2018 International Conference on Indoor Positioning and Indoor Navigation (IPIN), 24-27 September 2018, Nantes, France

Evaluation of indoor localisation systems: comments on the ISO/IEC 18305 standard

Francesco Potortì, Antonino Crivello, Paolo Barsocchi, Filippo Palumbo
Information Science and Technologies Institute - National Research Council of Italy
{Potorti, antonino.crivello, paolo.barsocchi, filippo.palumbo}@isti.cnr.it

***Abstract*—Indoor localisation systems have been studied in the literature for more than ten years and are starting to approach the market. The absence of standard evaluation methods is one of the obstacles to their adoption outside of customised environments. Specifically, the definition of benchmarking methodologies, common evaluation criteria, standardised methodologies useful to developers, testers, and end users is an open challenge. The need for common benchmarks has been tackled by some initiatives in recent years: EvAAL, EVARILOS, the Microsoft competition and the IPIN competition. The first formal attempt at defining a standard methodology to evaluate indoor localisation systems is the ISO/IEC 18305:2016 International Standard, which defines a complete framework for performing Test&Evaluation of localisation and tracking systems. This work is a first critical reading of the standard, intended to be a key contribution to the activities of the International Standards Committee of IPIN.**

***Index Terms*—Performance evaluation, Indoor localisation systems, ISO/IEC 18305:2016, IPIN-ISC**

I. INTRODUCTION

The goal of an Indoor Localisation System (ILS) is to detect a target, that is a person or object, inside an indoor environment. In contrast to a mature technology like Global Navigation Satellite System (GNSS), there is no de-facto standard for indoor environments. Proposed ILS technologies are very heterogeneous, spanning from inertial sensors [1] to radio-frequency (RF) based sensors [2], light sensors and computer vision [3], among others.

At the moment, no generally accepted solutions exist, but customised solutions are starting to hit the market. It may be argued that one of the gaps to market acceptance is the lack of standardisation in ILS design and ILS evaluation. IPIN-ISC[1], which stands for IPIN International Standards Committee, is an effort that aims at filling this gap. This paper is intended to give a head towards the definition of common concepts, metrics and procedures for ILS test and evaluation.

The first step towards defining a benchmark for ILS was the EvAAL[2] competition, launched by the European FP7 universAAL project in 2010 as a way of "Evaluating Ambient Assisted Living systems through competitive benchmarking". In 2011, the first EvAAL competition was held at the CIAmI Living Lab in Valencia (ES), with a single track devoted to Indoor Localisation and Tracking. The EvAAL project continued its activities until 2013. Metrics included accuracy both as the point error and zone detection, timely availability of estimates, installation complexity, user acceptance and interoperability with existing technologies [4].

In 2014, EVARILOS[3] marked a second milestone as a benchmarking methodology in the indoor localisation context [5]. The project was born as a European FP7 project and its main goal was to build a benchmarking tool able to consider accuracy, complexity, cost, energy and RF interference resilience metrics.

In that same year, two leading conferences in the field, the International Conference on Information Processing in Sensor Networks (IPSN) and the International conference on Indoor Positioning and Indoor Navigation (IPIN), launched each their own indoor localisation competitions: respectively the Microsoft indoor localisation competition [6] and the IPIN competition, the latter exploiting the legacy of EvAAL by adopting what is currently known as the *EvAAL framework* [7], [8]. Each year, both competitions offer the research community a real and challenging test site where independent evaluation is performed. Both, as a long-term goal, plan to identify useful comparison criteria for ILS solutions.

While technologies become more mature and first customised solutions start to appear, the need for common evaluation criteria and standardised procedures useful to developers, testers and end user is more and more apparent. Standard test & evaluation procedures are needed to add transparency to the market and to build and nurture stakeholders' trust. In other words, standards are needed for a key role: providing commerce and manufacturing with a common language and common evaluation & test tools.

In 2016 the first such standard appeared: the ISO/IEC 18305:2016 International Standard[4] provides a standard methodology for evaluating indoor localisation systems and detailed Test&Evaluation (T&E) procedures for Localisation and Tracking Systems (LTSs) [9].

This paper is a critical reading of the ISO/IEC 18305:2016 standard and provides an insight into its strengths and shortcomings. In the process, we evaluate the compliance of the criteria applied during the IPIN and Microsoft competitions with the ones proposed by the standard.

This work is intended as a basis for the activities of the International Standards Committee (ISC) of IPIN[5], specifically the Evaluation Methodologies subcommittee. IPIN-ISC

---

[1] http://ipin-standards.org/

[2] http://evaal.aaloa.org/

[3] http://www.evarilos.eu/

[4] https://www.iso.org/standard/62090.html

[5] http://www.ipin-standards.org/

is committed to develop and promote open standards for indoor positioning and indoor navigation through collaboration of academia, industry and government organisations around the globe.

II. ANALYSIS OF THE ISO/IEC 18305 STANDARD

The ISO/IEC 18305:2016 standard was prepared by the Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 31. ISO/IEC JTC 1/SC 31 is a standardisation subcommittee of the joint committee ISO/IEC JTC 1 of the International Organisation for Standardisation (ISO) and the International Electrotechnical Commission (IEC), which develops and facilitates international standards, technical reports, and technical specifications in the field of automatic identification and data capture techniques. Most work on the development of the analysed standard has been done by working group WG 5 *Real time location systems*, that in 2015 has been moved into the broader WG 4 *Radio communications (RFID, RTLS, Security)*.

The standard is one of the results of the dissemination and exploitation plan of the EU FP7 project EVARILOS – Evaluation of RF-based Indoor Localisation Solutions for the Future Internet [10]. The project was funded under the topic ICT-2011.1.6 – Future Internet Research and Experimentation (FIRE) with the aim of developing and validating standardised experiment-based benchmarks to allow a fair comparison of different localisation solutions, not only under ideal, but also under extreme conditions (e.g. environments with heavy RF interference from co-located wireless devices).

Leveraging the experience gained in the field of RF-based ILSs [11], [12], the purpose of the standard is to define a Test&Evaluation procedure for generic Indoor Localisation Systems under different scenarios. The procedure is quite detailed: for example it prescribes up to 14 different scenarios and 5 types of buildings where the test should be carried on, and up to 30 metrics for assessing the performance in each test case.

The standard does not address custom-made test and evaluation practices, but it is focused on a system-level, black-box testing approach. This means that T&E is performed without any knowledge of inner workings of the ILS under test, nor of the principles or sensors on which it is based. The standard does not consider other T&E and evaluation purposes, like those that would be useful to system developers or the testers themselves.

In the following, detailed comments are given on the most interesting areas covered by the standard.

A. Vocabulary, overview, failure modes

The standard defines an appropriate vocabulary, by distinguishing for example between *positioning*, for an *ELT* (entity to be localised / tracked) who needs to know its own position, and *locating*, when a different entity gets to know the position of an ELT. In the end, the document adopts the common *localisation* term for indicating both functionalities.

TABLE I
NUMBER OF TESTS FOR COMBINATIONS OF ELT TYPE AND BUILDING TYPES

	House	Office	Warehouse	High-rise	Subterranean
Object	2	5	6	5	3
Person	3	5(+3)	7	6(+3)	5
Robot	ND	ND	ND	ND	ND

The standard gives a good overview on the architecture of common localisation systems, methods and sensors used, and an interesting appendix that does a good job of enumerating various failure modes of many sensors. Both are rather comprehensive, but for some reason lack any references to radio-based Device-Free Localisation (DFL) systems [13]. This omission is especially noticeable given that the standard is particularly attentive to fire-fighter scenarios, where DFL systems, especially tomographic ones [14], are in principle attractive.

B. ELT and building types, scenarios, mobility

Object, *person* and *robot* are the three kinds of devised ELTs. Tests should be performed separately for each ELT type, if appropriate for the scenario and the system under test.

Five building types are defined: *single-family house*, *office building*, *warehouse*, *high-rise steel*, *subterranean*. Tests should be performed in all types of building, if appropriate

The standard defines 14 different scenarios for performing tests. Scenarios are definitions of the type of mobility and number of involved ELTs. For example, one scenario is about measuring coordinates statically (after standing still for 3 s), one other does the same while walking, others while pushing a cart, others while moving together with other people in a predefined way. Scenarios are well thought-out and include mobility specifications, such as definitions of walking, running and so on. A complete description of requirements for placing test points is given, including formal descriptions of how to choose points, what is intended by randomness, distance and space density of test points. However, no scenarios are defined for robots in the standards, so the robot ELT type is apparently a placeholder for future extensions.

A standard-compliant system test should perform a separate test for each scenario. This is a heavy requirement for some systems. For example, Table I lists how many scenarios are defined for each of the 15 combinations of ELT type and building type. The last row is empty because no scenarios are defined for robots. The numbers in parentheses show the additional number of tests required for scenarios where special movements are tested: running, walking backwards, crawling, which are especially useful for firefighter applications.

It should be noted that a high number of tests is required for very generic systems only. Most usually, a system will have some kind of target use case, like being installed in office buildings, or houses, or warehouses, so the number of tests required is much smaller than the sum of all numbers in table I. The same can be said of the ELT type: usually the ILS under test will be specialised for positioning objects, or for tracking persons, so only one row and one or few columns of table I need to be used for the evaluation of a given ILS.

It is interesting to note that apartments are missing from the list of test buildings. Distinguishing an apartment in a building from a single-family house is significant, both because of building differences (materials, structure) and the interaction between similar systems in adjacent apartments that should be evaluated in the case of apartments.

C. Metrics

While the overall framework described by the standard is comprehensive and well thought out, the section on metrics, which is arguably its core, needs some in-depth discussion.

The metrics considered by the standard are various statistics applied to point errors, that is, to errors measured from the ELT position to a series of test points which is defined as the ground truth. The standard does not consider other approaches related to the overall trajectory, such as the Fréchet distance [15]. We think that the reason behind using point error statistics instead of comparing trajectories [16], [17] is that the latter is less adequate to navigation purposes, for which the real-time identification of the position is more important than the path followed.

The density of test points suggested by the standard is one per 5–10 m^2 in single-family houses and one per 50–100 m^2 elsewhere, so the number of points per floor generally lies in the range of 50–200.

The considered metrics are based on various sets of statistics, as follows. We comment on the various metrics, by raising doubts on the usefulness of some of them and by suggesting possible modifications. When we propose a precise modification to the standard we use the term *should*.

While it is true that computing two statistics or thirty requires essentially the same amount of work, we believe that the usefulness of a standard is greatly improved if only a small set of results is produced, so that it is clear what is important for characterising a given system. This is especially true for normative references based on the standard, a case when some normative body asks for a system that satisfies at least some performance criteria: having 30 possible indexes to choose from makes it difficult to compare performance criteria.

In the following, we object to the usefulness of some metrics on the basis that those are not appropriate to a system-level, black-box type testing and evaluation, which is the focus of the standard. Some of the metrics to which we object are very useful for system debugging, others for system tuning after installation and periodically afterwards, others are appropriate to validate the correctness of the testing procedure itself: in none of these cases are the metric useful to the final user of the system.

In order to improve the credibility of the test report, the standard *should* mandate that the presentation of numerical results should include not only the number of samples, as already specified, but also the amplitude of a 90% confidence interval for each figure. The amplitude of the interval shall be discussed, a reasonable requirement would be for the interval to be not wider than $\pm$ 20% around the reported value.

As a last note on metrics in general, the standard *should* recommend minimum and maximum performance requirements for various types of systems. For example, for systems targeted to a specific type of ELT it makes no sense to require an accuracy smaller than approximately half the size of the ELT itself if it is static, or smaller that the size of the ELT it it can move; for example, requiring an accuracy of 20 cm (or anything smaller than 60 cm) for a moving person is unnecessarily strict. On the other hand, it makes no sense to require an accuracy larger than half the average size of the interesting areas of the building where the system is deployed; for example, requiring an accuracy of 20 m (or anything larger than 4 m) for a family house is excessively loose.

The next subsections contain comments on all the metrics defined by the standard, divided in five groups: probabilities of correct detection, first- and second-order statistics, quantiles, latencies, optional metrics.

1) Probabilities of correct detection: These are simple probabilities computed on floors or zones, the latter in case zones of interest are defined. For some use cases, this may be all that is needed, for example in the common case where one only needs to know the position of the target at the room level.

- Floor detection probability — for multi-floor buildings
- Zone detection probability — if zones are defined

The standard *should* allow to evaluate the system only on the basis of these two metrics in such cases, without requiring any of the subsequent metrics to be presented.

2) First- and second-order statistics: These are based on the *error*, which is a 3D vector error, and *absolute error*, which is a 3D vector whose components are the absolute values of the *error* components:

- Mean of *error* — overall 3D bias
- Covariance matrix of *error* — a 3-by-3 matrix
- Covariance of *error* — the trace of the above matrix
- Mean of *absolute error* — a 3D vector
- Mean and standard deviation of *error* magnitude
- Mean and standard deviation of horizontal *error* magnitude
- Mean and standard deviation of vertical *error* magnitude
- Root mean square of magnitude of *error* and its components — xyz, x, y, z, xy

These statistics are mostly inappropriate as the result of a system-level, black-box approach to test and evaluation of a localisation system, for various reasons.

An overall 3D bias which is significantly different from zero points to an installation problem. Indeed, as mentioned in the standard, it is normally null, and if different it indicates an installation or coordinates measurement problem. It may also indicate a system's weakness. In any case, while extremely useful to the system developer, to the installer and to the tester, this information is useless or even misleading from the point of view of the final user. It *should not* be presented.

Similar reasoning can be applied to almost all of the above metrics. Apart from the root mean squares, the first- and second-order statistics above *should not* be presented.

Root mean square of error magnitude may be an exception, because in fact it gives a good grasp on the interesting characteristics of the error with a single number. However, it is less useful than the quantiles of error, which are discussed next. In our opinion, from a final user point of view, presenting different statistics makes system comparison harder and more subjective, rather than easier and more objective. This is why we suggest that root mean square is a good and simple tool for system developers and installers, but is bad for the final user and as a final metric.

3) Quantiles of error: Quantiles are the way to go. They answer to the most important of user's questions: how often does this system give wrong results? Or more precisely: what is the fraction of cases when this system gives an estimate with an error bigger than X?

- Median of error magnitude — SEP (spherical error probable)
- 95th percentile of error magnitude — SE95
- Median of horizontal error magnitude — CEP (circular error probable)
- 95th percentile of horizontal error magnitude — CE95
- Median of vertical error magnitude — VEP (vertical error probable)
- 95th percentile of vertical error magnitude — VE95

Once it is decided that quantiles are important information, one should decide what is the base measurements on which quantiles are measured, and which quantile values should be presented. In the following, when we speak about error we mean the *error* magnitude, that is, the distance from the position estimated by the system under test and the position of the relative ground-truth test point.

First, it should be clear that a generic 3D error, in xyz coordinates measured using any reference system, while often easier to measure, is not generally useful. Vertical error should always be treated specially: if the ELT is a person, what is needed is the floor, rather than the height from a reference point. Computing the floor may be very simple or very complex, but anyway this is the information that matters. When the ELT is an object, height may matter, but again floor is the most important information, so height should be relative to the floor. Any information regarding the vertical position *should* include the floor. If the ELT is a person or a wheel robot, estimating the floor is all that is needed. If the ELT is an object or a flying or crawling robot, the needed information is the floor and the height from the floor.

There is a more subtle but more fundamental problem with vertical positioning error, a problem which is a consequence of how error is defined and which involves horizontal error as well. While in an empty space like a warehouse, a large corridor or an auditorium, horizontal error is an important and sufficient information for the error, in a general place where walls are present, horizontal error may be deceiving, because a 3 m error in the same room is much more tolerable than a 3 m error across a wall. This argument acquires a greater strength when applied to vertical error: a 3 m horizontal error, with or without an intervening wall, is almost always much more tolerable than a 3 m vertical error.

The standard prescribes the computation of floor detection probability, as mentioned in section II-C1: this is necessary but not sufficient. There are some ways to deal with this problem, which are summarised in table II. The standard *should* prescribe the use of the second and preferably also the third solution mentioned therein.

The last missing piece is which quantiles should be reported. The standard mentions 0.5 and 0.95 (that is, the median and the 95th percentile), which in principle look like good choices based on common usage and usefulness. Better yet would be to use more quantile values, so to give a good approximation of the CDF (cumulative distribution function). The standard *should* require to use four quantiles: 0.5, 0.75, 0.9, 0.95, which makes it easy to compare two systems and to set minimum requirements for normative reasons.

Note that measuring high quantiles like 0.95 is tricky when samples are extracted from distributions with long tails, such as those expected for the error of a reasonably well-working system. In those cases, the confidence interval of the sample quantile can be wide enough that it has little significance unless the number of samples is quite high. The standard recommends that 50–100 test points are set up per floor, and that at least half of them is used for each scenario. This means that, if high quantiles are to be measured, the course should be walked through at least several times to meet the requirements mentioned in section II-C.

4) Latencies: Latencies are defined in two dimensions. First dimension is whether the measurement information consumer is the ELT or an external tracking authority. Second dimension is whether a pull or push method is used, depending on whether the measurement is initiated by the consumer or the system, respectively. Given that each dimension is binary, we have a total of four cases for latency definitions.

Latency is defined as the time elapsed from measurement initiation to measurement reception by the consumer.

The requested metrics are the mean and standard deviation of latency for each of the four cases.

Latency is of great importance for static object localisation, that is, cases when the ELT is not normally moving. Systems dedicated to this use case may adopt a long measurement procedure in order to filter out noise and trade promptness for accuracy, for example by giving few estimates per minute or even at longer time intervals.

This is not true of systems that are intended for general use, when the ELT is an object that should be tracked in real time and is frequently moved, more so when a person is the ELT. In this case, latency appears to be of little value as a system-level, black-box evaluation metric, because from the end-user point of view it is just one of the parameters that concur to building the error performance. In other words, it is important for the system developer, not the end user and consequently *should not* be a required test result for generic systems.

TABLE II
POSSIBLE DEFINITIONS OF SCALAR ERROR

Method	Ease of measurement	Usefulness	References
3D Euclidean distance	Very high: requires coordinates and at most coordinate conversion	Low: vertical errors have the same weight as horizontal errors, which means overestimating their importance if less than floor height and grossly underestimating it if greater than that	Used in the standard under review and almost universally in the literature
2D Euclidean distance with floor penalty	High: requires floor estimate which is available anyway, and little more complication than the above	Medium: floor errors are accounted for, but the penalty is somehow arbitrary; horizontal obstacles like walls are not accounted for	The standard error definition for the IPIN competitions 2014–2018 [8]
Real distance	Low: requires accurate maps and appropriate implementation	High: estimates the length of the distance to be walked from the estimate to the ground truth	This metric is described in detail in [18] and is a candidate for use in the IPIN 2019 competition

5) Optional metrics: There are some metrics that are described by the standard, but their usage is not mandated, either because they are confined to specific use cases or because they are difficult to be defined in a general way.

- Set-up time — for emergency applications
- Coverage — fraction of test area where a minimum performance is met
- Availability — fraction of test time where minimum performance is met
- Relative accuracy — statistics on distance between two ELTs
- Susceptibility — performance degradation due to interference and such
- Resilience — performance degradation due to catastrophic events

In general, we agree with the standard developers in that the above metrics can be useful, but should only be required in special cases to be individually defined.

III. COMPLIANCE OF IPIN AND MICROSOFT COMPETITIONS WITH THE STANDARD

Table III summarises some key requirements defined by the standard and how the IPIN and Microsoft competitions comply with them. Here is a definition of the key requirement listed in table III:

System-level and black-box approach The criterion used to measure how the system performs and to verify whether it meets the user requirements. The system is tested without any knowledge of its inner workings (black-box) and measures are only based on the expected localisation output (system-level).

Type of building The standard defines five different types of building. Specifically, the office building used by the IPIN and Microsoft competitions is defined as a brick-and-mortar building with at least three 2000 m^2 levels above ground and one level below ground.

Type of scenario The standard defines 14 different scenarios. The most similar to those used by the IPIN and Microsoft competition are the static one, where a 3 s stop is made at each test point, and the walking one, where a person walks without stopping.

Test point choice and number The standard mandates 50–100 test points per floor, uniformly and irregularly distributed across all building areas.

Choice of course Courses are predefined and unchangeable sequences of test points to be walked by the tester in the given order. A course should not pass twice on a given point, all points should be used by at least one course, a course should use at least half of the test points.

Mobility It is defined as: stationary, walking, running, backward walking, sidestepping, crawling. Walking means human walking at a speed of about 5 km/h, which is in fact too fast for a generic indoor environment.

Point error and distance evaluation Point error statistics are based on error measured at a series of points, as opposed to, for example, Fréchet distance. Distance evaluation refers to what is the basic datum used to compute the metrics. The IPIN competition has used horizontal Euclidean distance between ground truth and estimated position, with a floor penalty of 15 m per each estimated floor difference with respect to ground truth.

Metrics Metrics are extensively discussed in section II-C. Both IPIN and Microsoft competition use a single metric to compare systems.

IV. CONCLUSIONS

Evaluation of indoor localisation is one of the main challenges for researchers in this field, and some initiatives have been conducted to cope with it. As a matter of fact, the definition of benchmarking methodologies, common evaluation criteria, standardised methodologies useful to developers, testers, and end users is still an open challenge.

In this paper, we reviewed the test & evaluation procedures defined by the ISO/IEC 18305:2016 International Standard, with a special attention at commenting each of the indicated metrics. Specifically, after an overview of the overall standard,

TABLE III
ANALYSIS OF COMPLIANCE WITH ISO-IEC-18305 FOR THE IPIN AND MICROSOFT COMPETITIONS 2014–2017

Standard requirement	IPIN competition	Microsoft competition
System testing and black-box approach	The competition is compliant with the standard. Results from all the tracks proposed are intended as system-level testing outputs. The method adopted is a black-box approach: an independent tester executes the test by wearing the competing system and walking a given course.	The competition is compliant with the standard. Results from all the tracks proposed are intended as system-level testing outputs. The method adopted is a black-box approach: an independent tester asks the system developers to make a positioning estimate for each individual test point.
Type of building	The competition is run in a concrete office building, generally the same building where the IPIN conference is held. Only one of the different types of building mandated by the standard is used, and no below-ground levels are considered.	The competition is run in a concrete office building, generally the same building where the IPSN conference is held. Only one of the different types of building mandated by the standard is used, with only one or two floors.
Type of scenario	Not compliant: only one scenario, which is a combination of stationary and walking scenarios as defined by the standard, including short and long stop.	Not compliant: only one scenario, which is very similar to the static scenario defined by the standard, the only difference being that the competing systems stops on each test point for 2 s instead of the 3 s mandated by the standard.
Test points choice and number	Not compliant, because uses about 60 test points in environments of 2–4 floors, while the standard mandates 50–100 points per floor.	Not compliant, because uses about 20 test points in environments of 1–2 floors, while the standard mandates 50–100 points per floor.
Choice of course	Compliant: the course include all test points, tester walks once over each test point in a predefined order.	Not Compliant: testers do not have a predetermined course, and test points may be visited in different order for different systems.
Mobility	Compliant with the walking mobility model. Non compliant with the walking speed defined by the standard, which is excessively high.	Compliant with the static mobility model.
Point error and distance	Non compliant: while point error is used, the error is defined as the horizontal distance augmented with a floor penalty, which is not contemplated by the standard.	Compliant: error is the 3D Euclidean distance at each test point between ground truth and estimated position.
Metrics	Non compliant, as it adopts a single statistics (75th percentile of error magnitude) which is not among the many ones mandated by the standard.	Not compliant, as it adopts a single statistics (mean of *error* magnitude), which is only one of the many mandated by the standard.

we concentrated on the usefulness of some of the proposed metrics and raised doubts on others, suggesting possible modifications.

The paper also analyses the compliance with the standard of the two currently active initiatives that deal with the evaluation of indoor localisation systems, that is, the IPIN and Microsoft indoor localisation competitions.

This work is a basic contribution to the activities of the International Standards Committee of IPIN, specifically the Evaluation Methodologies subcommittee

REFERENCES

[1] R. Harle, "A survey of indoor inertial positioning systems for pedestrians." *IEEE Communications Surveys and Tutorials*, vol. 15, no. 3, pp. 1281–1293, 2013.

[2] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 6, pp. 1067–1080, 2007.

[3] R. Mautz and S. Tilch, "Survey of optical indoor positioning systems," in *Indoor Positioning and Indoor Navigation (IPIN), 2011 International Conference on*. IEEE, 2011, pp. 1–7.

[4] P. Barsocchi, S. Chessa, F. Furfari, and F. Potortì, "Evaluating ambient assisted living solutions: The localization competition," *IEEE Pervasive Computing*, vol. 12, no. 4, pp. 72–79, 2013.

[5] T. Van Haute, E. De Poorter, F. Lemic, V. Handziski, N. Wirström, T. Voigt, A. Wolisz, and I. Moerman, "Platform for benchmarking of rf-based indoor localization solutions," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 126–133, 2015.

[6] D. Lymberopoulos, J. Liu, X. Yang, R. R. Choudhury, S. Sen, and V. Handziski, "Microsoft indoor localization competition: Experiences and lessons learned," *GetMobile: Mobile Computing and Communications*, vol. 18, no. 4, pp. 24–31, 2015.

[7] F. Potortì, P. Barsocchi, M. Girolami, J. Torres-Sospedra, and R. Montoliu, "Evaluating indoor localization solutions in large environments through competitive benchmarking: The evaal-etri competition," in *Indoor Positioning and Indoor Navigation (IPIN), 2015 International Conference on*. IEEE, 2015, pp. 1–10.

[8] F. Potortì, S. Park, A. R. Jiménez Ruiz, P. Barsocchi, M. Girolami, A. Crivello, S. Y. Lee, J. H. Lim, J. Torres-Sospedra, F. Seco, *et al.*, "Comparing the performance of indoor localization systems through the evaal framework," *Sensors*, vol. 17, no. 10, p. 2327, 2017.

[9] ISO/IEC 18305:2016, "Information technology – Real time locating systems – Test and evaluation of localization and tracking systems," ISO/IEC JTC 1/SC 31, Standard ISO/IEC 18305:2016, 2016.

[10] M. Mendoza, G. Glorioso, P. Crombez, E. De Poorter, T. Van Haute, I. Moerman, T. Voigt, N. Wirström, M. Nilsson, F. Lemić, V. Handziski, A. Wolisz, and I. Piens, "D1.4: Dissemination & exploitation report," 2015.

[11] T. Van Haute, E. De Poorter, J. Rossey, I. Moerman, V. Handziski, A. Behboodi, F. Lemic, A. Wolisz, N. Wirström, T. Voigt, *et al.*, "The evarilos benchmarking handbook: Evaluation of rf-based indoor localization solutions," in *2nd International Workshop on Measurement-based Experimental Research, Methodology and Tools*, 2013.

[12] T. Van Haute, E. De Poorter, F. Lemic, V. Handziski, N. Wirström, A. Wolisz, and I. Moerman, "Platform for benchmarking rf-based indoor localization solutions: demonstration abstract," in *Proceedings of the 15th International Conference on Information Processing in Sensor Networks*. IEEE Press, 2016, p. 63.

[13] N. Patwari and J. Wilson, "Rf sensor networks for device-free localization: Measurements, models, and algorithms," *Proceedings of the IEEE*, vol. 98, no. 11, pp. 1961–1973, 2010.

[14] J. Wilson and N. Patwari, "See-through walls: Motion tracking using variance-based radio tomography networks," *IEEE Transactions on Mobile Computing*, vol. 10, no. 5, pp. 612–621, 2011.

[15] K. Buchin, M. Buchin, R. Van Leusden, W. Meulemans, and W. Mulzer, "Computing the fréchet distance with a retractable leash," *Discrete & Computational Geometry*, vol. 56, no. 2, pp. 315–336, 2016.

[16] A. Mathisen, S. K. Sørensen, A. Stisen, H. Blunck, and K. Grønbæk, "A comparative analysis of indoor wifi positioning at a large building complex," in *Indoor Positioning and Indoor Navigation (IPIN), 2016 International Conference on*. IEEE, 2016, pp. 1–8.

[17] L. Schauer, P. Marcus, and C. Linnhoff-Popien, "Towards feasible wi-fi based indoor tracking systems using probabilistic methods," in *Indoor Positioning and Indoor Navigation (IPIN), 2016 International Conference on*. IEEE, 2016, pp. 1–8.

[18] G. M. Mendoza-Silva, J. Torres-Sospedra, and J. Huerta, "A more realistic error distance calculation for indoor positioning systems accuracy evaluation," in *Indoor Positioning and Indoor Navigation (IPIN), 2017 International Conference on*. IEEE, 2017, pp. 1–8.