core points and, transitively, those of connected core points.

(c) KBV. Stops are defined as low-speed segments, whose units have a velocity $<\theta$ (blue arrows), based on the Kalman filter. A stop segment ends at the first unit with velocity $\geq\theta$ (red arrows).

Fig. 3. Stop detection techniques and their parameters. The minimum stop duration ρ is the same and not shown. Units belonging to a stop are denoted by green/orange dots; move units are instead white.

- *Eliciting and defining practical notions of ground-truth* (§1.1) to reconcile the different constraints and opportunities arising when moving from an in-vitro setup to an in-vivo one.
- *Redesigning the quality metric,* originally based solely on temporal information. Here, we improve its definition and generalize it into a new *family* of metrics that characterize both temporal and spatial errors and connects them to the semantic application context (POIs) in a unified conceptual framework. As a consequence, although our in-vitro findings (§7) are based on the original dataset in [16], the analysis has been entirely redone and expanded to account for the new definition of metrics and the new dimensions they capture.
- *Providing evidence of real-world applicability.* Our in-vivo findings (§8) are not simply another evaluation. On the contrary, they fill the gap between *i)* the mere *feasibility* of our techniques and more generally of a UWB-based approach, assessed in [16] in the unrealistic yet fully controlled in-vitro setup where subjects stop by remaining *immobile in a designated point*, and *ii)* the actual *applicability* in a real context where visitors exhibit real stop-move patterns as they *move of their own volition.* These real patterns also prompted us to capture the frequent case of several stops around the same POI in the new concept of *visits*, whose quality we show can be assessed using the same family of metrics above (§9). Finally, our in-vivo campaign resulted in a dataset 3× larger than the in-vitro one (§5.2).

Overall, our technical, methodological, and experimental contributions offer positive, quantitative answers to our research questions, pushing the applicability of mobility analysis to an unprecedented fine-grained spatio-temporal resolution, as mentioned in our concluding remarks (§11).

2 STOP DETECTION: PROBLEM FORMULATION AND TECHNIQUES

The stop detection problem consists of extracting from a trajectory T the sequence of *estimated stop events* $\langle \hat{s}_1 \dots \hat{s}_m \rangle$, where $\hat{s}_i = \langle \mathbf{p}_i, [t_{start}, t_{end}]_i, o_i \rangle$. This segmentation process is performed by a detection technique that can be abstractly denoted as $\mathcal{D}(T, \Pi, \rho)$, where T is the input trajectory and Π is the set of configuration parameters specific to each technique and described in the following. Instead, ρ indicates the minimum allowed duration of estimated stops; a stop $\hat{s}_i$ for which $|\mathsf{int}(\hat{s}_i)| < \rho$ is ignored. The value of ρ is application-dependent and set based on domain knowledge. For instance, in our case the value of ρ is tied to the minimum stop denoting visitor attention; after consultation with the museum curators, we set $\rho = 10$ s. We use this value for all techniques under consideration, to enable their comparison.

The specific techniques we chose, embodying the detection function $\mathcal{D}$, are representative of state-of-the-art approaches (§10) with different complexity and tradeoffs. They are illustrated in Fig. 3 and can be summarized as follows:

- **Using spatial distance between units: SPD.** The Stay Point Detection (SPD) technique [48] identifies stops in GPS trajectories. It relies on a distance threshold δ representing the maximum spatial extent allowed between the first unit belonging to a stop and the other units belonging to the same stop; this threshold is the only configuration parameter, $\Pi = \{\delta\}$. Specifically, a stop segment $S = \langle(\mathbf{p}_{start}, t_{start}) \dots (\mathbf{p}_{end}, t_{end})\rangle$ contains all time-consecutive units $(\mathbf{p}_i, t_i) \in T$ whose spatial distance from the first segment unit is within the threshold, $\|\mathbf{p}_{start} - \mathbf{p}_j\| < \delta$. This technique is commonly used due to its simplicity. However, it is not well-suited when stops have different spatial size or are affected by outlier units, e.g., due to positioning noise, making it difficult to define a one-size-fits-all value for δ.
- **Using unit density: SeqScan.** These limitations can in principle be mitigated via density-based clustering. For instance, in DBSCAN [12] a *core point* in some abstract space has at least N neighbors within distance ϵ; a *cluster* contains these points and, transitively, those of neighboring core points. A direct application of this approach to stop detection maps the abstract space on the physical space associated to trajectories and clusters the *position* of their units. This yields stops as the final output clusters, whose properties above allow for outlier units temporarily escaping beyond ϵ. On the other hand, this naive approach alone cannot guarantee the temporal separation of the clusters representing stops except when spatially far from each other, as time is not considered; repeated stops (e.g., at the same exhibit but at different times) become indistinguishable. SeqScan [8] overcomes this limitation by using the same configuration parameters $\Pi = \{\epsilon, N\}$ of DBSCAN, but defining a stop as a cluster of units $\langle(\mathbf{p}_i, t_i) \dots (\mathbf{p}_j, t_j)\rangle$ where $\langle \mathbf{p}_i \dots \mathbf{p}_j \rangle$ is a DBSCAN cluster. Therefore, SeqScan clusters *i)* have arbitrary spatial shape and are robust against noise, unlike SPD, and *ii)* have *disjoint* time intervals $[t_i, t_j]$, unlike DBSCAN.
- **Using user velocity from Kalman filters.** Another way to look at stops is when the velocity of the tracked subject is (nearly) zero. Velocity can be derived from UWB trajectories; however, their noisy raw positions induce unacceptable velocity jitter if used directly. Interestingly, trajectories are commonly improved [45] via Kalman filters (§4) whose operation already entails hidden state variables representing the velocity associated with units $(\mathbf{p}_i, t_i)$. Segmentation then simply consists of identifying as stop segments those containing consecutive units whose velocity is below a threshold θ; the latter constitutes the only configuration parameter, $\Pi = \{\theta\}$. This Kalman-based velocity technique, KBV hereafter, is to our knowledge novel in stop-move detection and relevant here as a computationally cheap approach reusing the filtering commonly applied to trajectories.

3 QUALITY METRICS FOR FINE-GRAINED STOP DETECTION

Ideally, the set $\hat{S}$ of estimated stops output by stop detection techniques coincides with the set S of true stops; in practice, this is rarely the case. We illustrate the family of metrics we define to ascertain and compare the quality of stop-detection techniques. As mentioned (§1.1), hereafter we consider true stops rather than real stops, as using the latter would prevent us from separating and quantifying the error ② induced by segmentation from the error ① intrinsic in UWB positioning, analyzed later (§6).

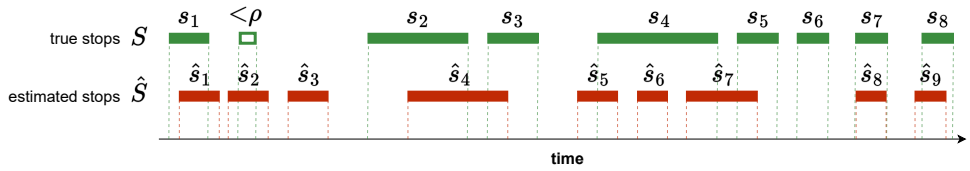


Fig. 4. An (artificial) example of segmentation. S is the set of of true stops, $\hat{S}$ the one for estimated stops. For graphical illustration, only the time intervals $\text{int}(s_i)$ and $\text{int}(\hat{s}_i)$ associated to the stops are shown.

3.1 Rationale and Novelty

The problem of defining metrics to compare stop-move detection techniques against ground truth has received surprisingly little attention in the literature, where it is often addressed only qualitatively. An exception are recent works [2, 27] proposing metrics to quantify quality at the unit level, based on the binary stop-move classification of trajectory units induced by segmentation. A unit $u_i = (\mathbf{p}_i, t_i)$ labeled as "stop", i.e., belonging to an estimated stop $\hat{s}$, is a true positive (TP) if it belongs also to a true stop s; otherwise, u_i is a false positive (FP). Similarly, a unit u_i labeled as "move", i.e., not belonging to an estimated stop $\hat{s}$, is a true negative (TN) if it also does not belong to a true stop s; otherwise, u_i is a false negative (FN).

Unfortunately, these *unit-centric* approaches are oblivious to the segmentation *structure*. By focusing on which *units* fall into the time interval of true stops, they do not capture properties of the estimated *stops* these units belong to. Even simple measures, e.g., the number of correctly identified stops, are lost in the flat, unit-centric view.

In contrast, the novel metrics we propose are based directly on the notion of stop, and aim at *matching* estimated stops $\hat{s}$ and true stops s. The correctness of the matching can then be defined *directly and intuitively based on stops rather than units*, accounting also for a notion of *spatio-temporal similarity* among stops that is intrinsically precluded to unit-centric approaches and is also useful to *quantify* how well an estimated stop matches a true one. We provide domain experts with the flexibility to define different instantiations of the metrics striking a custom balance between the spatial and temporal aspects. Moreover, our rich and expressive definition yields precious information about the *nature* of incorrect detections and is also applicable to the association between stops and POIs, reuniting this higher-level, semantic layer into the same methodological framework of basic stops.

We argue, and confirm quantitatively (§7.1), that the change from units to stops in the "lens" used to analyze segmentations increases expressiveness and practical relevance.

3.2 A Family of Stop-centric Metrics: Core Concepts

Our metrics revolve around the notion of matching estimated stops against true ones. We describe how this key notion is defined in our metrics.

Matching estimated and true stops. Consider the artificial example in Fig. 4, where S and $\hat{S}$ are the sets of true stops and estimated stops, respectively. Our goal is to establish a one-to-one *matching* relationship between true stops $s \in S$ and estimated stops $\hat{s} \in \hat{S}$, based on the intuition that if they represent the same stop, their time intervals must overlap and their centroids be close. Compared to unit-centric approaches, ours *directly* captures the quality of segmentation by quantifying how many true stops are

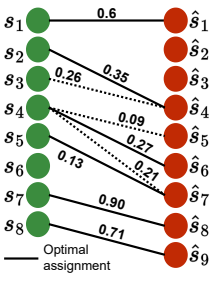


Fig. 5. Finding the matching set in Fig. 4 based on temporal similarity.

correctly reflected in estimated ones. Still, due to errors (Fig. 2), an estimated stop may overlap in time and/or be close in space with multiple true stops and vice versa. How to perform matching?

We cast the problem as an unbalanced assignment optimization over the bipartite graph $(S, \hat{S}, E, \mathcal{W})$ where E is the set of edges $e_{ij} = (s_i, \hat{s}_j)$ denoting a potential match between a true stop s_i and an estimated one $\hat{s}_j$, and $\mathcal{W}$ is a *stop similarity function* assigning a weight to the edges in E.

$\mathcal{W}$ is key to capture the spatio-temporal nature of segmentation quality and tailor the metric to application requirements. Applications focusing on spatial proximity may privilege spatial accuracy w.r.t. the temporal one; those focusing on dwelling time may privilege the opposite; finally, others may require the evaluation of both. We reconcile these concerns with the following definition:

$$\mathcal{W}(e_{ij}) = \begin{cases} 0 & \text{if } \mathsf{int}(s_i) \cap \mathsf{int}(\hat{s}_j) = \varnothing \\ \alpha \mathcal{W}_t(e_{ij}) + (1-\alpha)\mathcal{W}_s(e_{ij}) & \text{otherwise} \end{cases}$$

If the true and estimated stops are temporally disjoint the overall similarity is set to zero to avoid considering nonsensical cases, e.g., nearby stops occurring at entirely different times. Otherwise, $\mathcal{W}$ is a linear combination of the temporal similarity $\mathcal{W}_t$ and spatial similarity $\mathcal{W}_s$, whose relative importance is controlled by $\alpha \in [0, 1]$.

The temporal similarity

$$\mathcal{W}_t(e_{ij}) = \frac{|\mathsf{int}(s_i) \cap \mathsf{int}(\hat{s}_j)|}{|\mathsf{int}(s_i) \cup \mathsf{int}(\hat{s}_j)|}$$

is inspired by the Jaccard index over the time intervals for true and estimated stops, and quantifies the overlap between the two.

Instead, the spatial similarity

$$\mathcal{W}_s(e_{ij}) = \frac{1}{1 + h \, \|\mathsf{pos}(s_i) - \mathsf{pos}(\hat{s}_j)\|}$$

is inversely related to the Euclidean distance between their centroids, multiplied by a scaling factor h that enables tuning the metric w.r.t. the spatial resolution required by the application and the magnitude of positioning errors.

Formally, the problem is to determine the optimal assignment yielding the *matching set* $M \subseteq E$ such that:

$$M = \underset{E_M \subseteq E}{\operatorname{argmax}} \sum_{E_M} \mathcal{W}(e_{ij})$$

subject to:

$$e_{ij} \in M \Rightarrow \mathcal{W}(e_{ij}) \neq 0 \wedge \nexists k \mid e_{ik} \in M \wedge \nexists l \mid e_{lj} \in M, \forall i, j, k, l \mid k \neq j, l \neq i$$

i.e., maximizing the aggregated stop similarity of all non-zero links and ensuring each stop is matched at most once.

Fig. 5 illustrates how these concepts are applied, based on the example in Fig. 4 and exploiting only temporal similarity (α=1). The optimal assignment yields

$$M = \{(s_1, \hat{s}_1), (s_2, \hat{s}_4), (s_4, \hat{s}_6), (s_5, \hat{s}_7), (s_7, \hat{s}_8), (s_8, \hat{s}_9)\}.$$

Quantifying (in)correct detections (F-score). The matching set M captures which estimated stops best approximate the true stops while maximizing their similarity. An ideal segmentation would yield a one-to-one correspondence between each estimated stop $\hat{s} \in \hat{S}$ and each true stop $s \in S$, yielding a bijection between S and $\hat{S}$. In practice, some estimated stops may not have a correspondent true stop, and some true stops may go undetected. To ascertain the quality, it is essential to *quantify* this mix of correct and incorrect detections, which we achieve with the following definitions.

Given the sets of estimated stops $\hat{s} \in \hat{S}$ and true stops $s \in S$, and a matching set M, the set $\hat{S}_{TP} \subseteq \hat{S}$ contains the estimated stops belonging to M, $\hat{S}_{TP} = \{\hat{s}_j \in \hat{S} \mid \exists s_i : (s_i, \hat{s}_j) \in M\}$, i.e., the estimated

stops that have a corresponding true stop, indicating successful detection; we refer to the stops in $\hat{S}_{TP}$ as *true positives (TP)*. Along the same lines, we define the set $\hat{S}_{FP} \subseteq \hat{S}$ of *false positives (FP)* as $\hat{S}_{FP} = \{\hat{s}_i \in \hat{S} \mid \nexists s_j : (s_j, \hat{s}_i) \in M\}$, i.e., the estimated stops for which a corresponding true stop does not exists. Finally, the set $S_{FN} \subseteq S$ of *false negatives (FN)*, contains the true stops that do not have a corresponding estimated stop, $S_{FN} = \{s_i \in S \mid \nexists \hat{s}_j : (s_i, \hat{s}_j) \in M\}$. These three sets are disjoint; in our example, $\hat{S}_{TP} = \{\hat{s}_1, \hat{s}_4, \hat{s}_6, \hat{s}_7, \hat{s}_8, \hat{s}_9\}$, $\hat{S}_{FP} = \{\hat{s}_2, \hat{s}_3, \hat{s}_5\}$, $S_{FN} = \{s_3, s_6\}$. Moreover, by construction, the number of true positives is equal to the number of true stops in the matching set, which in turn is equal to the number of elements in it: $|\hat{S}_{TP}| = |M| = |\{s_i \mid \exists \hat{s}_j : (s_i, \hat{s}_j) \in M\}|$.

Based on these definitions, we evaluate the aggregate quality of estimated stops via the following well-known metrics:

- *Precision* is the fraction of estimated stops that match true stops: $P = \frac{|M|}{|\hat{S}|} = \frac{|\hat{S}_{TP}|}{|\hat{S}|} = \frac{|\hat{S}_{TP}|}{|\hat{S}_{TP}|+|\hat{S}_{FP}|}$.
- *Recall* is the fraction of true stops matched by estimated stops: $R = \frac{|M|}{|S|} = \frac{|\hat{S}_{TP}|}{|S|} = \frac{|\hat{S}_{TP}|}{|\hat{S}_{TP}|+|S_{FN}|}$.
- *F-score* is the harmonic mean $F = 2\frac{P \times R}{P+R}$ of P and R, and serves as a concise indicator capturing both.

These metrics capture concisely the quality of stop detection performance by combining the number of correct and incorrect detections. Their values range in $[0, 1]$; an ideal segmentation with a perfect one-to-one correspondence between true and estimated stops would yield $P = R = F = 1$. In our example, $P = 0.67$, $R = 0.75$, and $F = 0.71$.

Determining the nature of false detections. A distinguishing aspect of our approach is that the *nature* of incorrect detections in $\hat{S}_{FP}$ and S_{FN} can be inferred from the relationship between the time intervals associated to an estimated stop $\hat{s}_i$ and a true one s_j. This expressive feature enables a deeper understanding of the limitations of stop detection techniques and a further point of comparison among them.

A false positive $\hat{s}_i \in \hat{S}_{FP}$ can be either:

- A *fake* stop if it does not overlap in time with any true stop s_j, $\nexists s_j \in S \mid \mathsf{int}(\hat{s}_i) \cap \mathsf{int}(s_j) \neq \emptyset$ (e.g., $\hat{s}_3$ in Fig. 4); a stop has been detected where no true one exists.
- A *split* stop if it overlaps in time with a true stop, $\exists s_j \in S \mid \mathsf{int}(\hat{s}_i) \cap \mathsf{int}(s_j) \neq \emptyset \wedge (s_i, \hat{s}_j) \notin M$. This error occurs when the correspondence between true and estimated stops is one-to-many, resulting in the detection of one or more extra estimated stops; indeed, as $\hat{s}_i$ is a FP, it does not match any s_j but some other $\hat{s}_k \neq \hat{s}_i, k \neq i$ must, otherwise the assignment would not be optimal. In our example, s_4 overlaps with both $\hat{s}_5$ and $\hat{s}_6$ but is matched only by $\hat{s}_6$.
- A special case is a *short* stop, when the duration of s_j is $<\rho$ (i.e., irrelevant, §2) but the one of $\hat{s}_i$ is not (e.g., $\hat{s}_2$). In many use cases this is a "benign" FP as segmentation correctly identifies the stop, incorrectly estimating only its duration.

Similarly, a false negative $s_j \in S$ can be either:

- A *missing* stop if it does not overlap in time with any estimated stop, $\nexists \hat{s}_i \in \hat{S} \mid \mathsf{int}(\hat{s}_i) \cap \mathsf{int}(s_j) \neq \emptyset$, as in the case of s_6.
- A *merged* stop if it overlaps in time with an estimated stop, $\exists \hat{s}_i \in \hat{S} \mid \mathsf{int}(\hat{s}_i) \cap \mathsf{int}(s_j) \neq \emptyset \wedge (s_i, \hat{s}_j) \notin M$. This happens when two or more true stops are estimated as a single one. Indeed, s_j is not matched by $\hat{s}_i$ although they overlap; thus, $\hat{s}_i$ must match another $s_k \neq s_j, k \neq j$ otherwise, again, the assignment would not be optimal. In our example, s_3 is lumped into the same estimate $\hat{s}_4$ matched to s_2.
- As with false positives, a false negative *short* stop could in principle capture a true s_j incorrectly estimated by $\hat{s}_i$ with duration $<\rho$. However, these are automatically filtered by segmentation techniques (§2).

Quantifying the similarity of matched stops (S-score). In our running example, M contains both $(s_1, \hat{s}_1)$ and $(s_2, \hat{s}_4)$, whose temporal overlapping between estimated and true stops is remarkably different (Fig. 4). Still, another segmentation yielding the same M but with $(s_1, \hat{s}_1)$ and $(s_2, \hat{s}_4)$ perfectly aligned temporally would yield the same F-score. This indicator is therefore an expressive measure of the *correctness* of the segmentation, but does not capture how *similar* in space and time the individual matched stops are. To this end, we complement the F-score with the

$$\textit{S-score} = \frac{1}{|M|} \sum_{e_{ij} \in M} \mathcal{W}(e_{ij})$$

The S-score plays a key role with different α values, highlighting differences in the spatio-temporal features of the segmentation, e.g., focusing on how accurately stop durations ($\alpha = 1$) or positions ($\alpha = 0$) are estimated. This flexibility is useful when the target application is concerned with only one of the two dimensions, as mentioned earlier, or when the acquisition process for the ground truth of one of them is not as dependable as the other; intermediate configurations of the metric are also possible by properly setting α. In this respect, we note that if both temporal and spatial data are accurate, the difference in stop similarity with different α should be small; when this is not the case, the metric allows one to unveil critical issues in either the system under study or the acquired ground truth.

3.3 Incorporating Ground-truth Semantic Annotations

The metrics above ascertain segmentation quality w.r.t. true stops with input data from UWB trajectories. However, when ground truth is available for stops near POIs, as in both our setups, the stop-centric metric can be extended to answer a question of practical relevance: *Can we correctly* associate *the POI visited by the user based on the estimated stop position?* This is not obvious, as the combination of segmentation and positioning errors may be large enough to jeopardize the association between estimated stop and POI (Fig. 2) and therefore the usefulness of segmentation.

The association is application-dependent and encoded in the function O (§1.1). However, in many scenarios it consists simply of determining the POI *closest* to the estimated stop. The notion of proximity encoded in the mapping function O is therefore the Euclidean distance between an arbitrary stop $\hat{s}$ and the spatial objects $o \in O$, as

$$O(\hat{s}) = \begin{cases} \arg\min_{o \in O} \|\text{pos}(\hat{s}) - \text{pos}(o)\| & \text{if } \|\text{pos}(\hat{s}) - \text{pos}(o)\| \leq R \\ \varnothing & \text{otherwise} \end{cases} \tag{1}$$

where R is an application-dependent threshold defining the distance beyond which a stop can no longer be considered "near" the object. The above can be easily generalized via a proper redefinition, e.g., to the case where the object is a polygon and proximity is determined by the Euclidean distance between $\text{pos}(\hat{s})$ and the polygon sides. Finally, for true and real stops, the object $\text{obj}(s) = \text{obj}(s_{real})$ is defined by ground truth and not by O (§1.1).

True vs. estimated annotations. Even this simple definition of proximity is challenging with fine-grained stop-move detection; our scenario is no exception, due to *i)* POIs close to each other, and *ii)* non-negligible UWB positioning error. Nevertheless, we can easily adapt our metrics to the problem of ascertaining quality w.r.t. the ground-truth semantic annotations in true stops. To this end, we redefine the weights determining matching as

$$\mathcal{W}_o(e_{ij}) = \omega \mathcal{W}(e_{ij}), \quad \omega = \begin{cases} 1 & \text{if } \text{obj}(s) = \text{obj}(\hat{s}) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

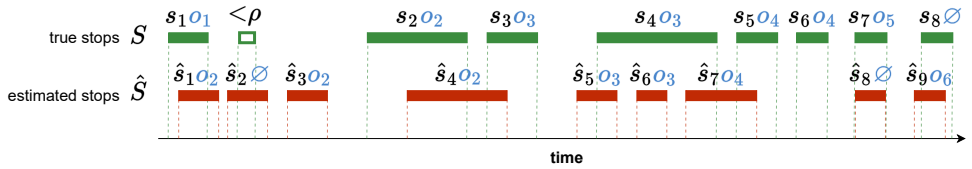


Fig. 6. Revisiting Fig. 4 with semantic annotation of stops (POIs).

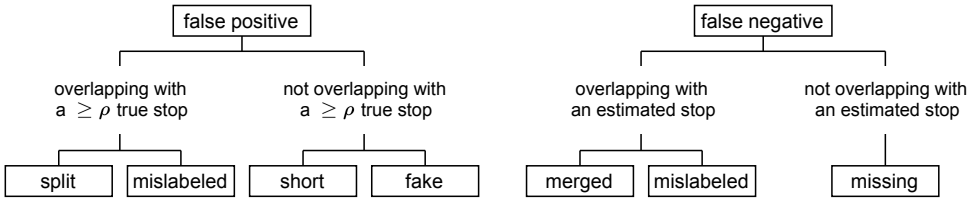


Fig. 7. A taxonomy of false detections.

This restricts purely spatio-temporal weights by allowing in the matching set M only the edges whose true and estimated stops are annotated with the same spatial object.

Fig. 6 revisits the example in Fig. 4 by considering also the semantic annotations. The matching set in this case becomes $M = \{(s_2, \hat{s}_4), (s_4, \hat{s}_6), (s_5, \hat{s}_7)\}$ where, compared to the one in §3.2, the three edges $(s_1, \hat{s}_1)$, $(s_7, \hat{s}_8)$ and $(s_8, \hat{s}_9)$ have been removed. In the first case, the best spatio-temporal match places the true and estimated stops near different spatial objects; in the other two, either the true or estimated stop is not near to *any* spatial object.

Revisiting the nature of false detections. Based on the matching set, we can determine the sets $\hat{S}_{TP}=\{\hat{s}_4, \hat{s}_6, \hat{s}_7\}$, $\hat{S}_{FP}=\{\hat{s}_1, \hat{s}_2, \hat{s}_3, \hat{s}_5, \hat{s}_8, \hat{s}_9\}$, $S_{FN}=\{s_1, s_3, s_6, s_7, s_8\}$ and the nature of false detections. However, annotations induce new cases w.r.t. those in §3.2. A FP estimated stop overlapping with a true one is a *split* stop only if both stops are annotated with the same object; otherwise, the FP is a *mislabeled* stop ($\hat{s}_1, \hat{s}_8, \hat{s}_9$ in Fig. 6). Dually, a FN true stop is a *merged* stop only if it overlaps with an estimated one at the same object; otherwise, is a *mislabeled* stop (s_1, s_3, s_6, s_7, s_8). The taxonomy of false detections (Fig. 7) witnesses the expressiveness of our metric.

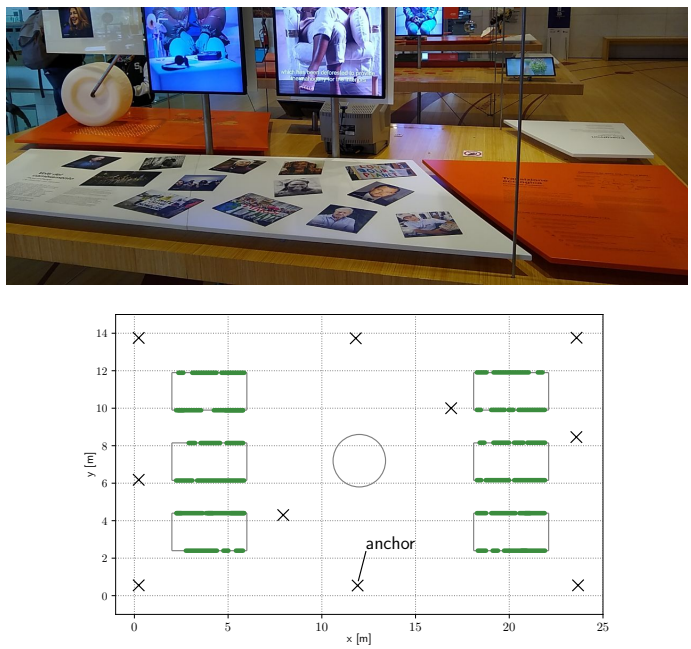
4 TRACKING VISITORS IN A SCIENCE MUSEUM

Before applying our quality metrics to the UWB datasets from our real-world museum use case we offer a description of the hw/sw infrastructure that enabled their collection.

UWB localization system. Each user wears a necklace with a UWB tag on the chest, a common deployment option. The position of the tag is the one actually recorded by UWB localization, for which we use Talla [40], a state-of-the-art system based on time-difference-of-arrival (TDoA).

Tags and localization anchors are battery- and mains-powered, respectively; both are Qorvo MDEK1001 devices hosting the popular DW1000 UWB radio [10]. Anchors are deployed in a 25×15 m^2 area (Fig. 8). Each anchor is connected via USB to a Raspberry Pi, relaying TDoA data to the localization server. The RPi is hidden in the false ceiling and connected to power and Ethernet via the existing PoE infrastructure. UWB devices are instead attached externally to the ceiling, to avoid its signal obstruction, via existing metallic supports hosting other equipment. This solution reuses the pre-existing cabling infrastructure and makes anchors visually non-invasive. The anchor position is chosen based on commonplace criteria for localization systems, to maximize the accuracy in tracking the tag position. The anchors placed on the perimeter fully include the target area and

(a) *In vitro.* The dots on the map mark the coordinates of the uni-dimensional point in front of each exhibit, representing the POI.



(b) *In vivo.* In the *same* area, the different layout of exhibits delimited by large colored tiles (top), implies a linear representation of POIs (bottom).

Fig. 8. Museum target area (top) and map (bottom) during the two experimental campaigns. The crosses in the maps mark the position of the UWB anchors placed on the ceiling.

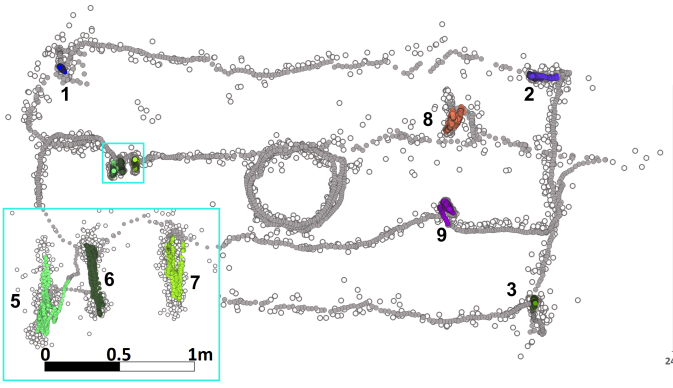


Fig. 9. Raw (white) vs. filtered (gray) trajectory. The colored points in the latter fall in ground-truth stop time intervals.

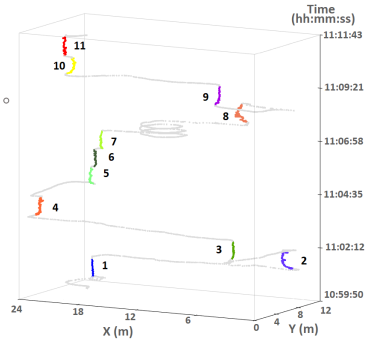


Fig. 10. Spatio-temporal view of the filtered trajectory in Fig. 9.

guarantee good geometrical properties, i.e., low geometric dilution of precision (GDOP). These would in principle be sufficient for localization; nevertheless, we placed two additional anchors near the center to mitigate the impact of radio signal occlusions by providing more opportunities for a clear line of sight between the tag and a sufficient number of anchors.

We configured Talla with 4 Hz time synchronization and 12 Hz position update rate, and the UWB radio with channel 5, 64 MHz PRF and 128 μs preamble.

Improving position estimates via Kalman filters. UWB position errors yield noisy trajectories commonly "smoothed" via Kalman filters [45]. Noise occurs for both moving and stationary tags; filters are usually optimized for either case. Still, stop-move detection requires efficient noise reduction in both cases *and* a fast switch between them to accurately determine a stop start/end times. Therefore, we combine two Unscented Kalman filters (UKF) representing the tag mode (stopped or moving) in the framework of Interacting Multiple Models (IMM) [3]. The output position is a linear combination of both filter estimates, weighted by the probability of each filter to match the current mode, i.e., tag behavior. The state of each filter includes 2D coordinates and the related velocity and acceleration. The IMM configuration M_{IMM}, along with the process noise Q and measurement noise R covariance matrices were determined experimentally:

$$Q = \text{diag}(Q_i, Q_i) \qquad Q_i = \begin{bmatrix} 5e{-}5 & 1.25e{-}4 & 1.67e{-}4 \\ 1.25e{-}4 & 3.33e{-}4 & 5e{-}4 \\ 1.67e{-}4 & 5e{-}4 & 1e{-}3 \end{bmatrix} \qquad R = \begin{bmatrix} 0.046 & 0 \\ 0 & 0.057 \end{bmatrix} \qquad M_{IMM} = \begin{bmatrix} 0.97 & 0.03 \\ 0.03 & 0.97 \end{bmatrix}$$

The IMM mode transition probabilities in M_{IMM} were set to favor the current mode of the system, with initial probabilities set to 0.5. The filter for the mobile mode uses the configuration above, while the one for stationary mode sets $Q = 0$ to imply that, in this case, measurements are the only source of uncertainty. Hereafter, we refer to the trajectories output by Talla as *raw* and to those post-processed via IMM-UKF as *filtered*. Fig. 9 exemplifies their difference, whose impact on stop positions (§6) and segmentations (§7) we analyze later.

5 EXPERIMENTAL SETUP: IN VITRO VS. IN VIVO

Ascertaining quantitatively the quality of stop-move detection entails conflicting dimensions we reconcile via two experimental campaigns with different goals. First, characterizing error sources (Fig. 2) is crucial to understand the attainable quality, the best configurations enabling it, and the related tradeoffs. Still, reliable findings require an *in-vitro* setup with known stop-move patterns and key variables fully captured by accurate ground truth. Second, an *in-vivo* setup with visitors

moving of their own volition is crucial to ascertain, albeit with slightly less accurate ground truth, whether the results above hold in a realistic setting, enabling a practical application of our findings and techniques.

5.1 Data Collection Methodology and Ground Truth

The target area contains a large globe in the center, surrounded by 6 tables hosting exhibits, whose layout was changed by the museum between the two campaigns. This prevents direct comparison of findings in vitro vs. in vivo, yet enables us to showcase the versatility of our metric.

In vitro. The tables host 44 exhibits of small size. Therefore, in this controlled campaign, we chose to model spatial objects (POIs) as points (Fig. 8a) that coincide with the ground-truth position where the user stops. Members of the research team emulate visitor behavior by alternating movement in the area with stops at given POIs, based on a known, pre-defined sequence. Each POI is visited only once.

Even in this highly controlled setup, collecting reliable ground truth is challenged by mobility. We obtained accurate *spatial* data by placing floor stickers near all designated POIs and acquiring their position with a laser meter. As for *temporal* ground truth, a smartphone application enables users to record arrival/departure times t_{start}, t_{end} at each POI. We synchronized the smartphone and Talla clocks, obtaining a common time reference for the timestamps in both ground truth and UWB trajectories. Moreover, we placed 2 tripod-mounted cameras with 180° angle on opposite sides covering the entire area, whose videos enabled cross-validation with the smartphone data.

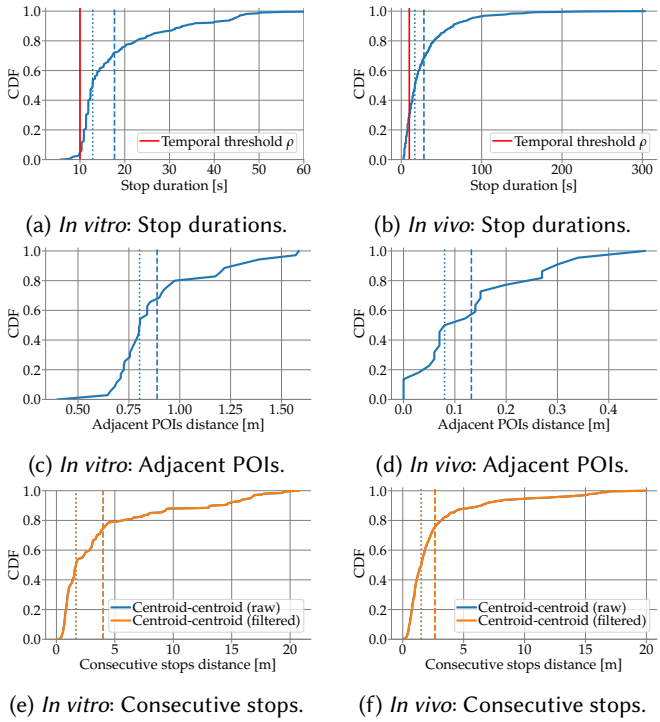
In vivo. In between the two campaigns, the museum refurbished the target area, now containing 35 exhibits delimited by colored rectangular tiles of variable size. This setup is more challenging, as tiles are often adjacent and rather large; POIs can no longer be modeled as points. However, tiles are meant to be observed by visitors from only one side (Fig. 8b); therefore, we modeled them as a linear spatial object.

We gather trajectories from real visitors recruited as volunteers by the museum and external to the research team. Informed consent and other standard procedures in place at MUSE for this type of studies were properly followed. The volunteers, who were all visiting the area for the first time, were given no constraints about their movement and asked to behave as they normally would during a museum visit. Therefore, unlike in vitro, people did not stop at designated coordinates, but anywhere in the target area. This unconstrained behavior also implies that a visitor may stop *away from all exhibits*, e.g., to talk with a friend or answer a phone call. To model this situation, we set the spatial object associated to a true stop to null ($\text{obj}(s) = \text{obj}(s_{real}) = \emptyset$, §1.1) whenever the distance $\|\text{pos}(s) - \text{pos}(o)\|$ between the true centroid from UWB trajectories and the POI exceeds the same threshold $R = 2$ m in Eq. (1), derived from empirical considerations based on the area layout.

Importantly, visitors were not involved in the gathering of ground truth. This was extracted solely from the videos recorded by the two cameras already exploited in the in vitro campaign, *without* the user-operated smartphone application. As a consequence of this and the fact that visitors are not supposed to be in a designated position as in the in-vitro case, temporal information is based on visual inspection, determining the time a visitor is near the exhibit tile. On the other hand, this setup greatly increases realism by minimizing the impact on the visitor behavior.

5.2 Dataset Description and Characterization

UWB trajectories, the input for segmentation, contain units in the form $(\mathbf{p}_i, t_i)$. For each timestamp t_i we collect both raw and filtered positions $\mathbf{p}_i = (x_i, y_i)$ (Fig. 9). The set S of true stops is a sequence of stops $s = \langle \mathbf{p}, [t_{start}, t_{end}], o \rangle$ whose position $\text{pos}(s)$ is derived as the centroid of the trajectory units within the real interval $[t_{start}, t_{end}]$, and whose associated POI is the real one (Fig. 2, §1.1).



(a) *In vitro*: Stop durations. (b) *In vivo*: Stop durations.

(c) *In vitro*: Adjacent POIs. (d) *In vivo*: Adjacent POIs.

(e) *In vitro*: Consecutive stops. (f) *In vivo*: Consecutive stops.

Fig. 11. Spatio-temporal characteristics of the dataset. In this figure and following ones, dashed lines represent the mean and dotted lines the median.

In vitro. We collected 9 trajectories of similar duration (~11 mins) for a total of 70,090 units over 100.03 mins. The number of true stops, known a priori, differs across trajectories and ranges from 11 to 29, for a total of 209 stops.

Fig. 11a shows *temporal* features via the cumulative distribution function (CDF) of true stop durations. In all the charts in Fig. 11 and in the rest of the paper, the mean and median are denoted by a dashed and dotted line, respectively. Our dataset deliberately contains very short stops, with a median of 12.4 s. The red line marks the threshold below which their duration is irrelevant for the application (§2), set to ρ=10 s based on requirements by the museum curators. The 10 stops (<5%) below ρ should not be detected, leaving 199 true stops as the ideal segmentation output (§2).

As for *spatial* features, POIs are very close (Fig. 8a). Fig. 11c shows that 80.6% of adjacent POIs are within 1 m from each other, and 1.59 m at most, demanding high spatial resolution in discerning stops with segmentation. Still, the distance between consecutive stops within a trajectory varies significantly (Fig. 11e) as exhibits are not necessarily visited in order. Fig. 10 offers an example by showing the same spatial information of the filtered trajectory in Fig. 9, this time augmented with temporal information. From this combined spatio-temporal view, we can see that the subject, mimicking visitors, mixes short strides to adjacent exhibits with longer ones, including one around the central globe.

In vivo. We tracked 10 volunteers across two rounds of 15 minutes each, yielding 219,937 units in the resulting 20 trajectories. Unlike in-vitro experiments, volunteers moved of their own volition, yielding more stops below ρ=10 s. Overall, trajectories contain between 11 and 29 true stops, for a

total of 392. Of these, 366 are associated to a POI, while the other 26 are not and therefore labeled as null (∅).

Fig. 11b shows the CDF of true stop durations, slightly longer w.r.t. in vitro, with a 17 s median and 303 s maximum. As for spatial features, in-vivo POIs are lines instead of points; their length ranges between 0.4 m and 2.7 m with a median of 1.09 m (Fig. 12). As a consequence of this relatively large size of POIs w.r.t. the tables hosting them (Fig. 8b), the distance between POIs is even smaller than in vitro. Adjacent POIs lay at a distance between 0 and 47 cm with a median of 8 cm (Fig. 11d), an order of magnitude lower than the median in vitro (81 cm). The distance between consecutive stops is also smaller in vivo, with a median of 14 cm (Fig. 11f) vs. 196 cm in vitro (Fig. 11e), due to closer adjacent POIs combined with real visitors often looking at nearby exhibits in sequence.

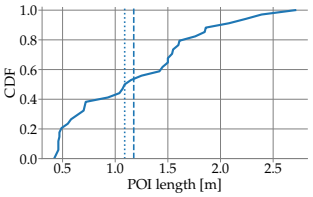


Fig. 12. *In vivo*: POI length.

6 UWB LOCALIZATION ERROR

Before delving into our findings, we exploit the controlled setup and accurate ground truth from the in-vitro one to ascertain the magnitude of the UWB positioning error, i.e., the distance $\|\text{pos}(s_{real}) - \text{pos}(s)\|$ between the real stops and the true ones derived as centroids from UWB trajectories, ① in Fig. 2.

Each UWB trajectory contains *several* positions for a stop segment (Fig. 9, in color), i.e., falling inside the interval $[t_{start}, t_{end}]$ whose ground-truth value is reliably determined via smartphone and cameras. Ideally, the UWB centroid $\text{pos}(s)$ matches exactly the real position $\text{pos}(s_{real})$ near a POI; in practice, this is not the case. In vitro, the $\text{pos}(s_{real})$ is manually measured very accurately, with a laser meter in a fixed and known position; $\text{pos}(s)$ is determined as the centroid of unit positions for a moving tag and with larger UWB errors. Their main source is the user body, creating non-line-of-sight (NLOS) conditions between the tag on the chest and the anchors behind the back. This is crucial when one of these anchors serves as time reference; manually changing it when in NLOS reduces the mean positioning error by 25%. NLOS mitigation techniques, an active research topic, could be incorporated in Talla and yield improvements. Notable approaches are *i)* the collection of multiple measurements, e.g., to estimate the noise distribution and incorporate it into the localization process [11], *ii)* NLOS detection, to dynamically select the best available anchors [39], and *iii)* error correction, to prevent NLOS from affecting localization in the first place. Detection and correction often exploit machine learning methods based on the channel impulse response (CIR) obtained from the radio [1, 21, 37]. Small models with short runtime have been recently proposed [13], which could be integrated in Talla without compromising its scalability. Nonetheless, the actual improvement they would unlock in our specific scenario is unclear, as most solutions target NLOS from walls or furniture. Few works have explored NLOS due to human occlusion specifically (e.g., [6, 25]), with promising results; however, their evaluation in the museum context is beyond the scope of this paper.

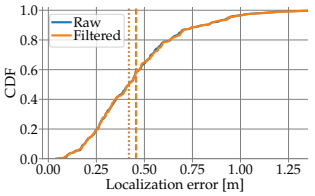


Fig. 13. *In vitro*: Positioning error.

Nevertheless, the error between $\text{pos}(s_{real})$ and $\text{pos}(s)$ (Fig. 13) remains sub-meter in 96.2% of the cases, i.e., significantly better than techniques based on WiFi and BLE, plagued by errors of several meters [45]. For both raw and filtered trajectories, the median and mean error are 42 cm and 46 cm, respectively; the commonly-used 75^{th} percentile is 57 cm in both cases. Interestingly, these metrics

are within few percents, i.e., *the smoothing induced by Kalman filters does not affect the position of the true stops.*

7 FINDINGS FROM IN-VITRO EXPERIMENTS

We study the quality of reference stop detection techniques (§2) against the yardstick of our novel metric (§3.2) in the controlled, in-vitro setup (§5) to distill quantitative findings eliciting the tradeoffs between configuration parameters and expected detection quality. Hereafter, we set equal weight of spatial and temporal dimensions, $\alpha = 0.5$, unless otherwise noted.

7.1 Novel Metric: Is It Worth?

We begin by confirming the higher expressiveness of our stop-centric metric w.r.t. unit-centric ones, observing that *segmentations with the same stop-centric quality can have different unit-centric quality, and vice versa.* For illustration only, we focus on SeqScan segmentations of sample trajectories.

Fig 14 shows the quality of two segmentations, obtained using different configuration parameters, of the same portion of a trajectory; the bottom pictures show the spatio-temporal views (akin to Fig. 10) for a sample stop. The unit-based metric assigns a higher F-score to $\langle 20, 36\rangle$; one could infer that it detects *more* stops than $\langle 10, 36\rangle$. Instead, both detect 21 true stops with one FN, as correctly captured by our metric by assigning the same F-score. The sample stop illustrates the reason; $\langle 20, 36\rangle$ detects a few more units matching the true stop, increasing F-score. However, this unit-centric perspective is misleading, as it mixes the correct association of estimated and true stops with their similarity. When considered separately, as in our metric, the overall quality of the two segmentations is the same, both for F-score and S-score. Further, with a purely temporal metric ($\alpha = 1$, not shown) the S-score is actually marginally higher for $\langle 10, 36\rangle$, accounting for the slightly more accurate overlapping between estimated and true stops exemplified in Fig 14.

Fig. 15 illustrates the dual situation on another trajectory and different segmentations; in each spatio-temporal view, different colors denote different stops. In this case, the unit-centric metric assigns them the same F-score = 0.90; yet, the spatio-temporal views, this time for the entire trajectory, show that their quality is *very* different. Segmentation $\langle 15, 24\rangle$ correctly detects all 28 true stops. Instead, $\langle 15, 12\rangle$ lumps 8 distinct true stops into 2 large estimated ones, representing incorrectly the user behavior. The unit-centric metric is oblivious to *structure*, considering whether individual units belong to *any* stop. Conversely, our stop-centric metric accounts for the 6 FN in $\langle 15, 12\rangle$ with a lower F-score than $\langle 15, 24\rangle$.

The higher expressiveness we concretely illustrated has practical implications. For instance, in our museum context, some analyses focus on *how many* exhibits are visited, others on *how long*

	Stop-centric					Unit-centric			
$\langle \epsilon, N \rangle$	TP	FP	FN	F-score	S-score	TP	FP	FN	F-score
$\langle 10, 36\rangle$	21	0	1	0.98	0.90	4348	483	1228	0.84
$\langle 20, 36\rangle$	21	0	1	0.98	0.90	5375	718	201	0.92

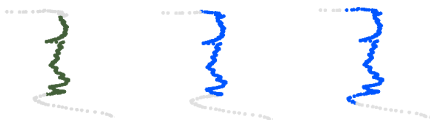
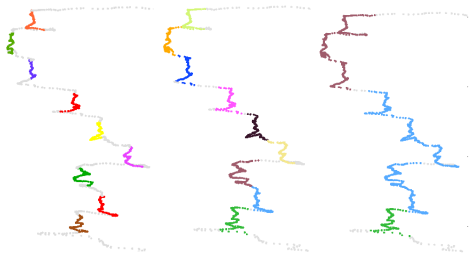


Fig. 14. *Same* stop properties, *different* unit quality. The figures represent the spatio-temporal view of a trajectory portion. Points with different colors represent units belonging to the true stop (green), or to the estimated ones (blue) determined with different SeqScan parameters; gray points are move units.

	Stop-centric					Unit-centric			
$\langle \epsilon, N \rangle$	TP	FP	FN	F-score	S-score	TP	FP	FN	F-score
$\langle 15, 24 \rangle$	28	0	0	1.00	0.88	4953	932	80	0.90
$\langle 15, 12 \rangle$	22	0	6	0.88	0.85	5013	1087	20	0.90



true $\langle 15, 24 \rangle$ $\langle 15, 12 \rangle$

Fig. 15. *Different* stop properties, *same* unit quality. The figures represent the spatio-temporal view of a trajectory portion, showing the true stops and two sets of estimated stops determined with different SeqScan parameters. In each picture, points with different colors are units belonging to different stops; same colors across pictures do not denote an association of stops across them; gray points are move units.

each visit is. Our metric captures and sharply separates the two via F-score and S-score, guiding the choice of the most appropriate technique and/or configuration.

7.2 Segmentation: Configuration and Quality

We dissect segmentation techniques through the lens of our stop-centric metrics, investigating at once the best configuration of each technique and the attainable quality.

Which parameters for what quality? We ascertained the impact on quality of several configurations for each technique; Table 2 shows a relevant subset for filtered trajectories. The highlighted best ones are those with highest F-score, e.g., $\langle 10, 12 \rangle$ and $\langle 15, 24 \rangle$ for SeqScan. Again, alternative criteria striking different quality tradeoffs are possible.

All methods yield good quality. KBV has the lowest and is the most sensitive to its θ parameter; yet, it is the cheapest computationally (§2). At the other extreme, SeqScan yields highest quality and its two-parameter configuration increases flexibility. Table 2 also reports the unit-centric metric, confirming its lower expressiveness. This is evident for KBV, whose highest unit-centric F-score is obtained with $\theta = 100$ that *i)* detects only 123 out of 199 true stops, yet *ii)* has nearly the same unit-centric F-score of the best SPD configuration, detecting 189. Therefore, hereafter we report only the results obtained with our stop-centric metric.

Raw vs. filtered trajectories: Does it matter? We first validate the behavior of IMM-UKF filters by inspecting the IMM mode probabilities and KF velocity in relation to true stop intervals. As expected, the predominant IMM mode in between stops is "moving" (Figure 16); instead, during a stop this probability rapidly drops and the KF velocity approaches zero.

Filtered trajectories reduce spatial jitter vs. raw ones (Fig. 9) yet induce the same stop-move structure (Fig. 11e). This is to be ascribed to the fact that most localization errors are due to NLOS. These typically causes a group of consecutive UWB units to spatially shift away significantly from previous ones, in a way that KF filters cannot correct; hence, similarly incorrect positions appear in both types of trajectories. Nonetheless, the "smoothing" induced by filtered trajectories may facilitate stop detection and provide better segmentation quality. To ascertain this, we performed for raw trajectories the same parameter exploration of Table 2, except for KBV. Table 3 shows

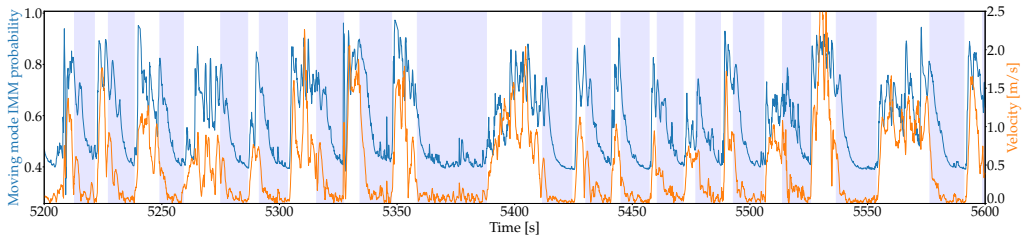


Fig. 16. IMM behavior on a trajectory excerpt. True stop intervals are highlighted.

Table 2. Exploring segmentation parameters and resulting quality for filtered trajectories; δ and ϵ are in cm, θ in cm/s.

	Stop-centric							Unit-centric
	TP	FP	FN	Precision	Recall	F-score	S-score	F-score
ϵ, N	SeqScan							
10, 12	186	8	13	0.959	0.935	0.947	0.891	0.871
15, 12	176	11	23	0.941	0.884	0.912	0.866	0.872
20, 12	165	13	34	0.927	0.829	0.875	0.842	0.869
10, 24	176	8	23	0.957	0.884	0.919	0.893	0.842
15, 24	187	9	12	0.954	0.940	0.947	0.883	0.870
20, 24	177	11	22	0.941	0.889	0.915	0.866	0.871
10, 36	166	6	33	0.965	0.834	0.895	0.893	0.824
15, 36	183	8	16	0.958	0.920	0.938	0.886	0.859
20, 36	178	10	21	0.947	0.894	0.920	0.871	0.869
δ	SPD							
20	117	16	82	0.880	0.588	0.705	0.837	0.630
30	161	16	38	0.910	0.809	0.856	0.873	0.777
40	175	19	24	0.902	0.879	0.891	0.882	0.822
50	181	19	18	0.905	0.910	0.907	0.885	0.835
60	189	18	10	0.913	0.950	0.931	0.871	0.854
70	187	21	12	0.899	0.940	0.919	0.855	0.850
80	187	18	12	0.912	0.940	0.926	0.847	0.850
90	183	17	16	0.915	0.920	0.917	0.835	0.849
100	176	22	23	0.889	0.884	0.887	0.820	0.849
110	173	24	26	0.878	0.869	0.874	0.803	0.847
130	161	26	38	0.861	0.809	0.834	0.772	0.840
150	155	21	44	0.881	0.779	0.827	0.756	0.833
θ	KBV							
10	31	4	168	0.886	0.156	0.265	0.783	0.42
20	110	6	89	0.948	0.553	0.698	0.853	0.658
30	154	4	45	0.975	0.774	0.863	0.870	0.719
40	168	11	31	0.939	0.844	0.889	0.875	0.801
50	175	10	24	0.946	0.879	0.911	0.873	0.831
60	168	10	31	0.944	0.844	0.891	0.849	0.844
70	162	9	37	0.947	0.814	0.876	0.826	0.846
80	142	7	57	0.953	0.714	0.816	0.798	0.847
90	129	9	70	0.935	0.648	0.766	0.768	0.851
100	123	11	76	0.918	0.618	0.739	0.751	0.855

how the best configurations for SPD and SeqScan respectively detect 8 and 10 fewer TP with an increase in FN. However, SPD requires a higher distance, $\delta = 80$, accounting for the higher position dispersion. In contrast, $\langle \epsilon = 15, N = 24 \rangle$ remains the best configuration in SeqScan, therefore more robust w.r.t. parameter sensitivity. Overall, these results confirm that filtered trajectories yield higher quality, although the difference is not dramatic.

What is the nature of false positives and negatives? Our metric offers further insights on quality by capturing the *type* of false detections (§3.2).

Table 3. *In vitro*: Quality of segmentation. The true stops are 199, plus 10 short ones (duration <10 s) that should not be detected.

Technique	Dataset	Best config.	TP	F-score	S-score	Nature of false detection					Spatio-temporal errors							
						FP			FN		t_{start} (s)		t_{end} (s)		Δt (s)		$\Delta\mathbf{p}$ (cm)	
						split	short	fake	merged	missing	μ	σ	μ	σ	μ	σ	μ	σ
SeqScan	filtered	⟨15, 24⟩	187	0.947	0.8833	2	6	1	5	7	-2.13	5.46	1.54	1.34	3.67	5.58	3.22	4.53
	raw	⟨15, 24⟩	176	0.921	0.8725	1	5	1	8	15	-3.37	5.50	1.63	2.30	5.00	6.24	5.76	32.63
SPD	filtered	60	189	0.931	0.8709	10	7	1	3	7	-0.67	4.29	1.32	3.93	1.99	6.63	4.09	6.07
	raw	80	181	0.903	0.8619	13	7	1	5	13	-0.97	4.64	1.36	4.34	2.32	7.38	4.52	6.33
KBV	filtered	50	175	0.911	0.8726	5	5	0	7	17	-0.17	6.81	1.53	2.63	1.71	7.97	4.19	8.08

The analysis of best segmentations (Table 2) shows that SPD detects nearly as many true stops as SeqScan (more, on raw trajectories) but with more false positives, lowering precision and F-score. Table 3 clearly shows that the culprit are split stops, a known weakness of the method. Also, all techniques are equally sensitive to stops with duration shorter than $\rho = 10$ s; this is actually the main source of mis-detection in SeqScan. However, this type of FP is particular and often benign (§3.2). At the other extreme, fake stops are rare, even absent with KBV.

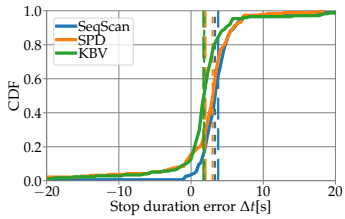
Dually, missing stops are the main source of FN for all techniques. SeqScan and SPD achieve similar results; the latter is more sensitive to spatial resolution. In SPD, a smaller δ does not affect split and merged stops but increases missing ones; with $\delta = 20$, they become the only source of FN. In contrast, the two-parameter structure of SeqScan achieves high quality with similar or even lower spatial resolution ϵ, slightly increasing merged stops in other configurations less performant than the one shown. Finally, the many missing stops in KBV are due to its reliance on velocity rather than distance, frequently changing around the threshold θ. This parameter crucially affects the nature of FN, dominant in KBV (Table 2); a value of 100 cm/s yields a majority (40) of merged stops, while 10 cm/s yields all missing stops.

What is the spatio-temporal error? The S-score is as a concise indicator of similarity between true and estimated stops. However, it does not account for the *actual* error in their temporal alignment and position, analyzed here.

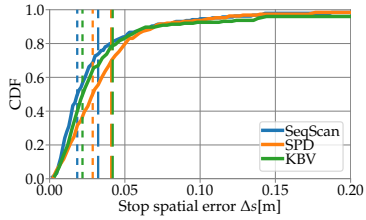
Given an estimated stop $\hat{s}$ and a true stop s, we compare their intervals $\mathsf{int}(\hat{s}) = [t_{\hat{s},start}, t_{\hat{s},end}]$ and $\mathsf{int}(s) = [t_{s,start}, t_{s,end}]$ by considering the errors in start time $t_{end} = t_{\hat{s},start} - t_{s,start}$, end time $t_{end} = t_{\hat{s},end} - t_{s,end}$, and duration $\Delta t = |\mathsf{int}(\hat{s})| - |\mathsf{int}(s)|$. Moreover, we consider the spatial error $\Delta\mathbf{p} = \|\mathrm{pos}(\hat{s}) - \mathrm{pos}(s)\|$, ② in Fig. 2. Table 3 reports their mean μ and standard deviation σ in the best configurations; Fig. 17 shows the CDFs of Δt and $\Delta\mathbf{p}$ of filtered trajectories only.

All techniques perform well, with errors of few seconds and centimeters, and small relative differences. KBV is the most accurate temporally, with a mean error μ=1.71 s. Yet, its mean spatial error $\Delta\mathbf{p}$ is the highest among filtered trajectories and close to SPD, whose median is significantly worse. At the other extreme, SeqScan yields the worst duration estimates; μ is nearly twice w.r.t. KBV, although the absolute difference is <2 s. Still, it is the most accurate spatially, a counterintuitive result explained by several factors: *i)* SeqScan is robust to outliers by design, intrinsically reducing spatial noise *ii)* temporal precision (σ) is the highest *iii)* t_{start} is *under*estimated and t_{end} *over*estimated, both in median (not shown) and mean, by nearly the same amount that tends to center the true stop inside the estimated one, reducing the distance between their centroids (Fig. 2).

Fig. 17a also shows that SPD often severely underestimates stop duration, likely the culprit for the many stop splits (Table 3). Nevertheless, its performance in terms of Δt and $\Delta\mathbf{p}$ does not change significantly when moving from filtered to raw trajectories. This is not the case for SeqScan, whose metrics for the latter (Table 3) are nonetheless heavily affected by a *single* outlier, caused by the

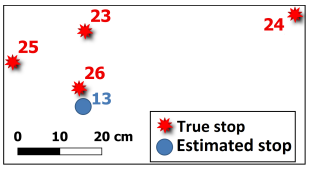


(a) Stop duration error.



(b) Stop spatial error.

Fig. 17. In-vitro: Duration and spatial error for filtered trajectories.



(a) True and estimated stops.

stop ID	start	end
True		
23	11:26:21:.047	11:26:35.972
24	11:26:36.043	11:27:01.965
25	11:27:02.037	11:27:15.962
26	11:27:16.033	11:27:28.959
Estimated		
13	11:26:19.976	11:27:29.709

(b) Stop durations.

Fig. 18. The spatio-temporal weight α may affect matching.

(a) True and estimated stops.

True stop ID	…	Estimated stop ID: 6	7
…	…	…	…
5	…	0.22	0.78
6	…	0	0.25

(b) The weight function $\mathcal{W}(e_{ij})$ with α=0 yields a *single* match and F-score = 0.703.

True stop ID	…	Estimated stop ID: 6	7
…	…	…	…
5	…	0.88	0.97
6	…	0	0.51

(c) The weight function $\mathcal{W}(e_{ij})$ with α=1 yields *two* matches and F-score = 0.757.

Fig. 19. The spatio-temporal weight α may affect the F-score.

merging of distant stops, whose removal yields μ=3.82 cm and σ=7.74 cm for $\Delta\mathbf{p}$. Anyway, this is in line with false detections (Table 3); while SPD is prone to stop splitting, SeqScan is to merging.

Sensitivity to the spatio-temporal weight α. We analyze the impact of values other than $\alpha = 0.5$, used until now, to weigh the spatial and temporal dimensions. Recall that α does not affect segmentation techniques and their *intrinsic* quality, but only their *evaluation* via our metrics (§3.2).

In our datasets, we observed two effects. First, *α may change the matching between estimated and true stop.* In Fig. 18, a small area contains multiple true stops. The purely spatial metric (α=0) matches the estimated stop to true stop #26. Instead, the purely temporal metric (α=1) matches it to #24, the one with the largest temporal overlapping. In this case, α affects the S-score, by capturing a different spatio-temporal similarity, but not the F-score, as the number of matches is unaltered. However, by affecting matching weights, *α may also affect the F-score.* In Fig. 19, the spatial metric yields an optimal assignment with a single match between the true stop #5 and the estimated one #7,

Table 4. *In vitro*: Quality of segmentation when associating estimated stops to POIs (filtered trajectories).

Technique	Best config.	TP	F-score	S-score	Nature of false detection							Spatio-temporal errors							
					FP				FN			t_{start} (s)		t_{end} (s)		Δt (s)		Δp (cm)	
					mislabeled	split	short	fake	mislabeled	merged	missing	μ	σ	μ	σ	μ	σ	μ	σ
SeqScan	$\langle 15, 24 \rangle$	180	0.911	0.883	8	0	6	2	11	1	7	-2.12	5.46	1.64	1.74	3.76	5.67	3.14	4.54
SPD	60	177	0.872	0.872	14	6	8	2	15	0	7	-0.78	4.56	1.33	3.65	2.11	6.64	4.08	6.20
KBV	50	168	0.875	0.873	8	2	6	1	13	1	17	-0.11	6.75	1.82	2.87	1.93	7.93	4.23	8.20

with F-score=0.703 (Fig. 19b). Instead, the temporal metric yields two matches (true stops #5 and #6 with estimated ones #6 and #7, respectively) and a higher F-score=0.757 (Fig. 19c).

Interestingly, when using the *best* segmentation configurations, these effects are very limited; only 1 change in matched stops for SeqScan, 0 for SPD, 3 for KBV, and an unchanged F-score. Instead, when using other configurations (Table 2) we see significant changes in our metrics and in matched stops, up to 9 for SeqScan, 13 for SPD, and 19 for KBV.

This confirms at once the soundness of our metrics and the good quality of the segmentation techniques evaluated through them. Indeed, an ideal segmentation would always find the *same* matches between estimated and true stops, yielding the same F-score irrespective of α. Dually, a marked change in F-score when changing α means that it swings matching across *many* candidates with wildly *different* spatio-temporal quality, possible only due to issues with true stops (e.g., poor temporal ground truth, high positioning noise) or estimated ones (e.g., poor segmentation).

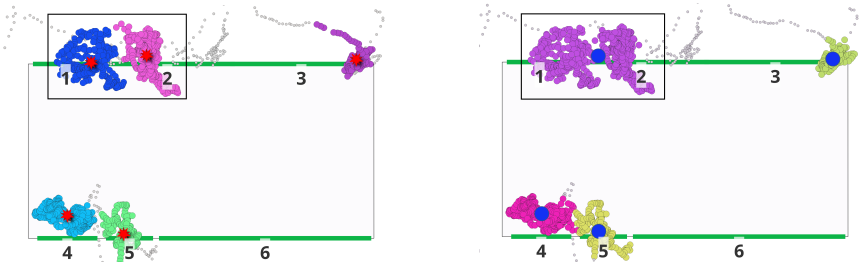
Consequently, spatio-temporal errors are also largely unaffected when using best configurations. Switching between purely temporal and spatial metrics yields differences in Δt and $\Delta \mathbf{p}$ of less than a second or centimeter, respectively. Still, as expected, these are always coherent with the dimension chosen, e.g., Δt is *always* lower with a temporal metric.

We observe that our approach, albeit motivated by the fine-grained spatio-temporal nature of UWB, is *general* and applicable to other positioning systems, likely affected by higher noise; however, this is outside the scope of this paper.

7.3 Associating Estimated Stops to POIs

We now close the circle and investigate the quality of segmentation techniques in correctly associating their estimated stops to the POIs represented as spatial objects with semantic annotations (§3.3). The mapping between stop and spatial objects is encoded by the O function in Eq. (1) that, once incorporated in the weight function $\mathcal{W}_o$ of Eq. (2), extends matching to account for semantic annotations. Note how, in the in-vitro campaign, a null association (obj(s) or obj($\hat{s}$) is $\varnothing$) never occurs in the pre-defined mobility pattern.

Table 4 reports the quality achieved by the best configurations with filtered trajectories. Quality remains very high: the best technique, SeqScan, correctly detects *and* associates 180 of the 199 true stops (90.5%) despite the challenges of *i)* POIs close to each other (Fig. 11c) and *ii)* non-negligible UWB positioning error (Fig. 13). Even KBV, the simplest, correctly detects 168 stops (84.4%). In all cases, quality is marginally lower w.r.t. Table 3; SPD shows a highest decrease in TP (-12) w.r.t. SeqScan and KBV (-7). As expected, mislabeling plays a big role in false detections, although its incidence is very limited. Finally, spatio-temporal errors are in line with those in Table 3.



(a) True stops: The visitor dwells in front of 5 exhibits.

(b) Estimated stops: Visits to two adjacent exhibits are lumped into a single one (top left).

Fig. 20. *In vivo*: Challenges of the POI layout. The numbers denote POIs, numbered from 1 to 6, representing exhibits on a table; the colored dots represent true and estimated stops.

Table 5. *In vitro* vs. *In vivo*: Top three configurations for each reference technique, in the same conditions as Table 2.

Dataset	Technique	Configurations	F-score	S-score
In vitro	SeqScan	$\langle 10, 12\rangle$,$\langle 15, 24\rangle$,$\langle 15, 36\rangle$	0.947, 0.947, 0.938	0.891, 0.883, 0.886
	SPD	60,80,70	0.931, 0.926, 0.919	0.871, 0.846, 0.855
	KBV	50,60,40	0.911, 0.891, 0.889	0.873, 0.848, 0.875
In vivo	SeqScan	$\langle 15, 24\rangle$,$\langle 15, 36\rangle$,$\langle 10, 12\rangle$	0.913, 0.909, 0.908	0.890, 0.896, 0.893
	SPD	130,120,100	0.834, 0.833, 0.825	0.833, 0.838, 0.852
	KBV	50,60,70	0.837, 0.821, 0.803	0.825, 0.824, 0.798

8 FINDINGS FROM IN-VIVO EXPERIMENTS

We ascertain to what extent our findings hold in the realistic in-vivo setup where users move of their own volition.

8.1 Challenges

The in-vivo setup poses significant challenges. First, the unconstrained user movement yields a *significantly higher fraction of true stops with duration <10 s*, to be discarded by segmentation. Compared to our in-vitro setting, where these accounted for 4.7% of the total (10 out of 209), in vivo they account for 28.7% (158 out of 550). When these are detected as estimated stops (with an incorrect duration ≥10 s) a FP occurs, albeit relatively benign. The question is whether this order-of-magnitude increase degrades quality with a corresponding increase in short FP.

Second, the layout of POIs, close to each other and of linear shape (§5.2), significantly complicates stop detection. Fig. 20 offers an example from our dataset, where the visitor true stops occur in front of 5 exhibits (Fig. 20a). Segmentation correctly estimates 3 of them; however, it lumps the other 2 into a single stop, yielding a false negative. A related issue affects the semantic matching of stops to POIs. For instance, the stop near POI #5 in the bottom left is correctly estimated (Fig. 20b); still, even a small spatial error could bring it closer to #4 or even #6, causing a mislabeled FP or FN error. These situations occur in reality, as visitors are not necessarily immobile when visiting an exhibit and the inherent noise from UWB localization further "spreads" trajectory units. True stops and their association to POIs can be manually identified based on ground truth, a luxury that cannot be afforded by *automated* segmentation techniques.

Table 6. *In vivo*. Quality of segmentation. The true stops are 392, plus 158 short ones (duration <10 s) that should not be detected.

Technique	Best config.	TP	F-score	S-score	Nature of false detection					Spatio-temporal errors							
					FP			FN		t_{start} (s)		t_{end} (s)		Δt (s)		Δp (cm)	
					split	short	fake	merged	missing	μ	σ	μ	σ	μ	σ	μ	σ
SeqScan	$\langle 15, 24 \rangle$	343	0.913	0.890	10	5	1	37	12	-2.68	7.63	1.06	9.40	3.75	12.27	7.31	9.11
SPD	130	362	0.834	0.833	59	53	2	30	0	-2.08	10.19	0.45	11.79	2.52	15.43	11.61	13.83
KBV	50	329	0.837	0.825	54	9	2	34	29	2.59	13.6	-1.31	12.86	-3.91	19.81	10.34	12.10

Table 7. *In vivo*: Quality of segmentation when associating estimated stops to POIs (filtered trajectories).

Technique	Best config.	TP	F-score	S-score	Nature of false detection							Spatio-temporal errors							
					FP				FN			t_{start} (s)		t_{end} (s)		Δt (s)		Δp (cm)	
					mislabeled	split	short	fake	mislabeled	merged	missing	μ	σ	μ	σ	μ	σ	μ	σ
SeqScan	$\langle 15, 24 \rangle$	309	0.823	0.890	37	7	5	1	51	20	12	-3.25	9.48	0.73	9.24	3.98	12.98	7.39	12.62
SPD	130	319	0.735	0.832	61	41	53	2	57	16	0	-2.86	14.15	-0.81	13.05	2.05	18.94	11.71	14.90
KBV	50	289	0.735	0.833	53	41	9	2	56	18	29	1.48	13.79	-2.18	12.60	-3.66	19.55	9.75	12.11

8.2 Segmentation: Configuration and Quality

Given these challenges, a crucial question is: *Can the shift to a real-world, uncontrolled scenario be tackled with the same configurations determined for the controlled one?* A positive answer would witness the robustness of segmentation techniques to input data, and validate our parameter tuning methodology.

The answer depends on the technique. Table 5 reports the top three configurations for each, ranked based on our metrics, for both in vitro and in vivo. SeqScan is very robust; its best configurations are the *same* with minor differences in F-score and S-score affecting only the ranking among configurations. Notably, $\langle 15, 24 \rangle$ was also the best one when using raw trajectories (Table 3); hence, it is the most versatile among those tested and the one we use hereafter. KBV is also affected only marginally, with two configurations providing the best results in both cases. In contrast, *none* of the best SPD in-vitro configurations remain the same in vivo. SPD is very sensitive to the "smoothness" of trajectories and requires an increase in the distance δ, already evident with raw trajectories and exacerbated here: the best in-vivo SPD configuration *doubles* the value of δ w.r.t. the in-vitro one.

Table 5 also directly compares quality that, as expected, degrades in vivo, yet remains very good. Again, SeqScan is the most robust, with only a minor F-score decrease (−3.6%). Instead, SPD shows a much higher decrease (−10.4%), making KBV (−8.1%) the second best technique.

Table 6 focuses on the best in-vivo configurations, where false detections offer interesting observations. First, despite the order-of-magnitude increase in true short stops, the short FP with estimated duration $\geq \rho$=10 s do not follow the same trend. Indeed, for SeqScan and KBV the absolute number of short FP remains nearly the same as in vitro; however, relative to the 158 true in-vivo short stops, these misdetections are a small fraction, 3.1% and 5.7%, respectively. The reason is that the (real) distribution of the duration of true short stops in vivo is more uniform, with durations $\ll$10 s; instead, the (artificial) in-vitro durations of true short stops are very close to the threshold, thus easier to estimate incorrectly. In contrast, SPD is very sensitive to short FP; the increase in misdetections (from 7 to 53) mirrors the order-of-magnitude increase in true short stops.

Table 8. Breakdown of the mislabeled errors in Table 7.

FP					
Technique	**Same table**			**Other table**	**Either is ∅**
	Same side		**Opposite side**		
	Adjacent	**Non-adjacent**			
SeqScan	32 (86.49%)	0	2 (5.4%)	2 (5.4%)	1 (2.7%)
SPD	52 (85.24%)	1(1.64%)	2 (3.28%)	4 (6.56%)	2 (3.28%)
KBV	43 (81.13%)	0	5 (9.43%)	2 (3.77%)	3 (5.66%)
FN					
Technique	**Same table**			**Other table**	**Either is ∅**
	Same side		**Opposite side**		
	Adjacent	**Non-adjacent**			
SeqScan	45 (88.23%)	0	2 (3.92%)	3 (5.88%)	1 (1.97%)
SPD	50 (87.72%)	3 (5.26%)	1 (1.75%)	1 (1.75%)	2 (3.51%)
KBV	47 (83.93%)	1 (1.78%)	5 (8.93%)	1 (1.79%)	2 (3.57%)

As for the other misdetections, fake FP remain very low, comparable to in vitro despite the twofold increase in true stops: *estimated stops that did not actually happen are extremely rare in practice*. Dually, *true stops are rarely missed by stop-move detection.* For SeqScan and KBV, the increase in missing FN (from 5 to 12 and from 17 to 29, respectively) almost directly mirrors the twofold increase in true stops, but is very small in absolute: only 3% and 7.4% of true stops are missed; missing stops are even absent with SPD.

This is not the case for split FP and merged FN. In the former case, SPD suffers from a 6× increase in misdetections, second only to the 10× of KBV. In SeqScan, despite the 5× increase the 10 split FP are a negligible fraction w.r.t. the 392 true stops, much smaller than the other techniques, in line with earlier findings (§7). As for merged FN, all techniques behave similarly, in line with in vitro albeit again with more errors. In both cases, the increase of these misdetections is likely due to the challenging layout (Fig. 20).

The slight decrease in quality affects the F-score but also the S-score, as reflected in spatio-temporal errors (Table 6). The average values μ remain similar to in vitro, apart from the spatial error $\Delta\mathbf{p}$ which shows a marginal increase, more marked for SPD and KBV. The standard deviation σ is instead 2–3 times higher, likely the culprit for the slight quality decrease in POI association, discussed next.

As for the impact of the spatio-temporal weight α, the findings in vitro are confirmed, with negligible changes in metrics, spatio-temporal errors, and matched stops: 2 for SeqScan, 9 for SPD, and 5 for KBV, out of 392 true stops.

8.3 Associating Estimated Stops to POIs

Table 7 shows the results of our analysis, analogous to the one in vitro (§7.3). At first, quality appears slightly worse, in line with the decrease above. In terms of F-score, SeqScan remains the best technique and the one with the smallest decrease (−9.6%); in contrast, both SPD and KBV degrade significantly (around −16% in both cases).

The culprit is clearly the increased number of mislabeled FP and FN; however, this aspect deserves further investigation. Table 8 analyzes the fraction of mislabeled FP and FN in Table 7 and differentiates the nature of mislabeling by distinguishing the cases where true and estimated POIs are *i)* adjacent *ii)* non-adjacent but on the same side *iii)* on different sides but on the same table *iv)* on different tables *v)* null (∅) in either case. For all techniques, the vast majority (between 81.1% and 88.2%) of mislabeled detections are to be ascribed to *adjacent* POIs. This is coherent with POIs being very close to each other (§5.2, Fig. 8b), leading segmentation techniques to mistake

Table 9. *In vivo.* Quality of segmentation when applied to detecting visits. The true visits are 310, of which 24 associated to a null POI.

Technique	config.	TP	F-score	S-score	Nature of false detection							Temporal errors					
					FP				FN			t_{start}(s)		t_{end}(s)		Δt(s)	
					mislabeled	split	short	fake	mislabeled	merged	missing	μ	σ	μ	σ	μ	σ
SeqScan	$\langle 15, 24 \rangle$	264	0.874	0.873	24	2	3	1	35	0	11	-2.59	11.84	0.95	13.77	3.54	17.76
SPD	70	270	0.840	0.838	38	12	11	2	32	0	8	-1.00	10.00	-1.13	13.77	-0.13	17.22
	130	266	0.799	0.834	38	7	43	2	44	0	0	-3.92	12.60	1.49	13.52	5.41	18.60
KBV	60	250	0.821	0.834	31	4	12	2	44	0	16	-0.48	12.34	1.18	14.37	1.65	19.12
	50	250	0.820	0.837	35	5	8	2	39	0	21	1.06	12.48	-1.50	13.31	-2.56	18.27

one for the other. Arguably, situations where a visitor stands across two exhibits (e.g., Fig. 20) are sometimes ambiguous even for a human observer.

Adjacent POIs aside, the content of the other columns shows that the fraction of grossly mislabeled POIs is negligible across all techniques. This is remarkable, given the practical relevance of associating stops with POIs in stop-move applications, for which our use case is no exception.

9 FROM STOPS TO VISITS

In vivo, we often noticed visitors making consecutive stops at the same POI, e.g., observing an exhibit from different points. Detecting these individual stops, our focus thus far, is useful for many analyses. However, when focusing on the time visitors *continuously* spend at a POI, e.g., to find the most engaging exhibits, individual stops are less relevant; the overall, higher-level *visit* is instead key. This intuitive notion is useful beyond our museum use case, e.g., in retail and recommender systems [27, 48]. Visits can be obtained via a simple aggregation of stops. Our contribution is to show that *i)* we can apply the *same* metrics defined for stops also to visits, and *ii)* ascertain the quality of the *visit detection* obtained with the reference segmentation techniques.

Modeling visits and their quality. Given a generic sequence of stops $Q = \langle s_1, \dots s_m \rangle$, we define a visit $v = \langle s_{start} \dots s_{end} \rangle \subseteq Q$ as a sequence of *consecutive* stops at the *same* spatial object o, $\forall s_i \in v \mid \text{obj}(s_i) = o$. A visit at a spatial object o is captured as $v = \langle o, [t_{start}, t_{end}] \rangle$, where the time interval extends from the beginning of the first stop s_{start} in the visit to the end of the last stop s_{end}. This applies to true and estimated stops, yielding *true visits* and *estimated visits*.

A visit can be regarded as a higher-level stop aggregating many others, where their stop centroids are replaced by the single POI associated to the visit. Crucially, this means that the POI-annotated version of our metric can be *directly* applied to the problem of matching true and estimated visits, including the quantitative measure of quality and the ability to determine the nature of false detection, both investigated next.

Quality of visit detection. Table 9 shows the results. For each technique, we report the best configuration used for stop detection (Table 7) along with the one yielding the best results for visit detection; notably, the two coincide for SeqScan, once again the most robust to parameter selection.

Interestingly, the quality of visit detection is generally higher than stop detection (Table 7). The reason is twofold. First, incorrectly fragmented stops (split FP) are now likely to be re-aggregated into a visit. Dually, *i)* merged stops are now part of the same visit, and *ii)* it is unlikely that a visit, typically much larger than a stop, is merged with another; merged FN are indeed absent for most configurations across segmentation techniques. Temporal errors are also entirely in line with those for stop detection. Note how we do not report the spatial error as it is meaningless in this context, as position is already abstracted away by the POI.

Table 10. Breakdown of the mislabeled errors in Table 9.

FP						
Technique	Config.	Same table			Other table	Either is ∅
		Same side		Opposite side		
		Adjacent	Non-adjacent			
SeqScan	⟨15, 24⟩	21 (87.5%)	0	0	2 (8.33%)	1 (4.17%)
SPD	70	26 (68.42%)	0	8 (21.05%)	2 (5.26%)	2 (5.26%)
	130	30 (78.95%)	1(2.63%)	2 (5.26%)	3 (7.89%)	2 (5.26%)
KBV	60	22 (70.97%)	0	6 (19.35%)	1 (3.23%)	2 (6.45%)
	50	27 (77.14%)	0	5 (14.29%)	1 (2.86%)	2 (5.71%)
FN						
Technique	Config.	Same table			Other table	Either is ∅
		Same side		Opposite side		
		Adjacent	Non-adjacent			
SeqScan	⟨15, 24⟩	32 (91.43%)	0	0	2 (5.71%)	1 (2.86%)
SPD	70	28 (87.5%)	0	1 (3.12%)	2 (6.25%)	1 (3.12%)
	130	39 (88.64%)	2 (4.54%)	0	1 (2.27%)	2 (4.54%)
KBV	60	39 (88.64%)	2 (4.54%)	2 (4.54%)	0	1 (2.27%)
	50	33 (84.61%)	1(2.56%)	2 (5.13%)	1(2.56%)	2 (5.13%)

Mislabeled false detections are again the main error contributor, although significantly fewer w.r.t. stops. For instance, when moving from stops to visits in SeqScan, mislabeled FP and FN decrease by −35% and −31%, respectively. Table 10 shows that these are dominated by POIs adjacent to the correct one, confirming our earlier observations (§8.3).

10 RELATED WORK

Stop-move detection. Segmentation has been applied to GPS trajectories for a long time [22]. Here, we considered techniques (§2) representative of mainstream classes: *criteria-based* define stops based on, e.g., distance, time duration, velocity, as in SPD, KBV and e.g., [4, 22]; *cluster-based* include SeqScan and, e.g., [20, 29, 33]. Others are based on statistical models [2] or do not even rely on segmentation [27, 28, 50]. Our focus is not on exhaustive comparison, rather on *enabling* a methodologically grounded comparison by *quantifying* the quality attainable when applying stop-move techniques to UWB and, via our novel metrics, elicit tradeoffs and challenges.

Semantic enrichment of trajectories. Often, stop-move detection is used as part of a larger process known as semantic enrichment [34], enhancing trajectory data with POIs and/or other context-dependent information, enabling higher-level analyses for outdoor trajectories, e.g., from GPS [18, 43]. Indoor trajectories are generally deemed problematic due their low quality [27]. We fill this gap by exploiting the higher-level accuracy of UWB for the semantic enrichment based on small-size POIs. The stops output, or their aggregation into visits, can supply the input for, e.g., symbolic trajectories [15].

Quality metrics. The evaluation of stop-move patterns vs. ground truth is hindered by the complexity induced by mobile targets. Validation is often only qualitative, e.g., through visual inspection of trajectories or simple metrics like the number of stops [19, 33]. This may suffice over large-scale areas with well-separated stops, but is challenging when stops are close in space and time, as in our setup.

In [2], units are assigned a stop/move label; detection quality is evaluated against ground-truth labels manually derived from GPS trajectories. However, due to the unit-centric binary labeling, the metric is oblivious to the number of stops. The use of F-score on stops instead of units was also proposed in [20, 29] but without specifying how true stops are accrued, likely deferring to qualitative considerations.

In these works, the relation to spatial objects, enabling the semantic dimension, is not considered. In contrast, the work in [27] defines a metric based on the percentage of units correctly labeled in indoor trajectories. Labels contain the event type along with the room containing the unit. In addition to this coarse granularity, the temporal extent of stop/move patterns are ignored, along with the impact of false observations. Similarity metrics have also been proposed [26] compare the movement of two individuals, e.g., based on event ordering; instead, we exploit a different notion of similarity focusing on segmentation correctness.

Stop-move detection in museums and beyond. Understanding the behavior of visitors is crucial in museums: the number, location, and duration of the visitors' stops are indicators of attention and hence interest [35]. This was recognized as early as 1935, when Melton [31], an experimental psychologist, pioneered the practice of "timing and tracking", a systematic observation of visitors by an external observer manually recording their movements. Simple and accurate, this practice is still used today [24, 42], but the significant human effort needed for large areas or visitor numbers limits the analyses building atop it.

Recently, sensors enabled automation and scalability at the price of accuracy. Reported experiences rely on Bluetooth, whose meter-level positioning [45] forces coarse spatio-temporal granularity, room-level and with uncertain timing. The overall stay (hours) and frequency of visits to key areas of the Louvre is analyzed in [44]; others analyzed the stay in a room [5] or "hotspots" and other macro-indicators [30]. These approaches extract spatio-temporal features from the Bluetooth signal, increasing complexity and yielding coarse accuracy. In contrast, we showed quantitatively how the higher accuracy of UWB directly yields much greater spatio-temporal resolution in discriminating stops, achieving automation *and* accuracy at once.

Finally, although our experimental evaluation focuses only on a museum scenario due to the prohibitive administrative, logistic, and deployment effort to replicate it elsewhere, any other domain where measuring the presence of an individual with near POIs is important could benefit from our work. Among these, retail [7] is an obvious example bearing a strong resemblance to museums; the quantitative insights offered by fine-grained stop-move detection near products on display, akin to visitors of museum exhibits, once combined with results by consumer research [38] would provide store owners with actionable information about their customers.

11 CONCLUSIONS

The high spatio-temporal accuracy of UWB localization intuitively enables fine-grained detection of stop-move patterns, key to many applications. Yet, this opportunity has not been studied, let apart quantitatively and experimentally. This is our goal, exploiting two experimental campaigns with different characteristics in a museum deployment with accurate ground truth, also rare in the literature.

Key to our contribution is a novel, expressive *family* of metrics whose *stop-centric* view is *significantly more expressive* than the state of the art. We provide definitions and methods enabling automated quantitative analysis and interpretation of false detections. Moreover, we encompass both the temporal and spatial aspects of stops, their semantic enrichment based on association to spatial objects, and the corresponding higher-level notions of visit, *reuniting all relevant dimensions in a single methodological framework.*

We apply these metrics to configure and compare representative segmentation techniques originally targeting coarser-grained scenarios. We show that, when used with UWB trajectories, they induce spatio-temporal errors of only few centimeters and seconds, enabling very high correspondence of estimated and true stops despite the inevitable UWB positioning error.

This high quality shows only marginal differences between the controlled in-vitro setting and the uncontrolled in-vivo one. This confirms that stop-move detection can be successfully exploited

towards the fine-grained resolution enabled by UWB. Of course, our findings are limited to a given museum use case and specific environment; the effort to gather ground truth with real visitors is daunting. Nevertheless, we hope that our technical and methodological contributions inspire similar studies in other application domains and environments, ultimately helping pave the way to a new generation of mobility analysis techniques seizing the new opportunities offered by UWB, a goal we also facilitate by releasing our datasets publicly [17].

ACKNOWLEDGMENTS

We are grateful to M. Lanzingher, Director of MUSE, and V. Cozzio, Head of IT services, for making this study possible, and to D. Dal Piaz and D. Tombolato for their support. At our institutions, we thank A. Bacchiega, M. Fenu, A. Giovannone, T. Istomin, and D. Molteni for their help on technical and experimental issues. This work is partially supported by the Italian government via the NG-UWB project (MUR PRIN 2017), and by the project SERICS (PE00000014) and the ICSC National Research Centre for High Performance Computing, Big Data and Quantum Computing (CN00000013), both under the NRRP MUR program funded by the NextGenerationEU. The views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the Italian MUR. Neither the European Union nor the Italian MUR can be held responsible for them.

REFERENCES

[1] S. Angarano, V. Mazzia, F. Salvetti, G. Fantin, et al. 2021. Robust ultra-wideband range error mitigation with deep learning at the edge. *Engineering Applications of Artificial Intelligence* 102 (2021), 104278.

[2] L. Bermingham and I. Lee. 2018. A Probabilistic Stop and Move Classifier for Noisy GPS Trajectories. *Data Mining and Knowledge Discovery* 32, 6 (2018), 1634–1662.

[3] H. A. Blom and Y. Bar-Shalom. 1988. The Interacting Multiple Model Algorithm for Systems with Markovian Switching Coefficients. *IEEE Transactions on Automatic Control* 33, 8 (1988), 780 –783.

[4] M. Buchin, A. Driemel, M. van Kreveld, and V. Sacristan. 2011. Segmenting trajectories: A framework and algorithms using spatiotemporal criteria. *Journal on Spatial Information Science* 3 (2011), 33–63.

[5] P. Centorrino, A. Corbetta, E. Cristiani, and E. Onofri. 2021. Managing Crowded Museums: Visitors Flow Masurement, Analysis, Modeling, and Optimization. *Journal of Computational Science* 53 (2021), 101357.

[6] Y. Chen, J. Wang, and J. Yang. 2024. Exploiting anchor links for NLOS combating in UWB localization. *ACM Transactions on Sensor Networks* 20, 3 (2024), 1–22.

[7] FIRA Consortium. 2024. Smart Retail. https://www.firaconsortium.org/discover/use-cases/smart-retail. Last access: November 1, 2024.

[8] M.L. Damiani, F. Hachem, H. Issa, N. Ranc, et al. 2018. Cluster-based Trajectory Segmentation with Local Noise. *Data Mining and Knowledge Discovery* 32 (2018), 1017–1055.

[9] M.L. Damiani, H. Issa, G. Fotino, M. Heurich, et al. 2016. Introducing 'Presence' and 'Stationarity Index' to Study Partial Migration Patterns: an Application of a Spatio-temporal Clustering Technique. *International Journal of Geographical Information Science* 30, 5 (2016), 907–928.

[10] DecaWave Ltd. 2017. DW1000 Data Sheet, version 2.19.

[11] C. Di Franco, A. Prorok, N. Atanasov, B. Kempke, et al. 2017. Calibration-free Network Localization Using Non-line-of-sight Ultra-wideband Measurements. In *Proc. of ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*.

[12] M. Ester, H.P. Kriegel, J. Sander, and X.Xu. 1996. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proc. of ACM International Conference on Knowledge Discovery & Data Mining (KDD)*.

[13] M. Gallacher, M. Stocker, M. Baddeley, K. Römer, et al. 2023. InSight: Enabling NLOS Classification, Error Correction, and Anchor Selection on Resource-Constrained UWB Devices. In *Proc. of International Conference on Embedded Wireless Systems and Networks (EWSN)*. Association of Computing Machinery.

[14] R.H. Güting, M.H. Böhlen, M. Erwig, C.S. Jensen, et al. 2000. A foundation for representing and querying moving objects. *ACM Transactions on Database Systems (TODS)* 25, 1 (2000), 1–42.

[15] R.H. Güting, F. Valdés, and M.L. Damiani. 2015. Symbolic trajectories. *ACM Transactions on Spatial Algorithms and Systems (TSAS)* 1, 2 (2015), 1–51.

[16] F. Hachem, D. Vecchia, M.L. Damiani, and G.P. Picco. 2022. Fine-grained Stop-Move Detection in UWB-based Trajectories. In *Proc. of IEEE International Conference on Pervasive Computing and Communications (PerCom).*

[17] F. Hachem, D. Vecchia, M.L. Damiani, and G.P. Picco. 2025. UWB Trajectories and Fine-grained Stop-move Detection: A Museum Dataset. Zenodo. https://doi.org/10.5281/zenodo.14918763 Version 1.

[18] Y. Hu, S. Ruan, Y. Ni, H. He, et al. 2021. SALON: A Universal Stay Point-Based Location Analysis Platform. In *Proc. of International Conference on Advances in Geographic Information Systems (SIGSPATIAL).*

[19] S. Hwang, C. Evans, and T. Hanke. 2017. Detecting Stop Episodes from GPS Trajectories with Gaps. In *Seeing Cities Through Big Data: Research, Methods and Applications in Urban Informatics.* Springer, 427–439.

[20] S. Hwang, C. VanDeMark, N. Dhatt, S. V. Yalla, et al. 2018. Segmenting human trajectory data by movement states while addressing signal loss and signal noise. *International Journal of Geographical Information Science* 32, 7 (2018), 1391–1412.

[21] C. Jiang, J. Shen, S. Chen, Y. Chen, et al. 2020. UWB NLOS/LOS classification using deep learning method. *IEEE Communications Letters* 24, 10 (2020), 2226–2230.

[22] J. H. Kang, W. Welbourne, B. Stewart, and G. Borriello. 2004. Extracting places from traces of locations. In *Proc. of ACM International Workshop on Wireless Mobile Applications and Services on WLAN Hotspots.*

[23] A. Kontarinis, K. Zeitouni, C. Marinica, D. Vodislav, et al. 2021. Towards a semantic indoor trajectory model: application to museum visits. *Geoinformatica* 25, 2 (2021), 311–352.

[24] D. Kosmopoulos and K. Tzortzi. 2021. Visitor Behavior Analysis for an Ancient Greek Technology Exhibition. In *Proc. of IFIP International Conference on Artificial Intelligence Applications and Innovations.*

[25] V. A. Minh Le, M. Trobinger, D. Vecchia, and G. P. Picco. 2022. Human Occlusion in Ultra-wideband Ranging: What Can the Radio Do for You?. In *Proc. of International Conference on Mobility, Sensing and Networking (MSN).*

[26] A. L. Lehmann, L. O. Alvares, and V. Bogorny. 2019. SMSM: A similarity measure for trajectory stops and moves. *International Journal of Geographical Information Science* 33, 9 (2019), 1847–1872.

[27] H. Li, H. Lu, G. Chen, K. Chen, et al. 2020. Toward Translating Raw Indoor Positioning Data into Mobility Semantics. *ACM Transactions on Data Science* 1, 4 (2020), 1–32.

[28] Z. Li, J. Han, M. Ji, L. Tang, et al. 2011. MoveMine: mining moving object data for discovery of animal movement patterns. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2, 4 (2011), 1–32.

[29] T. Luo, X. Zheng, G. Xu, K. Fu, et al. 2017. An Improved DBSCAN Algorithm to Detect Stops in Individual Trajectories. *ISPRS International Journal of Geo-Information* 6, 3 (2017), 63.

[30] C. Martella, A. Miraglia, J. Frost, M. Cattani, et al. 2017. Visualizing, Clustering, and Predicting the Behavior of Museum Visitors. *Pervasive and Mobile Computing* 38 (2017), 430–443.

[31] A. Melton. 1935. Problems of Installation in Museums of Art. *American Association of Museums* (1935).

[32] OptiTrack. 2024. OptiTrack. https://optitrack.com. Last access: October 16, 2024.

[33] A. Palma, V. Bogorny, B. Kuijpers, and L.O. Alvares. 2008. A Clustering-based Approach for Discovering Interesting Places in Trajectories. In *Proc. of ACM Symposium on Applied Computing.*

[34] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, et al. 2013. Semantic trajectories modeling and analysis. *ACM Computing Surveys (CSUR)* 45, 4 (2013), 1–32.

[35] B. Serrell. 1997. Paying Attention: The Duration and Allocation of Visitors' Time in Museum Exhibitions. *Curator: The Museum Journal* 40, 2 (1997), 108–125.

[36] S. Spaccapietra, C. Parent, M. L. Damiani, J. A. de Macedo, et al. 2008. A conceptual view on trajectories. *Data & Knowledge Egineering* 65, 1 (2008), 126–146.

[37] M. Stocker, M. Gallacher, C.-A. Boano, and K Römer. 2021. Performance of support vector regression in correcting UWB ranging measurements under LOS/NLOS conditions. In *Proc. of Workshop on Benchmarking Cyber-Physical Systems and Internet of Things.*

[38] P. Underhill. 1999. *Why We Buy: the Science of Shopping.* Simon & Schuster.

[39] B. Van Herbruggen, J. Fontaine, and E. De Poorter. 2021. Anchor pair selection for error correction in Time Difference of Arrival (TDoA) Ultra Wideband (UWB) positioning systems. In *Proc. of IEEE International Conference on Indoor Positioning and Indoor Navigation (IPIN).*

[40] D. Vecchia, P. Corbalán, T. Istomin, and G. P. Picco. 2019. TALLA: Large-scale TDoA Localization with Ultra-wideband Radios. In *Proc. of International Conference on Indoor Positioning and Indoor Navigation (IPIN).*

[41] S. Wang, Z. Bao, J.S. Culpepper, and G. Cong. 2021. A Survey on Trajectory Data Management, Analytics, and Learning. *ACM Computing Surveys (CSUR)* 52, 2 (2021), 1–36.

[42] S. Yalowitz and K. Bronnenkant. 2009. Timing and Tracking: Unlocking Visitor Behavior. *Visitor Studies* 12, 1 (2009), 47–64.

[43] Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra, et al. 2011. SeMiTri: A Framework for Semantic Annotation of Heterogeneous Trajectories. In *Proc. of International Conference on Extending Database Technology (EDBT).*

[44] A. Yoshimura, A. Krebs, and C. Ratti. 2017. Noninvasive Bluetooth monitoring of visitors' length of stay at the Louvre. *IEEE Pervasive Computing* 16, 2 (2017), 26–34.
[45] F. Zafari, A. Gkelias, and K. K. Leung. 2019. A Survey of Indoor Localization Systems and Technologies. *IEEE Communications Surveys & Tutorials* 21, 3 (2019), 2568–2599.
[46] F. Zafari, I. Papapanagiotou, and K. Christidis. 2015. Microlocation for Internet-of-Things-Equipped Smart Buildings. *IEEE Internet of Things Journal* 3, 1 (2015), 96–112.
[47] Y. Zheng. 2015. Trajectory Data Mining: An Overview. *ACM Transactions on Intelligent Systems and Technology* 6, 3 (2015), 1–41.
[48] Y. Zheng, L. Zhang, Z. Ma, X. Xie, et al. 2011. Recommending Friends and Locations Based on Individual Location History. *ACM Transactions on the Web (TWEB)* 5, 1 (2011), 1–44.
[49] Y. Zheng and X. Zhou. 2011. *Computing with spatial trajectories.* Springer.
[50] Y. Zhuang, J. Yang, Y. Li, L. Qi, et al. 2016. Smartphone-based indoor localization with Bluetooth Low Energy beacons. *Sensors* 16, 5 (2016), 596.
[51] E. Zimányi and M. Sakr. 2020. MobilityDB: A Mobility Database Based on PostgreSQL and PostGIS. *ACM Transactions on Database Systems* 45, 4 (2020), 1–42.