ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/282848677

A Survey of Experimental Evaluation in Indoor Localization Research

Conference Paper · October 2015

DOI: 10.1109/IPIN.2015.7346749

CITATIONS
64

READS
539

4 authors:

Stephan Adler
Freie Universität Berlin
21 PUBLICATIONS **572** CITATIONS
SEE PROFILE

Simon Schmitt
Freie Universität Berlin
14 PUBLICATIONS **518** CITATIONS
SEE PROFILE

Katinka Wolter
Freie Universität Berlin
162 PUBLICATIONS **1,697** CITATIONS
SEE PROFILE

Marcel Kyas
Reykjavik University
74 PUBLICATIONS **760** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project AVS Extrem View project

Project Model-based quantitative analysis of cryptocurrency systems View project

All content following this page was uploaded by Simon Schmitt on 14 October 2015.

The user has requested enhancement of the downloaded file.

2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN), 13-16 October 2015, Banff, Alberta, Canada

A Survey of Experimental Evaluation in Indoor Localization Research

A Look Back on IPIN conferences 2010, 2011, 2012, 2013, and 2014

Stephan Adler, Simon Schmitt, and Katinka Wolter
Freie Universität Berlin
AG Computer Systems & Telematics
Berlin, Germany
{stephan.adler, simon.schmitt, katinka.wolter}@fu-berlin.de

Marcel Kyas
Reykjavík University
School of Computer Science
Reykjavík, Iceland
marcel@ru.is

***Abstract*—During the last decade, research in indoor localization and navigation has focused on techniques, protocols, and algorithms. The first International Conference on Indoor Positioning and Indoor Navigation (IPIN) was held in 2010. Since then, this annual conference showed the progress of research and technology. The variations of evaluation methods are significant in this field: they range from none, to extensive simulations, and real-world experiments under non-lab conditions. We look at the articles published in the proceedings of IPIN by IEEE Xplore from 2010 to 2014, and analyze the development of evaluation methods. We categorized 183 randomly selected papers, in respect to five different aspects. Namely: (1) the underlying system/technology in use, (2) the evaluation method for the proposed technique, (3) the method of ground truth data gathering, (4) the applied metrics, and (5) whether the authors establish a baseline for their work.**

I. Introduction

We survey the current state of research in indoor localization by exploring the kind of systems that are commonly used and how they are evaluated. We focus on comparability between different evaluations and show important factors that have to be considered when evaluating a system for use in indoor localization.

The motivation for this survey is rooted in our attendance to the fifth international conference on Indoor Positioning and Indoor Navigation (IPIN), where we wondered whether our impressions about a visible improvement in the methodology of evaluation of indoor localization systems over the last years could be confirmed in a systematic analysis of prior publications.

We analyzed the evaluation methods used in most contributed publications of the first five installments of this conference [1]–[5]. We examine if there is a shift in technologies in use, how these technologies are evaluated, and if the publications are based on scientifically and empirically sound methods. Finally, we suggest improvements to the methods of evaluating indoor positioning and indoor navigation methods.

This paper may also be viewed as a contribution to the debate on reproducibility in computer science research. Reproducibility is a corner stone of science, especially physics. A theory gets accepted by the community, if it has been confirmed by independent researches, e.g. by reproducing the results of an experiment.

978-1-4673-8402-5/15/$31.00 ©2015 IEEE

A. Related Work

Tichy et al. [6] survey papers in computer science from 1993 to determine whether computer scientists support their results with experimental evaluation. Their result shows that authors did not support their results well: 40 % of their sample did no evaluation at all, even though it would be required. Tichy et al. further motivate more experimentation in all computer science disciplines [7].

This result of Tichy et al. was later corroborated by Wainer et al. [8], who looked at publications from 2005. They indicate that the lack of empirical or experimental components did not change during 15 years of research.

More recent studies also stress the importance of experimentation and lack thereof in various sub disciplines of computer science, e.g. Andujar et al. [9] and Tedre and Moisseinen [10].

We do not contribute to the debate on reproducibility in computer science research, as initiated with WaveLab by Buckheit and Donoho [11] and later by Collberg, Proebsting, and Warren [12].

Reproducibility is the ability of an experiment or study to be duplicated, either by the same researcher or by someone else working independently. Reproducibility is one of the main principles of the scientific method. A theory should only be accepted by the community, if it has been confirmed by independent researches, e.g. by reproducing the results of an experiment [13].

In this paper, we do not measure the reproducibility of IPIN results. First, no experiment was replicated by us for this study. Neither did we ask for code or additional information. Thus, we cannot make any statement about reproducibility.

As experimental data is analyzed statistically, the suggestions of Gentleman and Lang [14] should apply. The paper should at least reference the data and the methods that were used to generate all statistical data, like the numbers and figures. For simulations, that includes all code to run the simulator, which would make this research fully reproducible.

Physical experiments are harder to reproduce. Using the method of Collberg, Proebsting, and Warren [12], we would have to exclude almost all papers, because they need special hardware to reproduce. Still, such experiments should be explained in as much detail that someone with access to this hardware can reproduce the experiment. We did not try to repeat any experiment.

We can still make a statement about repeatability of experiments. We counted the number of papers that stated a number of repetitions (see Sect. II-D). The purpose of repeating experiments is to increase confidence in the observations and improve the precision of measurements, which is proportional to the reciprocal square root of the number of repetitions.

Repeatability should satisfy the following conditions: the experiment should be done using the same experimental tools, by the same observer using the same measuring instrument, used under the same conditions at the same location. The experiment should be repeated over a short period of time and the experiment should have the same objective [15]–[17].

We study the proposed metrics and best practices described by Raj [17] for our approach to classify the experimental setups and to assess applied methods found in the publications.

We focus on statistical comparability. We define *comparability* of results as: can we estimate a probability that we have (or have not) reproduced the same experimental result. As the published evaluations are obtained using statistical methods (usually identified by supplying an average, a cumulative distribution function (c.d.f.) or some other sample of measurements), the published result should be usable for tests of statistical comparison. The same notion of comparability is used to judge whether one method is effectively superior to any other method.

II. Research Method

The proceedings of the first five IPIN conferences published by IEEE Xplore contain 626 papers (2010: 129, 2011: 95, 2012: 157, 2013: 144, and 2014: 101). We randomized the order of the publication list and drew at least 30 % of the publications published by IEEE Xplore per conference to ensure impartiality.

To avoid an emotional discussion about our ratings we avoided to cite single publications as positive or negative examples in this paper. Nevertheless, our results can easily be verified or extended by looking at the conference proceedings directly.

We divided the corpus of publications into five categories. First, we looked at what type of localization technique was used by the authors (system category). Second, we distinguished the method of evaluation and categorized by the chosen scenario (evaluation category). Third, we distinguished methods for ground truth data gathering (reference category). Fourth, we distinguished methods by the way metrics are calculated and presented. And fifth, we distinguished methods by the choice of a baseline to compare against.

A. System Categories

The system categories describe the underlying technology of the system under test (SUT). If a publication describes a system which uses multimodal sensoring and therefore can be categorized into multiple categories, we marked it as *multiple*. We also count it as a member of all distinct categories that fit the SUT. In most cases these are inertial systems combined with Received Signal Strength (RSS) and/or time-of-flight (TOF) systems for the purpose of recalibration of the Inertial Measurement Unit (IMU). Other multimodal approaches describe systems which carry a global navigation satellite system (GNSS) sensor for outdoor localization and another type of (sub-) system as a fallback solution for indoor purposes. If a sensor in use is only for ground truth data gathering, it does not belong to the SUT and is therefore not counted in the system category.

There are publications that do not fit into any of our categories. These are mostly publications that do not describe systems or algorithms in indoor scenarios directly (e.g. surveys). We marked them as *unrelated*.

We define the following system categories:

Inertial Systems that use pedestrian dead reckoning (PDR) or other IMU tracking techniques.

Map Matching Systems that use any kind of a priori generated or recorded maps, including patterns of environmental characteristics for the position estimation, such as received signal strength indicator (RSSI).

RSS Systems that use RSS for range estimation.

TOF The TOF category contains all approaches which use some form of TOF estimation to calculate the distance to another network member.

Sound Systems that are based on sound, for example ultrasound beacons or other sound sources with known position to estimate their distance to known anchor beacons.

Other Systems that use different spatial depended environmental properties than described above, e.g. light, magnetic fields, visual object recognition.

Multiple Systems using multimodal sensoring.

Unrelated Publications which neither describe a localization system nor a localization algorithm but other localization related topics.

B. Evaluation Categories

The evaluation categories describe the method used for evaluation for the related papers. If multiple evaluation methods are used, we marked the publication as *multiple*. However, we also count it as a member of all distinct evaluation categories. In most cases simulations of the method is combined with a physical evaluation of a system.

Since the error model of inertial based systems is very different from the error model of range based systems, the evaluation techniques in use are also different for both cases. For that reason, we examined these systems separately.

We define the following evaluation categories:

No Evaluation Publications that propose localization algorithms or systems but are neither evaluated by simulation nor by experiment.

Simple Simulation Simulations which use simple signal propagation models that do not recognize obstacle-dependent effects such as non-line-of-sight (NLOS) or assume Gaussian distributed errors of the measurement system.

Complex Simulation Simulations which take effects introduced by obstacles into account, adjust the error model accordingly, or use experimentally gathered data as a source for an error distribution.

Discrete Point Experiments with selected discrete points for evaluation. The reference points are either discrete points in a 1-dimensional structure (path or straight line) or non-systematically scattered in an evaluation area.

Grid-like Grid experiments are simple setups usually in an empty room without any obstacles. The reference points form a grid-like shape.

Office Walk An office walk experiment uses a specific path as reference. The target is moved along this path while estimating its location or collecting measurements. Afterwards the (mostly predefined) ground truth path can be compared to its estimated counterpart.

Outdoor Experiments that are only performed in outdoor environments.

Real World Experiments and even simulations, where the SUT is exposed to different building structures under non-lab conditions, e.g. in populated buildings or over several days.

Multiple Publications that use more than one of the evaluation methods above.

Unrelated Publications that do not cover localization itself and are thereby not subject to an evaluation process.

C. Reference Categories

The following categories describe the kind of reference system or method in use in the selected publications. We marked papers that do not disclose their ground truth gathering method as *undefined*.

We define the following reference categories:

Landmarks Single reference points with varying degree of accuracy.

Path A path, e.g. fixed points, on the floor or landmarks in the vicinity. The target is directed along these points while the time may be observed.

Optical An optical system that tracks the target using cameras, e.g. mounted on the ceiling.

Infrastructure A measurement system that often takes considerable effort to install, e.g. robot localization using predefined maps.

GNSS Reference data collected by global navigation satellite systems outside of buildings, e.g. GPS.

Undefined The use of reference data or the method to gather it is not mentioned or very poorly described.

D. Metric Categories

Every evaluation of a systems performance should result in a quantity. That quantity should have a unit. It should also be clear how this quantity is measured and what parameters influence the measurements. Finally, measurements ought to be reported together with the number of measurements, the accuracy and the precision [17]. The terminology follows ISO 5275-1: 1994 [16].

No Metric No metric is stated at all.

Trueness The closeness between the average value of a number of measurements and an accepted reference value. It is usually expressed as the bias, but sometimes also by the mean average error (MAE) of the measurements.

Precision The closeness of agreement between independent measurements. It is usually expressed as the variance, standard deviation, or quantiles.

Accuracy A measure of the systems combined trueness and precision is stated, e.g. the mean squared error (MSE) or root-mean-square error (RMSE).

Distribution The data distribution is provided, either as a histogram, a c.d.f. or similar.

Sample Size The number of experiments that are considered in above values is stated.

To motivate the choice of categories, we point out that usually, the exact performance of a system is not known. Experiments and measurements collect data points that allow us to estimate the performance of a system. The *accuracy* of a system is the closeness between the experimental results and the reference data. It is composed of the two components *trueness* and *precision*. Accuracy, trueness and precision can be related by

$$\text{MSE} = \text{MAE}^2 + \text{Variance} \tag{1}$$

An increase of sample size N, e.g., by repeated measurements, influences the value of the variance and thus the accuracy of an estimate. Assuming the statistical independence of the measurements, the sample variance of the measurements depends on the number of samples, it decreases proportional to $N^{-1/2}$. This means that our estimate of the MAE and the variance improves with increased sample size.

An alternative presentation of the data is the c.d.f., where the frequencies of the measurement errors are displayed in a graphical manner. We include histograms of errors and display of the raw data in this category, because the c.d.f. can be derived from the histogram.

We only include papers that present their experimental results in a graphical fashion if the data is clearly legible from the graphics. Papers that provide a depiction of a track only are not counted, because the time of each value is missing, among others.

We also consider two derived categories in our analysis:

Statistics All measures to compare two experiments or simulations are provided in a way that allows a comparison. This means that a paper is counted in at least two of

TABLE I
SYSTEM CATEGORIES SUMMARY IN [COUNT (%)]

	2010		2011		2012		2013		2014		Avg.	
Inertial	7	(21)	7	(35)	8	(25)	10	(29)	7	(41)	39	(28)
only Inertial	2	(6)	4	(20)	4	(13)	4	(11)	3	(18)	17	(12)
Map Matching	4	(12)	6	(30)	9	(28)	6	(17)	5	(29)	30	(22)
only Map Matching	3	(9)	5	(25)	7	(22)	4	(11)	3	(18)	22	(16)
RSS	7	(21)	3	(15)	4	(13)	10	(29)	3	(18)	27	(20)
only RSS	3	(9)	3	(15)	3	(9)	4	(11)	1	(6)	14	(10)
TOF	16	(47)	2	(10)	4	(13)	10	(29)	3	(18)	35	(25)
only TOF	10	(29)	1	(5)	1	(3)	7	(20)	0	(0)	19	(14)
Sound	5	(15)	0	(0)	5	(16)	2	(6)	1	(6)	13	(9)
only Sound	1	(3)	0	(0)	5	(16)	2	(6)	1	(6)	9	(7)
Other	8	(24)	5	(25)	8	(25)	7	(20)	4	(24)	32	(23)
only Other	6	(18)	4	(20)	6	(19)	5	(14)	3	(18)	24	(17)
Multiple	9	(26)	3	(15)	6	(19)	9	(26)	6	(35)	33	(24)
Related Paper	34	(100)	20	(100)	32	(100)	35	(100)	17	(100)	138	(100)

TABLE II
EVALUATION CATEGORIES (ALL SYSTEMS) IN [COUNT (%)]

	2010		2011		2012		2013		2014		Avg.	
No Evaluation	3	(9)	1	(5)	1	(3)	1	(3)	0	(0)	6	(4)
Simple Simulation	10	(29)	4	(20)	1	(3)	4	(11)	4	(24)	23	(17)
only Simple Simulation	7	(21)	3	(15)	1	(3)	3	(9)	3	(18)	17	(12)
Complex Simulation	1	(3)	2	(10)	2	(6)	7	(20)	2	(12)	14	(10)
only Complex Simulation	1	(3)	1	(5)	1	(3)	5	(14)	2	(12)	10	(7)
Discrete Point	5	(15)	3	(15)	9	(28)	9	(26)	3	(18)	29	(21)
only Discrete Point	4	(12)	2	(10)	8	(25)	7	(20)	3	(18)	24	(17)
Grid-like	5	(15)	2	(10)	5	(16)	4	(11)	1	(6)	17	(12)
only Grid-like	4	(12)	2	(10)	5	(16)	3	(9)	1	(6)	15	(11)
Office Walk	9	(26)	8	(40)	11	(34)	10	(29)	6	(35)	44	(32)
only Office Walk	8	(24)	7	(35)	9	(28)	9	(26)	5	(29)	38	(28)
Outdoor	1	(3)	1	(5)	5	(16)	2	(6)	0	(0)	9	(7)
only Outdoor	0	(0)	1	(5)	3	(9)	1	(3)	0	(0)	5	(4)
Real World	4	(12)	1	(5)	2	(6)	2	(6)	1	(6)	10	(7)
only Real World	4	(12)	1	(5)	1	(3)	2	(6)	1	(6)	9	(7)
Multiple	3	(9)	2	(10)	3	(9)	4	(11)	1	(6)	13	(9)

average, variance, and squared error. With two quantities given, Eq. 1 is fully determined.

Complete All measures to compare two experiments or simulation are provided in a way that allows a comparison. This means that a paper is counted in at least two of average, variance, and squared error, and it includes the sample size.

Note that the required variance mentioned above is important to evaluate whether the reproduced measurements are close enough to the published measurements and the experiment can be considered reproduced. The RMSE cannot be used for this purpose [18].

E. Baseline Category

The basic question of this category is: do the authors argue an improvement with respect to previously published work. In order to do this, the experiment should be evaluated by a previous method, thus establishing a baseline. Then, the proposed method is evaluated using the same data. While this method is established scientific practice, we consider the age of the youngest alternative the method is compared against. Ideally, this age is small and references one year old publications. This way, actual progress becomes visible.

Papers are not counted in this category, if some well known method like linear least squares (LLS) is used as the youngest baseline. First, LLS is a well-known algorithm that is easily improved on. Second, if the only merit is to improve on LLS, an improvement is not exactly established.

Additionally, we wish that the improvement is soundly demonstrated, e.g. by estimating the probability that the new method is really giving results different from the baseline.

III. RESULTS

We examined 183 papers and marked 45 papers as *unrelated*. The following tables show the absolute number of publications that were assigned to a category and the relative number in percent in parenthesis. The following percentage numbers in this paper have always the related papers we found as basis, not including *unrelated* papers if not stated otherwise.

A. System Category

Table I shows the localization techniques in use (system category). Between 68 and 85 % of all publications per year fit in our categorization system. The remaining papers are categorized as *unrelated*. All categories except *sound* (only 9 %) share nearly equal portions of all contributions (20 to 28 %).

The range based solutions using TOF and/or RSS add up to 41 % over the years. Publications using range based solutions are dropping from 62 % in 2010 to 25 % in 2011 and 2012, and come back to 49 % in 2013 and 35 % in 2014. This is accompanied by a rise in the *unrelated* and *map matching* categories in 2012 and 2013. Although we expected to see a trend towards a favored technolgy before surveying the publications the data fails to show such a development.

We observe that 76 % of the publications focus on one type of sensing, while the remaining papers use multimodal approaches. The majority of the multimodal systems we examined are IMU (67 % of all *multiple* papers) with recalibration by another system. They are most often combined with *RSS*, *TOF*, *map matching*, and *other* systems, in descending order. We did not find any IMU combined with *sound*. The other *multiple* systems combine other range based and map matching approaches to solve the indoor localization problem.

B. Evaluation Category

Table II shows the distribution of the evaluation categories. 96 % of all papers perform some kind of evaluation whereas 4 % perform no evaluation at all. The three most often used evaluation methods are *office walk* (32 %), all kinds of *simulation* approaches (27 %), and *discrete point* (21 %) followed by evaluation using *grid-like* experiments (12 %). We did not expect that only 7 % of all papers meet our *real world* requirements. Looking at the *simulation* categories, we see that *simple simulation* declined from 29 % in 2010 to 3 % in 2012, and rose again to 24 % in 2014, while *complex simulation* was on the highest in 2013 with 20 %. 20 % of all papers perform only simulation, whereas 68 % perform only physical experiments. 7 % perform simulations and physical experiments.

TABLE III
EVALUATION CATEGORIES (INERTIAL SYSTEMS) IN [COUNT (%)]

	2010		2011		2012		2013		2014		Avg.	
No Evaluation	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)
Simple Simulation	1	(14)	1	(14)	0	(0)	0	(0)	0	(0)	2	(5)
only Simple Simulation	0	(0)	1	(14)	0	(0)	0	(0)	0	(0)	1	(3)
Complex Simulation	0	(0)	1	(14)	1	(13)	2	(20)	1	(14)	5	(13)
only Complex Simulation	0	(0)	1	(14)	0	(0)	1	(10)	1	(14)	3	(8)
Discrete Point	0	(0)	0	(0)	1	(13)	0	(0)	0	(0)	1	(3)
only Discrete Point	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)
Grid-like	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)
only Grid-like	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)
Office Walk	4	(57)	4	(57)	6	(75)	7	(70)	4	(57)	25	(64)
only Office Walk	3	(43)	4	(57)	5	(63)	6	(60)	4	(57)	22	(56)
Outdoor	1	(14)	1	(14)	2	(25)	1	(10)	0	(0)	5	(13)
only Outdoor	0	(0)	1	(14)	1	(13)	1	(10)	0	(0)	3	(8)
Real World	3	(43)	0	(0)	1	(13)	1	(10)	1	(14)	6	(15)
only Real World	3	(43)	0	(0)	0	(0)	1	(10)	1	(14)	5	(13)
Multiple	1	(14)	0	(0)	2	(25)	1	(10)	0	(0)	4	(10)

TABLE IV
EVALUATION CATEGORIES (NON-INERTIAL) IN [COUNT (%)]

	2010		2011		2012		2013		2014		Avg.	
No Evaluation	3	(11)	1	(8)	1	(4)	1	(4)	0	(0)	6	(6)
Simple Simulation	9	(33)	3	(23)	1	(4)	4	(16)	4	(40)	21	(21)
only Simple Simulation	7	(26)	2	(15)	1	(4)	3	(12)	3	(30)	16	(16)
Complex Simulation	1	(4)	1	(8)	1	(4)	5	(20)	1	(10)	9	(9)
only Complex Simulation	1	(4)	0	(0)	1	(4)	4	(16)	1	(10)	7	(7)
Discrete Point	5	(19)	3	(23)	8	(33)	9	(36)	3	(30)	28	(28)
only Discrete Point	4	(15)	2	(15)	8	(33)	7	(28)	3	(30)	24	(24)
Grid-like	5	(19)	2	(15)	5	(21)	4	(16)	1	(10)	17	(17)
only Grid-like	4	(15)	2	(15)	5	(21)	3	(12)	1	(10)	15	(15)
Office Walk	5	(19)	4	(31)	5	(21)	3	(12)	2	(20)	19	(19)
only Office Walk	5	(19)	3	(23)	4	(17)	3	(12)	1	(10)	16	(16)
Outdoor	0	(0)	0	(0)	3	(13)	1	(4)	0	(0)	4	(4)
only Outdoor	0	(0)	0	(0)	2	(8)	0	(0)	0	(0)	2	(2)
Real World	1	(4)	1	(8)	1	(4)	1	(4)	0	(0)	4	(4)
only Real World	1	(4)	1	(8)	1	(4)	1	(4)	0	(0)	4	(4)
Multiple	2	(7)	2	(15)	1	(4)	3	(12)	1	(10)	9	(9)

TABLE V
REFERENCE CATEGORIES SUMMARY IN [COUNT (%)]

	2010		2011		2012		2013		2014		Avg.	
Landmarks	10	(29)	4	(20)	15	(47)	14	(40)	5	(29)	48	(35)
Path	4	(12)	6	(30)	7	(22)	6	(17)	2	(12)	25	(18)
Optical	2	(6)	1	(5)	1	(3)	2	(6)	2	(12)	8	(6)
Infrastructure	2	(6)	1	(5)	1	(3)	2	(6)	0	(0)	6	(4)
GNSS	0	(0)	0	(0)	4	(13)	1	(3)	0	(0)	5	(4)
Undefined	8	(24)	2	(10)	3	(9)	1	(3)	4	(24)	18	(13)

Since inertial based systems are often evaluated using *office walks*, we also split the results into inertial based and non-inertial based systems. Table III shows the percentages of all inertial based systems. The majority of all evaluations used predefined paths, i.e. an *office walk*, in office environments (64 %). It is also interesting, that more papers fit into the *real world* (15 %) and *outdoor* category (13 %). In 2010, 43 % of the papers we looked at were *real world* experiments, which is remarkably high when looking at other years (0 to 14 %). When we look at the non-inertial based systems in Tab. IV, we see that *discrete point* evaluation is the dominant method (28 %), followed by *simple simulation* (21 %), *office walk* (19 %), and *grid like* (17 %) experiments.

C. Reference Category

Table V lists the percentages of reference methods in use. We did not find any reference method that did not fit into the categories. Most of all papers describe their setup very well, but tend to neglect the information on how the ground truth information was gathered. Therefore it was very hard for us to categorize them. If it was implicitly clear what method was used (mostly *landmarks*), we did not categorize it as *undefined*. However, there are publications, which do not describe their method at all. Most of the time, images depict a ground truth trajectory as a straight line behind measurement information, with no information on how the gathered data was fitted to the line at all.

TABLE VI
METRICS CATEGORY SUMAMRY IN [COUNT (%)]

	2010	2011	2012	2013	2014	Avg.
No metric	17 (50)	3 (15)	8 (25)	7 (20)	6 (35)	41 (30)
Trueness	15 (44)	15 (75)	17 (53)	20 (57)	8 (47)	75 (54)
only trueness	10 (29)	7 (35)	6 (19)	7 (20)	4 (24)	34 (25)
Precision	5 (15)	8 (40)	12 (38)	14 (40)	4 (24)	43 (31)
only precision	0 (0)	0 (0)	1 (3)	1 (3)	0 (0)	2 (1)
Accuracy	7 (21)	9 (45)	14 (44)	16 (46)	7 (41)	53 (38)
only accuracy	2 (6)	1 (5)	3 (9)	3 (9)	3 (18)	12 (9)
Distribution	0 (0)	2 (10)	4 (13)	13 (37)	3 (18)	22 (16)
only distribution	0 (0)	1 (5)	3 (9)	4 (11)	0 (0)	8 (6)
Sample size	5 (15)	6 (30)	15 (47)	16 (46)	6 (35)	48 (35)
only sample size	1 (3)	0 (0)	1 (3)	0 (0)	1 (6)	3 (2)
Trueness&Precision	5 (15)	8 (40)	11 (34)	13 (37)	4 (24)	41 (30)
Complete	2 (6)	4 (20)	9 (28)	10 (29)	3 (18)	28 (20)

Most often the authors choose *landmarks* as their resource of ground truth information (27 %). We found papers that explicitly state the accuracy of their manual measurements from millimeters to centimeters, as well as papers in which only a rough description of the true position is given (e.g. by a dot in a picture). The difference to *path* related reference methods (15 %) is only that a consecutive series of points in space is considered as ground truth. Unfortunately, we saw no paper that explicitly stated that they evaluate their approach by looking not only at spatial distance to their ground truth but also distance in time.

To our surprise, we did find little about reference methods like *optical*, *infrastructure*, and *GNSS* (3 to 4 % each). Using *GNSS* as ground truth information only makes sense e.g. when evaluating inertial based systems outdoor on a long track, which was not often represented in our findings (6 %). But the other two categories are able to evaluate a wide range of systems in a repeatable manner. The most likely reason are the perceived setup costs of such solutions. However, there are also cheap ways of doing an evaluation with e.g. mobile robots, as we showed previously [19]. We did not find such a solution in our sample.

D. Metric Category

Table VI summarizes the distribution of papers to the metric categories. In total, about 73 % of the publications provide a kind of metric or data of their results. Of those, 70 % present a quantifiable metric, and 3 papers present a metric in a non-numeric way. The other papers rely on anecdotal evidence, in that they demonstrate the feasibility of their approach and don't measure the performance.

A measure of trueness is reported in 54 % of the publications, a measure of accuracy is reported in 38 % of the publications. In total, 9 % of the papers only report a measure of accuracy. A c.d.f., histogram or the raw data is provided in 16 % of all papers. Half of those, 8 % only report a c.d.f. without a numerical summary. Two papers (2 %) only report variances.

Accuracy is the combination of trueness and precision. A combined measure is less useful for comparison, because the systematic error, measured by a lack of trueness, is conflated with the lack of precision. Thus, we indicate the number of papers that report both an estimation of trueness and estimation of precision. These are 30 % (41 papers).

Finally, as the reported data is an estimate that depends on the sample size considered in the estimation, we also calculated the number of papers that report a measure of trueness, a measure of precision, and the sample size used to estimate those values. These are 20 %. If one is to compare his results to a published result, the sample size has to be considered in the estimation of the similarity/difference, because the sample size is often an input to the calculation.

We note 3 papers that report a sample size but do not report any other metric from our selection. The papers present raw data in a figure. By visually inspecting the results, the effectiveness of the method is conveyed. However, the raw data is not published in the paper and cannot be compared.

1) Metrics in systems: Can we identify a reason for the choice of reporting of metrics in IPIN? Can we explain why only one in five papers report metrics that allow a meaningful comparison? Can we explain why almost one in three publications does not report any metrics? To answer these questions, we look at the distribution of reporting per system.

After looking at the summarized data, we look at preferences of methods in different fields of research. Since we see that the method of presenting metrics varies, and that the reporting is sometimes incomplete according to certain standards, we wonder if we can identify any preference that explains this observation.

Note that the data reported here has a large error margin, as the data refers to a smaller subset of papers only. But while the numbers may not be accurate, they still indicate the presence of problems. Consequently, we do not provide a formal model or look for correlations. Variations in a single paper result in rather large changes of the percentages reported in the tables and our data considers only about a third of each year. For this reason, we exclude papers on sound based systems from the evaluation.

Table VII summarizes the presentation of metrics in papers presenting inertial systems. Because we consider only 39 papers on inertial systems, the margin of error is bounded from above by 0.15 for a confidence level of 95 %.

Except for 2010 and 2014, the use of metrics is estimated to be similar to the totals in Table VI. The high percentage difference in 2010 and 2014 is a single paper difference from the expected value.

TABLE VII
METRICS FOR INERTIAL SYSTEMS IN [COUNT (%)]

	2010	2011	2012	2013	2014	Avg.
No metric	3 (43)	1 (14)	2 (25)	2 (20)	3 (43)	11 (28)
Trueness	4 (57)	6 (86)	4 (50)	5 (50)	3 (43)	22 (56)
only trueness	3 (43)	6 (86)	0 (0)	3 (30)	1 (14)	13 (33)
Precision	1 (14)	0 (0)	5 (63)	3 (30)	2 (29)	11 (28)
only precision	0 (0)	0 (0)	1 (13)	1 (10)	0 (0)	2 (5)
Accuracy	1 (14)	0 (0)	5 (63)	2 (20)	3 (43)	11 (28)
only accuracy	0 (0)	0 (0)	1 (13)	0 (0)	1 (14)	2 (5)
Distribution	0 (0)	0 (0)	0 (0)	4 (40)	2 (29)	6 (15)
only distribution	0 (0)	0 (0)	0 (0)	2 (20)	0 (0)	2 (5)
Sample size	1 (14)	2 (29)	5 (63)	4 (40)	3 (43)	15 (38)
only sample size	0 (0)	0 (0)	1 (13)	0 (0)	1 (14)	2 (5)
Trueness&Precision	1 (14)	0 (0)	4 (50)	2 (20)	2 (29)	9 (23)
Complete	0 (0)	0 (0)	3 (38)	2 (20)	1 (14)	6 (15)

TABLE VIII
METRICS FOR MAP MATCHING SYSTEMS IN [COUNT (%)]

	2010	2011	2012	2013	2014	Avg.
No metric	1 (25)	1 (17)	0 (0)	2 (33)	0 (0)	4 (13)
Trueness	3 (75)	5 (83)	7 (78)	2 (33)	4 (80)	21 (70)
only trueness	1 (25)	2 (33)	3 (33)	0 (0)	3 (60)	9 (30)
Precision	2 (50)	3 (50)	4 (44)	2 (33)	1 (20)	12 (40)
only precision	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Accuracy	2 (50)	3 (50)	4 (44)	2 (33)	2 (40)	13 (43)
only accuracy	0 (0)	0 (0)	0 (0)	0 (0)	1 (20)	1 (3)
Distribution	0 (0)	0 (0)	2 (22)	3 (50)	1 (20)	6 (20)
only distribution	0 (0)	0 (0)	2 (22)	2 (33)	0 (0)	4 (13)
Sample size	0 (0)	3 (50)	5 (56)	1 (17)	1 (20)	10 (33)
only sample size	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Trueness&Precision	2 (50)	3 (50)	4 (44)	2 (33)	1 (20)	12 (40)
Complete	0 (0)	3 (50)	3 (33)	1 (17)	1 (20)	8 (27)

Our first observation is that it is probably the case that papers on inertial system are as like to report metrics as the statistical population. It may be less likely that papers on inertial systems report trueness and precision or even comparable results. In our sample, more authors state the sample size than the precision.

We believe that these metrics are not reported, because walking people get tired and experiments are consequently hard to repeat. Thus, the precision is hard to estimate. The papers that report a sample size report a rather low number. There, it actually makes sense not to report those numbers, as their significance is weak.

Despite the large error margins we have to expect, we are able to explain the numbers. Still, we think that reporting the number of trials should be pervasive, because this number serves as the justification to the lack of reports on precision.

Table VIII summarizes the presentation of metrics in papers presenting map matching systems. There are 30 papers on map matching in our sample.

Most papers on map matching systems provide metrics on the performance. Also, 40 % of all publications report trueness and precision separately, thus we can properly evaluate the accuracy of the method. The number of samples in the evaluation is reported in only 33 % of the papers, which makes it hard to evaluate the quality of those estimates. Overall, performance evaluation of map matching algorithms is well done in many cases.

Table IX summarizes the presentation of metrics in papers presenting systems based on RSS values. There are 27 papers in our sample. The data is hard to interpret: There are 3 papers

TABLE IX
METRICS FOR RSS SYSTEMS IN [COUNT (%)]

	2010	2011	2012	2013	2014	Avg.
No metric	3 (43)	0 (0)	2 (50)	1 (10)	1 (33)	7 (26)
Trueness	4 (57)	1 (33)	2 (50)	5 (50)	0 (0)	12 (44)
only trueness	3 (43)	0 (0)	1 (25)	3 (30)	0 (0)	7 (26)
Precision	1 (14)	1 (33)	1 (25)	3 (30)	0 (0)	6 (22)
only precision	0 (0)	0 (0)	0 (0)	1 (10)	0 (0)	1 (4)
Accuracy	1 (14)	2 (67)	1 (25)	4 (40)	2 (67)	10 (37)
only accuracy	0 (0)	1 (33)	0 (0)	2 (20)	2 (67)	5 (19)
Distribution	0 (0)	1 (33)	0 (0)	5 (50)	0 (0)	6 (22)
only distribution	0 (0)	1 (33)	0 (0)	1 (10)	0 (0)	2 (7)
Sample size	0 (0)	0 (0)	1 (25)	4 (40)	1 (33)	6 (22)
only sample size	0 (0)	0 (0)	0 (0)	0 (0)	1 (33)	1 (4)
Trueness&Precision	1 (14)	1 (33)	1 (25)	2 (20)	0 (0)	5 (19)
Complete	0 (0)	0 (0)	1 (25)	2 (20)	0 (0)	3 (11)

TABLE X
METRICS FOR TOF SYSTEMS IN [COUNT (%)]

	2010	2011	2012	2013	2014	Avg.
No metric	8 (50)	1 (50)	2 (50)	1 (10)	2 (67)	14 (40)
Trueness	7 (44)	1 (50)	1 (25)	7 (70)	1 (33)	17 (49)
only trueness	5 (31)	0 (0)	0 (0)	0 (0)	0 (0)	5 (14)
Precision	2 (13)	1 (50)	1 (25)	7 (70)	1 (33)	12 (34)
only precision	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Accuracy	3 (19)	1 (50)	2 (50)	8 (80)	1 (33)	15 (43)
only accuracy	1 (6)	0 (0)	1 (25)	1 (10)	0 (0)	3 (9)
Distribution	0 (0)	0 (0)	0 (0)	4 (40)	1 (33)	5 (14)
only distribution	0 (0)	0 (0)	0 (0)	1 (10)	0 (0)	1 (3)
Sample size	3 (19)	0 (0)	0 (0)	6 (60)	1 (33)	10 (29)
only sample size	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Trueness&Precision	2 (13)	1 (50)	1 (25)	7 (70)	1 (33)	12 (34)
Complete	2 (13)	0 (0)	0 (0)	5 (50)	1 (33)	8 (23)

in 2011, 4 papers in our sample from 2012, and 3 papers in 2014. RSS systems were popular in 2010 (7 papers) and 2013 (10 papers).

In 2010, almost half did not report any metrics, while most papers reported one in 2013. However, the data was usually not presented in an accessible manner, relying on pictorial representations of a c.d.f. in 5 papers in 2013 and on reports of trueness in 2010. Probably, authors do not report precision, because RSS based systems are often very imprecise, experiments can take a lot of time, and readings may be difficult to repeat because of the imprecision.

Table X summarizes the presentation of metrics in papers presenting systems based on TOF values. There are 35 papers on this topic on our sample. Again, systems using TOF were under-represented in 2011, 2012, and 2014, and we refrain from interpreting the data in those years. In 2010, reporting of metrics was not adequate. Half of the publications did not report any metric, no paper reported any data on the distribution. Just two papers allow meaningful evaluation and comparison of the metrics. This improved greatly in 2013, where almost all papers report a metric and half of them allow a meaningful evaluation and comparison.

Overall, we do not see a clear correlation between metrics and systems. This can in part be explained by the low sample sizes in each category. But we think it is more likely that there is no such correlation. Only in the case of inertial systems, the data could be used to explain a model. Next, we check whether we can see a correlation between metrics and evaluation methods.

TABLE XI
METRICS FOR SIMULATIONS IN [COUNT (%)]

	2010	2011	2012	2013	2014	Avg.
No metric	6 (55)	1 (17)	0 (0)	1 (9)	3 (50)	11 (30)
Trueness	4 (36)	4 (67)	2 (67)	6 (55)	1 (17)	17 (46)
only trueness	3 (27)	1 (17)	1 (33)	1 (9)	1 (17)	7 (19)
Precision	1 (9)	3 (50)	1 (33)	6 (55)	0 (0)	11 (30)
only precision	0 (0)	0 (0)	0 (0)	1 (9)	0 (0)	1 (3)
Accuracy	2 (18)	4 (67)	2 (67)	7 (64)	2 (33)	17 (46)
only accuracy	1 (9)	1 (17)	1 (33)	2 (18)	2 (33)	7 (19)
Distribution	0 (0)	0 (0)	0 (0)	6 (55)	0 (0)	6 (16)
only distribution	0 (0)	0 (0)	0 (0)	1 (9)	0 (0)	1 (3)
Sample size	1 (9)	1 (17)	1 (33)	7 (64)	1 (17)	11 (30)
only sample size	1 (9)	0 (0)	0 (0)	0 (0)	0 (0)	1 (3)
Trueness&Precision	1 (9)	3 (50)	1 (33)	5 (45)	0 (0)	10 (27)
Complete	0 (0)	1 (17)	1 (33)	5 (45)	0 (0)	7 (19)

2) Metrics in evaluation methods:

After looking at the summarized data, we look at preferences of methods by evaluation methods. Since we see that the method of presenting metrics varies and we could not explain this variation by the kind of system used in the research, we try to see whether there is a correlation between evaluation method and metric reporting.

Again, since we look at the relation of metrics to evaluation methods, the number of samples in each view can be rather small. This makes it hard to conclude anything definite. Again, this is the reason why we do not calculate any correlations. Also, since there are so few experiments in the grid category, the outdoor category, and the real world category, we do not evaluate these.

Table XI displays the use of metrics for papers that use simulation. There are 37 papers using simulation to evaluate the proposed method, of which 23 papers with simple simulations and 14 papers with complex simulations. Therefore, we do not distinguish between simple and complex simulations in this table. Of those papers, 10 papers report both a simulation and an experiment.

First, we are surprised by the number of papers that simulate and do not provide any metric. Especially in simulations, it should be simple to obtain metrics. Then, we were surprised that the sample size is seldom reported in 2010, 2011, 2012, and 2014. As simulations are very easy to repeat, and one should repeat simulations with varying starting conditions, we wonder why the number was not reported.

Table XII displays the use of metrics for papers that use experiments for evaluation. There are 104 papers in our sample that use some experiment to evaluate its proposal. Of those papers, 10 papers report both a simulation and an experiment.

It shows that about two-thirds demonstrate the trueness of their system, and about one third also indicate the precision of the method. About one third indicates a metric of accuracy, either directly or by indicating both trueness and precision. About a quarter of all experiments do not report a metric. This is still in the margin of error for simulations and the sample populations proportions. However, it is rather likely that experimental work generally presents a measure of trueness and probably also a measure of precision.

Table XIII displays the use of metrics for papers that use discrete points for evaluation. There are 29 papers that

TABLE XII
METRICS FOR EXPERIMENTS IN [COUNT (%)]

	2010	2011	2012	2013	2014	Avg.
No metric	9 (39)	1 (7)	7 (24)	5 (19)	2 (18)	24 (23)
Trueness	13 (57)	13 (87)	16 (55)	16 (62)	7 (64)	65 (63)
only trueness	8 (35)	6 (40)	5 (17)	6 (23)	3 (27)	28 (27)
Precision	5 (22)	7 (47)	12 (41)	11 (42)	4 (36)	39 (38)
only precision	0 (0)	0 (0)	1 (3)	1 (4)	0 (0)	2 (2)
Accuracy	6 (26)	7 (47)	13 (45)	11 (42)	6 (55)	43 (41)
only accuracy	1 (4)	0 (0)	2 (7)	1 (4)	2 (18)	6 (6)
Distribution	0 (0)	2 (13)	4 (14)	9 (35)	3 (27)	18 (17)
only distribution	0 (0)	1 (7)	3 (10)	3 (12)	0 (0)	7 (7)
Sample size	4 (17)	6 (40)	15 (52)	11 (42)	4 (36)	40 (38)
only sample size	0 (0)	0 (0)	1 (3)	0 (0)	0 (0)	1 (1)
Trueness&Precision	5 (22)	7 (47)	11 (38)	10 (38)	4 (36)	37 (36)
Complete	2 (9)	4 (27)	9 (31)	7 (27)	3 (27)	25 (24)

TABLE XIII
METRICS FOR DISCRETE POINTS EXPERIMENTS IN [COUNT (%)]

	2010	2011	2012	2013	2014	Avg.
No metric	1 (20)	0 (0)	1 (11)	2 (22)	1 (33)	5 (17)
Trueness	3 (60)	3 (100)	7 (78)	7 (78)	2 (67)	22 (76)
only trueness	2 (40)	0 (0)	3 (33)	2 (22)	1 (33)	8 (28)
Precision	1 (20)	3 (100)	4 (44)	5 (56)	1 (33)	14 (48)
only precision	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Accuracy	2 (40)	3 (100)	5 (56)	5 (56)	1 (33)	16 (55)
only accuracy	1 (20)	0 (0)	1 (11)	0 (0)	0 (0)	2 (7)
Distribution	0 (0)	1 (33)	1 (11)	2 (22)	1 (33)	5 (17)
only distribution	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Sample size	0 (0)	2 (67)	6 (67)	5 (56)	1 (33)	14 (48)
only sample size	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Trueness&Precision	1 (20)	3 (100)	4 (44)	5 (56)	1 (33)	14 (48)
Complete	0 (0)	2 (67)	4 (44)	4 (44)	1 (33)	11 (38)

report experiments using discrete points. Overall, we see that experimental work is usually characterized by providing metrics. About three-quarters indicate the trueness, whereas a sixth does not provide metrics.

Table XIV displays the use of metrics for papers that use office walks for evaluation. There are 44 papers that report experiments using office walks. About two-thirds indicate the trueness, whereas a sixth does not provide metrics.

Table XV displays the use of metrics for papers that use multiple methods for evaluation, most importantly simulations and experiments. While there are only 13 papers in this category, we still report it, because the overall distribution is pretty close to the distribution of methods used in experiments.

3) Summary: Overall, reporting and proper reporting of metrics seems to depend on the use of actual experiments for evaluation. This surprises a bit, because it is harder to measure in actual experiments.

One should not judge the quality of the papers from these criteria alone. It is justifiable to not to report some or all metrics. For example, reporting a variance does not make much sense if the sample size is small. Some authors base their results on three repetitions of an experiment.

Some experiments were reported to demonstrate the system and not to measure its performance. On the other hand, once the principles of a method are properly understood, a performance evaluation is needed to demonstrate that one contributed to the state of the art. For example, if a new localization algorithm is suggested, its performance should be evaluated and compared to other methods to understand the utility of the newly proposed method.

Still, we are surprised that less than a third of the papers report an estimate of trueness and precision, and only about a fifth report values that are meaningful to compare statistically. This is especially surprising, since given the experimental setup and the collected data, it ought to be simple to calculate and present those values, or at least explain why the values were not calculated.

TABLE XIV
METRICS FOR OFFICE WALKS IN [COUNT (%)]

	2010	2011	2012	2013	2014	Avg.
No metric	3 (33)	1 (13)	2 (18)	1 (10)	1 (17)	8 (18)
Trueness	6 (67)	7 (88)	7 (64)	6 (60)	4 (67)	30 (68)
only trueness	5 (56)	4 (50)	1 (9)	3 (30)	1 (17)	14 (32)
Precision	1 (11)	3 (38)	7 (64)	4 (40)	3 (50)	18 (41)
only precision	0 (0)	0 (0)	1 (9)	1 (10)	0 (0)	2 (5)
Accuracy	1 (11)	3 (38)	7 (64)	3 (30)	4 (67)	18 (41)
only accuracy	0 (0)	0 (0)	1 (9)	0 (0)	1 (17)	2 (5)
Distribution	0 (0)	0 (0)	0 (0)	5 (50)	1 (17)	6 (14)
only distribution	0 (0)	0 (0)	0 (0)	2 (20)	0 (0)	2 (5)
Sample size	2 (22)	3 (38)	6 (55)	3 (30)	2 (33)	16 (36)
only sample size	0 (0)	0 (0)	1 (9)	0 (0)	0 (0)	1 (2)
Trueness&Precision	1 (11)	3 (38)	6 (55)	3 (30)	3 (50)	16 (36)
Complete	0 (0)	2 (25)	4 (36)	1 (10)	2 (33)	9 (20)

TABLE XV
METRICS FOR MULTIPLE METHODS IN [COUNT (%)]

	2010	2011	2012	2013	2014	Avg.
No metric	1 (33)	0 (0)	1 (33)	0 (0)	0 (0)	2 (15)
Trueness	2 (67)	2 (100)	1 (33)	3 (75)	0 (0)	8 (62)
only trueness	1 (33)	0 (0)	0 (0)	0 (0)	0 (0)	1 (8)
Precision	1 (33)	2 (100)	1 (33)	4 (100)	0 (0)	8 (62)
only precision	0 (0)	0 (0)	0 (0)	1 (25)	0 (0)	1 (8)
Accuracy	1 (33)	2 (100)	2 (67)	3 (75)	1 (100)	9 (69)
only accuracy	0 (0)	0 (0)	1 (33)	0 (0)	1 (100)	2 (15)
Distribution	0 (0)	0 (0)	0 (0)	2 (50)	0 (0)	2 (15)
only distribution	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Sample size	0 (0)	1 (50)	1 (33)	3 (75)	0 (0)	5 (38)
only sample size	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Trueness&Precision	1 (33)	2 (100)	1 (33)	3 (75)	0 (0)	7 (54)
Complete	0 (0)	1 (50)	1 (33)	3 (75)	0 (0)	5 (38)

TABLE XVI
USE OF BASELINES IN [COUNT (%)]

	2010	2011	2012	2013	2014	Avg.
Total	0 (0)	2 (10)	2 (6)	5 (14)	1 (6)	10 (7)
Inertial	0 (0)	0 (0)	1 (3)	1 (3)	0 (0)	2 (1)
Map matching	0 (0)	2 (10)	1 (3)	0 (0)	0 (0)	3 (2)
RSS	0 (0)	0 (0)	0 (0)	2 (6)	1 (6)	3 (2)
TOF	0 (0)	0 (0)	0 (0)	1 (3)	0 (0)	1 (1)
Sound	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Other	0 (0)	0 (0)	0 (0)	2 (6)	0 (0)	2 (1)

E. Baseline Category

Table XVI summarizes the distribution of papers to the metric categories. We see that few evaluations (about 7.1 %, or 13 publications) define an external baseline for the evaluation. Of those publications, 10 are considered related to our work. As we found so few papers that actually did, we cannot identify any additional properties of the choice of baselines. Distributed to the system categories, we find 2 papers about inertial systems, 3 papers about map matching, 3 papers about RSS methods, 1 paper about TOF methods, 0 about sound based systems and 2 about other systems; 1 paper combines an inertial system with an RSS based system.

This low figure is very surprising at first. We offer the following explanations for this observation:

1) Many proposals are new and there is little that may serve as a baseline to compare to. The main purpose of the evaluation is to demonstrate that the proposed method is effective.
2) Methods that may serve as a baseline are ill-defined. To compare, one has to establish the baseline within ones own environment and using the available equipment to make sure that only the performance of the system under evaluation (SUE) is compared and not some other parameter. This requires the ability of the recreation that fails with incomplete disclosure.
3) Effort is too large. For a proper evaluation, each experiment should be repeated multiple times to increase the precision. Experiments may take many hours including set-up and tear-down. Adding a baseline easily doubles the effort.

Our figure does not consider an internal baseline, e.g. if filters are applied. In such papers, we see the unfiltered results as a baseline and filtered or processed results to establish an improvement.

We observed that few quantify or report the effect of their proposed method. It is not established whether the improvement can be explained by noise or whether there is a statistical significance to the improvement. This is especially true in papers that propose combinations of methods, where each additional method shows less effect. We cannot tell whether the addition has actually any effect, because no variance is reported, or if c.d.f. are compared, then no Kolmogorov-Smirnov (K-S) statistics is provided. The K-S statistics would help in understanding and quantifying the differences between the c.d.f..

The lack of a baseline makes it difficult to quantify the progress made over the years.

IV. Accuracy of this study

Neither did we have the time nor the resources to analyze all papers published in the IPIN conference series. We also did not investigate other big conferences and symposia on indoor positioning, like UPIN-LBS or WPNC, or other conferences where papers on indoor positioning or indoor navigation were published. Thus, this study is strictly about the publication and selection culture of IPIN.

A. Systematic error

The main sources of systematic error are misclassification and publication selection bias.

1) Classification errors: Each paper was read and classified by at least one person. Usually, the papers make a strong statement on what they are about. But the reader may still have misclassified the publication.

Reasons for misclassification are ambiguities between the classes definition. It is, e.g., not so clear where we draw the line between simple and complex simulations.

Some papers did not fit clearly into one of the categories and are therefore not counted in any. If, for example, a paper describes an experiment and does not give any numerical metrics, it may be counted as a paper without any metrics. This happens when the paper displays its measurements in diagrams only, and no clear value can be obtained from the figure.

As we did not classify each paper by more than one person, we have no estimate of classification errors.

2) Publication selection bias: The second source of systematic error is that the selection of articles we reviewed could be biased towards a particular style or quality.

We selected papers for this study in the following way: We wrote a program that downloaded an index of all papers that were published in IEEE Xplore in some IPIN conference, and have it pseudo-randomly select 30% of the papers from this set. No interference by the authors happened in this step.

Concerning the publications of IPIN 2014, those were not available on IEEE Xplore as of this writing. We wrote a similar program using the papers made available at the conference, while excluding all papers associated to the poster session.

We are confident that the selection of papers is unbiased. We claim that this method resulted in a fairly representative selection of papers published at IPIN.

B. Statistical errors

Statistical errors in our study are random classification mistakes and the accuracy of our estimations.

Since we have no measure of the classification errors, we cannot estimate it. We are confident that not many papers were misclassified.

Concerning the accuracy of our estimations, we can estimate the margin of errors of our values. With a confidence of 95%, a margin of error of about 0.03 for the proportions reported for the whole conference series and a margin of error of about 0.05 to 0.07 for the proportions reported for individual years. We could have selected proportions such that the margin of error is the same for each year. That would have implied varying proportions of papers in each year and it would have increased and increased the total proportion of papers, thus decreasing its margin of error. We opted for the more economical approach and note that all proportions have a margin of errors of up to 0.07 with 95% confidence.

Per category, the margin of error is even larger. First, we have only an estimate of the total number of papers for each category, and second, the sample may even be lower.

We tried to refrain from evaluating data sets where the margin of error increases above 0.15 at a confidence level of 95%. It would be impossible to make any statement in this situation, if the margin of error allows also a contradicting explanation. This probably happens, if the number of papers is below 30.

C. Overall accuracy

Concerning the data on all papers selected in IPIN, we are confident that we show a pretty accurate picture of all publications. We guess that the margin of error of misclassifying is about the same as the margin of statistical estimation errors. Thus, the proportions are probably pretty accurate.

The data per year and per category is less accurate. However, this does not invalidate the overall conclusion we draw.

V. Recommendations & Conclusion

We find that 95 % of all related papers of IPIN we have looked at perform some kind of evaluation. Physical experimentation plays an important role and is used in 77 % of all related papers. We see a trend towards more complex simulations and a steady share between different physical evaluation methods. We challenge the relevance of outdoor only experiments to indoor systems, which we see in 3 % of related publications.

While the reliance on evaluation looks high, the quality of description is marbled. A high percentage of publications describe their methods of ground truth data gathering poorly at best. Although many authors do not write about that process, we assume manual measurements using rulers and distance meters were used for ground truth positions.

We did not evaluate the quality of all experiments. We also do not claim that experiments that do not follow our metrics are necessarily bad. A publication may fail our metrics because of the authors intention. Authors demonstrating only the concept and its feasibility give examples of this. Some publications fail our criteria, because they measure and evaluate differently.

The scientific standard of an experiment is the formulation of a hypothesis about a system's performance, a description how the hypothesis is tested, and what steps were taken to eliminate all forms of bias. For example, it is known that PDR is subject to drift, i.e. the observed error increases with distance. We seldom see statements that relate the observed error to the distance or time of evaluation. Only 35% report how often an experiment was performed. We are not made aware of data that may contradict the hypothesis. We never see a quantitative estimate of the systematic errors in the experimental evaluation. For example, with GPS based reference systems in cities, we should track the estimation precision of the GPS. If this precision is 10 meters and we see an error of 5 meters, the experiment says little about the performance of the SUT. Lack of full disclosure of the experiments makes it hard to validate the results independently.

Experimentation is central to scientific processes. We appeal to authors to write about their method of ground truth data gathering in the spirit of comparison, reproducibility, and explanation. This concerns not only the true position itself, but also the time of a measurement, which is often not mentioned in publications. We also feel that the research community should ask for more real world experiments. It is important to test an approach under different conditions. For repeatability we encourage the use of automated reference systems, like robots or optical systems. Simulation could be improved by sharing data sets to create a set of standard benchmarks for indoor localization systems.

References

[1] *2010 International Conference on Indoor Positioning and Indoor Navigation, IPIN 2010, Zurich, Switzerland, September 15-17, 2010*. IEEE, 2010. [Online]. Available: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=5637226

[2] *2011 International Conference on Indoor Positioning and Indoor Navigation, IPIN 2011, Guimaraes, Portugal, September 21-23, 2011*. IEEE, 2011. [Online]. Available: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6062621

[3] *2012 International Conference on Indoor Positioning and Indoor Navigation, IPIN 2012, Sydney, Australia, November 13-15, 2012*. IEEE, 2012. [Online]. Available: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6409516

[4] *2013 International Conference on Indoor Positioning and Indoor Navigation, IPIN 2013, Montbeliard, France, October 28-31, 2013*. IEEE, 2013. [Online]. Available: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6811041

[5] *2014 International Conference on Indoor Positioning and Indoor Navigation, IPIN 2014, Busan, Korea, October 27-30, 2014*. IEEE, 2014.

[6] W. F. Tichy, P. Lukowicz, L. Prechelt, and E. A. Heinz, “Experimental evaluation in computer science: A quantitative study,” *Journal of Systems and Software*, vol. 28, no. 1, pp. 9–18, Jan. 1995.

[7] W. F. Tichy, “Should computer scientists experiment more?” *Computer*, vol. 31, no. 5, pp. 32–40, May 1998. [Online]. Available: http://dx.doi.org/10.1109/2.675631

[8] J. Wainer, C. G. N. Barsottini, D. Lacerda, and L. R. M. de Marco, “Empirical evaluation in computer science research published by ACM,” *Information and Software Technology*, vol. 51, no. 6, pp. 1081–1085, Jun. 2009.

[9] C. Andujar, V. Schiaffonati, F. A. Schreiber, L. Tanca, M. Tedre, K. van Hee, and J. van Leeuwen, “The role and relevance of experimentation in informatics,” *Informatics Europe, Report*, 2013.

[10] M. Tedre and N. Moisseinen, “Experiments in computing: A survey,” *The Scientific World Journal*, vol. 2014, 2014.

[11] J. B. Buckheit and D. L. Donoho, “WaveLab and reproducible research,” in *Wavelets and Statistics*, ser. Lecture Notes in Statistics, A. Antoniadis and G. Oppenheim, Eds., New York, 1995, vol. 103, pp. 55–81.

[12] C. Collberg, T. Proebsting, and A. M. Warren, “Repeatability and benefaction in computer science research,” University of Arizona, Arizona, Tech. Rep. TR 14-05, Feb. 2015. [Online]. Available: http://reproducibility.cs.arizona.edu/v2/RepeatabilityTR.pdf

[13] K. Popper, *The Logic of Scientific Discovery*. Hutchinson & Co., 1959, cited after edition by Routledge, 2002.

[14] R. Gentleman and D. T. Lang, “Statistical analyses and reproducible research,” *J. Computational and Graphical Statistics*, vol. 16, no. 9, pp. 1–23, 2007.

[15] R. A. Fisher, *The Design of Experiments*, 8th ed. Edinburgh: Oliver and Boyd, 1966.

[16] *ISO 5725-1 : 1994 ”Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitionsÂt’Ât’*, International Organization for Standardization, Geneva, Switzerland, 1994.

[17] R. Jain, *The art of computer systems performance analysis - techniques for experimental design, measurement, simulation, and modeling.*, ser. Wiley professional computing. Wiley, 1991.

[18] C. J. Willmott and K. Matsuura, “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance,” *Climate Research*, vol. 30, no. 1, p. 79, 2005.

[19] S. Schmitt, H. Will, B. Aschenbrenner, T. Hillebrandt, and M. Kyas, “A reference system for indoor localization testbeds,” in *Indoor Positioning and Indoor Navigation (IPIN), 2012 International Conference on*, 2012, pp. 1–8.

View publication stats