Practical Evaluation and Tuning Methodology for Indoor Positioning Systems

Grigorios G. Anagnostopoulos, Carlos Martínez de la Osa, Tiago Nunes, Abbass Hammoud, Michel Deriaz and Dimitri Konstantas

Information Science Institute, GSEM/CUI
University of Geneva
Geneva, Switzerland
Email: $\{grigorios.anagnostopoulos, carlos.martinez, tiago.nunes\}$
$\{abbass.hammoud, michel.deriaz, dimitri.konstantas\}$@unige.ch

***Abstract*— How do we evaluate the performance of an indoor positioning system? In addition, in which way can the system be optimally tuned for a certain environment? These are the questions addressed in this study. We propose a practical, cost efficient methodology for evaluating and tuning indoor positioning systems. The methodology has two main phases. In the first online recording phase, the ground truth information is gathered, and raw signals are recorded. In the second phase, offline positioning algorithms utilise the recorded information to infer position estimations which can then be precisely evaluated. An automatic parameter optimization methodology, which recommends optimal tunings for the positioning algorithm, is presented as a key utility of this work. An overall advantage of the proposed method is the fact that the recorded data guarantee the repeatability of tests, and allow consistent comparisons among different algorithms, creating the perspective of a testbed based on real data. The implementation of the methodology is exemplified with the presentation of the GpmLab Android application and the GpmStudio desktop platform, tools which assist our main positioning framework, the Global Positioning Module (GPM).**

Index Terms*—*Indoor positioning, Tuning, Ground truth definition, Positioning evaluation, Tracking

I. Introduction

During the last years, the domain of indoor positioning has attracted a lot of attention of both the academic community and the industry. So far, no universal standard has been established defining the way of measuring and evaluating the performance of a positioning system.

Having a consistent way of measuring the performance of positioning systems is indispensable. It is evident that a reliable evaluation methodology allows a research group to share with the community the performance of their positioning system. Moreover, it provides the group with the ability of having a continuous overview of the impact of the improvements that they attempt on the existing methods, as well as evaluating new algorithms and comparing them with others.

One of the main challenging issues in the evaluation procedure is the ground truth acquisition. In a relevant recent study [1], the authors studied the publications from five recent indoor positioning conferences, and found that a high percentage of the authors describe their methods of ground truth data gathering poorly or do not describe them at all. Furthermore, they express their concern about the reproducibility of the experiments of the field. Given the above, we believe that the discussion of the ways that we test, measure and evaluate indoor positioning systems is indispensable.

An interesting study in this domain [2], describes the evaluation methods by grouping them in three categories. The first method is (i) *Static evaluation*, in which a device is left at a specific location for a sufficient amount of time. Then, the evaluation is made by analysing the position estimates that the System Under Test (SUT) provided during this time. A major drawback of this category is that it does not reflect most real use cases, which are expected to be more dynamic. This requirement is addressed by the other two categories: (ii) *Dynamic evaluation using a reference positioning system* and (iii) *Dynamic evaluation with predefined geometrical paths*. Category (ii) methods use a positioning system with high accuracy in order to evaluate the SUT. The requirement of using another system may imply significant additional deployment effort and cost. In the third category (iii), specific paths are defined in advance at a test field, and then followed by a person that records the estimations of the SUT. The method presented in this paper belongs in this category, which combines many advantages, such as realistic capture of real-life usage scenarios, ease of deployment, and cost efficiency.

Furthermore, apart from the evaluation methodologies that have started to draw significant attention by the indoor positioning community, the methodologies of optimally tuning a system is also a domain worth discussing, sharing ideas on and improving. Assume a positioning expert called to deploy a positioning system, tune it properly, present it and evaluate its performance. It is common in the community to discuss the parts of deploying and its requirements, the positioning methods, as well as the ways of evaluating the system's performance. It is not rare though, that the step of

This work was co-funded by the State Secretariat for Education, Research and Innovation of the Swiss federal government and the European Union, in the frame of the EU AAL project SmartHeat (aal-2014-153).

how the system is tuned accordingly is left blurry, mentioning for example an empirical selection of the appropriate settings. The question of how to tune a positioning system at a new deployment has drawn our interest. Investigating this issue, gave as fruit the present work. This study is the natural and significant expansion of a previous work [3], where we focused on ground truth definition and position evaluation of online positioning solutions. In the current study, we present a practical evaluation and tuning methodology, which simplifies the procedure of optimally tuning a positioning system at a new deployment. A more detailed description is given at the end of the following section, where after presenting the related work we proceed in explaining the contributions of the current work.

The rest of this paper is organized as follows. In Section II, the related work is discussed, along with the contributions of this study. After presenting the proposed method theoretically in Section III, the implementation of the method with the related software tools are shown in Section IV. In Section V, we provide an overview of the use cases of the methodology and discuss its particularities. Lastly, the future directions of this work are discussed in Section VI.

II. Related Work

The definition of the ground truth information used to evaluate positioning systems is a crucial subject for the indoor positioning community. As mentioned earlier, a recent study [1] expresses the concern of its authors about the lack of strictly defined procedures of ground truth acquisition. Furthermore, during their analysis, the authors of that work found several studies that refer only to the spatial information of the ground truth. Hence, they highlight that the temporal information of both the ground truth and the position estimates is also indispensable.

There are studies [4] that try to simplify the procedure of evaluating positioning systems, by defining the ground truth only with spatial (and no temporal) information, targeting to approximate the cumulative statistics of the positional error. Nevertheless, even for achieving this approximation, strict requirements are set concerning the symmetry of the deployed access points and the form of the path followed for the recording. More precisely, it is required that all base stations are deployed symmetrically, so that the cross-track and along-track errors can be considered statistically independent. Then, an approximation of the statistical distribution of the overall positioning error is given, based on the cross-track error (vertical deviation from the predefined path).

Recent indoor positioning competitions, like the Microsoft Indoor Localisation Competition [5] (performed in the context of the ACM IPSN conferences), and the EvAAL competition [6] (in the context of the IPIN conferences) define strict methodologies of evaluating positioning systems. In IPSN, a list of specific evaluation points are defined in the test area. A tester carries a device above each of the evaluation points, waits for a couple of seconds at that point, and records the location reported by the SUT. This methodology evaluates precisely the accuracy of position estimates statically, as discrete points, though it does not aim to evaluate a continuous dynamic movement of users.

The competition of IPIN defines a dynamic way of evaluation, aiming to cover this aspect. Again, a list of specific evaluation points are defined in the test area. The tester follows the path defined by a sequence of evaluation points with a natural pace, and without performing an artificial pause at the evaluation points. The tester records at the device the timestamp of the moment he passed from each evaluation point, so that the most recent position estimate is compared to the position of the evaluation point. This methodology covers the more realistic use cases of moving users, as all measurements are taken dynamically. It is noteworthy though, that in contrast with the IPSN competition, the ground truth position and the position estimate are not taken simultaneously, introducing implicitly as a factor in the evaluation, the frequency of production of position estimates. Thus, this method has the feature of dynamic recording, and the feature of evaluating at random moments of the walk (when passing by an evaluation point) the latest position estimate that a user of the SUT would see in his screen.

An impressive contribution in the direction of benchmarking of indoor positioning systems has been the work performed in the context of the EVARILOS project [7],[8]. This project identifies the pitfall of reproducing research results of indoor localization in real life scenarios, as they suffer from uncontrolled RF interference and from the weakness of numerous published solutions being evaluated under individual, not comparable and not repeatable conditions. EVARILOS, as well as other works [2], [9], [10], [11] perform their evaluation with the use of another system of high accuracy, such as a robot or cameras, as a reference system. Using a precise robotic system as reference is a valuable solution for strictly measuring and evaluating systems in controlled test areas. Nevertheless, this approach has some limitations, such as the cost, the speed with which a broad area can be measured, and its mobility limitations (obstacles, stairs, etc.). Furthermore, it might be much less complex for a human tester to perform the task in case that an evaluation is needed in the context of tuning a positioning system, when deploying it in a crowded public space (like a shopping mall), that might also have the above mentioned limitations. Nevertheless, when focusing purely on a precise evaluation, methods that utilise high accuracy reference systems are precious.

Apart from the definition of the ground truth, recording simultaneously the raw data received from the technologies used for positioning, is of high importance. Having this information combined (ground truth and raw data), facilitates the reproducibility of tests since using the same data allows consistent comparisons among systems. In an influential sequence of publications [11],[12],[13],[14],[15], the authors from the Free University of Berlin have worked in this direction. After defining a reference system for indoor localisation systems [11], they present the concept of a visual testbed [12], in which a robot spans an area recording both the raw signals and the ground truth information throughout the area. These data are offered to be used as a testbed for positioning systems. Following, they perform several experimental evaluations of systems [13],[14],[15], using their robotic reference system.

In the current work, we have tried to address most of the problems discussed above. Similarly to both the above mentioned competitions (IPSN and IPIN), the proposed methodology utilizes a predefined path in which a tester passes through predefined checkpoints. As our intention was to evaluate the continuous performance of the SUT, the evaluation is not limited only to the checkpoints, but concerns the continuous production of estimations of the system (as opposed to the aforementioned competitions). We do so, by introducing two ways of interpolating the ground truth information. Furthermore, opposite to the IPIN competition, the ground truth position and the position estimate are simultaneous, without having to perform the artificial stop required in the IPSN competition. In addition, since the data collected with the proposed methodology can be repeatedly used to test offline positioning algorithms and evaluate their performance, the perspective of a testbed is created. Lastly, we introduce the perspective of utilising the previous steps, to create a tuning tool that facilitates the task of optimally tuning a system.

III. Proposed Method

In this section, we present the proposed methodology in its full extent. Firstly, we offer an overview of the global architecture. Then, we proceed with a detailed presentation of the online data recording method, followed by the offline positioning and evaluation part. Lastly, we analyse the procedure of optimally tuning the positioning algorithm.

A. Architecture

In this subsection, we describe the workflow of our proposed methodology. It consists of two main phases: an online recording phase which takes place at the environment of the deployment, and an offline phase. The workflow is summarized in Figure 1. During the first online phase, the spatio-temporal ground truth information is gathered (the spatial information (x, y), plus the time t), along with the received raw signals (the signal s, at time t). In the second phase, an offline positioning algorithm is used to infer position estimates (the spatial information (x, y), along with the time t), using the recorded raw data. Different settings of the parameters of the tested algorithm, result to different lists of position estimates (containing spatial information (x, y), plus the time t) which can be directly compared with the ground truth information, and evaluated. By comparing the position estimates of the offline positioning algorithm with the ground truth, the parameters of the positioning algorithm can also be properly tuned, and its performance can be optimized.

B. Online data recording

For the online phase, a predefined path needs to be designed. This path consists of a list of positions that will be followed, one after the other, by a tester and will serve as checkpoints.

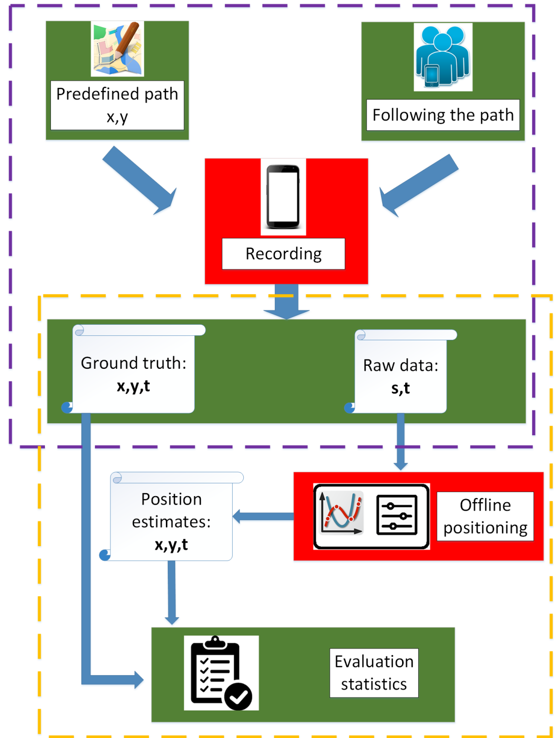

Fig. 1. The overall workflow of the proposed methodology. The purple box highlights the online part, while the yellow one indicates the offline part.

The tester will have to follow this predefined path, holding the mobile device that will continuously record all the raw signals that it receives during the path, along with the timestamp of each reception. The tester has simply to indicate to the recording application the moment at which he passes over each checkpoint. In this way, every predefined checkpoint is marked with the exact timestamp indicating when the tester passed by it. The tester should move on the path, which is composed of the straight linear segments that connect these checkpoints, at a steady pace. Note that the pace should be steady during a linear segment connecting two checkpoints, but it is not required to be the same among different segments.

With this procedure, the exact time that each checkpoint was crossed is recorded. The spatio-temporal ground truth information for every moment between the crossing of two checkpoints can be inferred with linear interpolation. Under the assumption of a steady pace between consecutive checkpoints, the result of the linear interpolation is the accurate ground truth. The density of the checkpoints handles a trade-off between minimising the effort included in the recording procedure and minimising the interpolation error. Having dense checkpoints allows the user to frequently inform the recording application about his true position, reducing the length of the segments in which the interpolation is done. In this way, the error introduced by the interpolation is minimized. On the other hand, a more sparse placement of checkpoints simplifies the procedure, with the risk of introducing a higher error due to interpolating over bigger segments.

There are different approaches that utilize a predefined path. A common method used (for example by the IPSN competition) is the one in which the tester has to follow a path with checkpoints, at which he is called to halt for a couple of seconds, in order to receive a position estimate, and then continue the path. Even if the SUT can continuously utilize information received between the checkpoints (such as inertial sensor data), the evaluation is actually a static one, as the tester stops for taking a measurement. On the other hand, there exist more dynamic evaluation methods (as the one performed at the IPIN competition), where the user follows a path, without stopping at the checkpoints. However, even if the measuring procedure has the element of a continuously moving user, the evaluation of the SUT is not continuous, but sporadic since only a sampling of the most recent estimations at the checkpoints is evaluated. Therefore, despite being dynamic, the method does not fully capture the experience of a user that would actually follow the path while receiving position estimates.

In this work, we introduce two methods of interpolating the recorded checkpoint indications of the tester, in order to obtain the spatio-temporal ground truth information, and based on this, to evaluate the position estimates. These methods are discussed in the following subsection.

C. Offline positioning and ground truth interpolation

In the preceding online phase, all the necessary data needed for feeding an offline positioning algorithm were recorded. The collected raw data are utilized by the offline positioning algorithms, which produce position estimates from the recorded raw signals. Thus, for any signal s received at time t, a position estimate x, y is inferred. The same data can feed different positioning algorithms, or the same algorithm with several different settings of its parameters, resulting in different position estimates.

As a simple verification method of the correct functioning of the implementation of the methodology, apart from recording only the raw data that feed an offline positioning algorithm, it is recommended to also run the online version of the algorithms, and record the position estimates. The results of the online and the offline algorithms should be identical, under the same parameter tuning and the same data.

The spatio-temporal information of the position estimates inferred, can be compared with the ground truth, in order to evaluate the accuracy of the estimation. In order to evaluate an estimate x, y for the time t, the ground truth position at time t is needed. The ground truth information available from the offline phase is limited to the positions of the checkpoints and the corresponding time the tester was there. On the other hand, as the signal recording was continuous, the position estimates will have timestamps that chronologically lay between the timestamps of two consecutive checkpoints. In order to infer the true position of the user at a time between two consecutive checkpoints, we have proposed two interpolation approaches [3], which are presented bellow.

1) Interpolation per estimation update: This first solution is the most intuitive aproach, which is to calculate an interpolated ground truth point in the path, at the time that every new position estimate was received. In order to do this, every time the algorithm processes a position estimation update, it checks its timestamp, and calculates the corresponding interpolated point in the ground truth path for this timestamp. For example, assume that a tester was at checkpoint A at time $t_A = 0$ and at checkpoint B at time $t_B = 10$. If he receives a position estimate at time $t = 1$, his ground truth position is inferred by linear interpolation, and is assumed to be at the 10% of the distance of the linear segment linking A to B.

This approach is useful for evaluating each position estimate individually, and checking if the logic of the algorithm that calculates the position estimates is accurate. On the other hand, there is a drawback in this technique, in the case that a position provider is updating the position estimations with a very low frequency. In those cases, only a few points will be taken into account for the statistical analysis. A big time lapse between consecutive position estimates can significantly deteriorate the perceived utility for the user, who will have the feeling that the estimations are lagging, and thus, the estimates will not be representative for the whole path. For example, assume a scenario where a tester starts recording a path which is one kilometre long, and he receives only two position estimates, one at very the beginning and another one at the very end of the path, both of them being very accurate estimates. Using this method, the evaluation will later conclude that the position provider is very accurate, despite the fact that during the whole path the latest position estimate was the first one, placing the user at the beginning of the path. With this motivation, we proceed by discussing the second interpolation method.

2) Periodical interpolation: In order to better evaluate the end user's perception of the positioning provider, a different interpolation method is proposed. In this case, the tester will choose a fixed time interval, the period T_i. Then, the system creates an interpolated ground truth point in the path periodically every T_i time units $(0, T_i, 2T_i, 3T_i, \ldots)$. These points will be compared to the most recent position estimation received at the corresponding timestamp. In this way, we have a periodical evaluation of how close the estimation that would appear on the user's screen is to the true position. The shortest the time interval T_i is, the more representative of the continuous user experience the evaluation will be, increasing though proportionally the processing time. Let T_p be the expected period with which the position provider offers its position estimates. Thus, selecting a $T_i \ll T_p$ is recommended. This method copes with the limitation of the first algorithm, as it very frequently compares the position of the tester with the estimates.

D. Evaluation

After the interpolation part, pairs of ground truth positions (x_g, y_g) and position estimates (x_e, y_e) will have been created, regardless of which interpolation method was selected. With this information, the evaluation of the positioning algorithm can take place. The most popular way of evaluating positioning systems is by utilizing the 2D euclidean distance error d_e

(Equation 1), and the relative statistics that can be inferred based on it.

$$d_e = \sqrt{(x_g - x_e)^2 + (y_g - y_e)^2} \tag{1}$$

It is common in evaluating positioning systems to present the Cumulative Distribution Function (CDF) of the error, and report basic statistical metrics like:

- the mean
- the median
- the 75th percentile
- the standard deviation

All these metrics can be easily calculated from the recorded data. According to the kind of information that are included in a position estimate, other metrics can be also added. For example, if the position estimate and the corresponding ground truth include extra information like: room, floor, building, etc., this allows an estimation of room accuracy, floor accuracy, and building accuracy as a ratio of correct detection. Furthermore, other metrics regarding the smoothness of the estimated path, based on the full sequence of position estimates of a moving tester (as discussed in [3],[16]) could also be used. Potential metrics that may be used have been extensively discussed in relevant studies [7],[16].

The tester can use the metric of his choice in order to evaluate the performance of the positioning SUTs, according to his needs. The key point of the evaluation with the proposed method is that the results of different algorithms, or different tunings of the same algorithm, can be consistently compared as they are produced by the same raw data recordings.

E. Tuning and optimization

The tuning of the positioning algorithm can be done either manually, by a person that is experienced in the particularities of the positioning algorithm, or automatically by using optimization techniques. The logic of the tuning procedure is presented in Figure 2.

Initially, the offline positioning module gets as input the recorded raw signals with their timestamp (s, t), and based on some setting P of its algorithm's parameters, the module produces its outcome: a list of position estimates with their corresponding timestamp (x_e, y_e, t_e). Then, by comparing the estimates to the ground truth (x_g, y_g, t_g), the tester can characterise the performance of the setting P, using a statistical metric $f(P)$ (e.g. mean error, median error, etc.) that evaluates the outcome of the estimation. This procedure can be repeated for several parameter settings P, obtaining an evaluation for each of them, in search of the optimal one P^*.

$$P^* = \arg\min_P f(P) \tag{2}$$

Without the possibility of running the positioning offline while having the spatio-temporal ground truth information, the tuning of a positioning system can become a very tedious task. The tester would need to repeatedly traverse the environment with an online positioning application. The performance of the online positioning algorithm should be evaluated for many candidate parameter settings. In that case, one formal solution would be to repeat this procedure many times, and to record every time the position estimates, while also gathering ground truth information. The complexity of this task can increase exponentially with the number of parameters, due to the possible combinations. Furthermore, the environmental conditions can change among the different recordings, introducing a potential bias in the selection of the optimal setting. A less formal, but undeniably existing method in practical deployments, is the approach of testing-and-setting with visual evaluation, in which an experienced engineer tries empirically several parameter settings, evaluating visually, in real time, the position estimations. The sketchiness of this method is evident, but due to lack of formal evaluation and tuning methodologies, it is sometimes met.

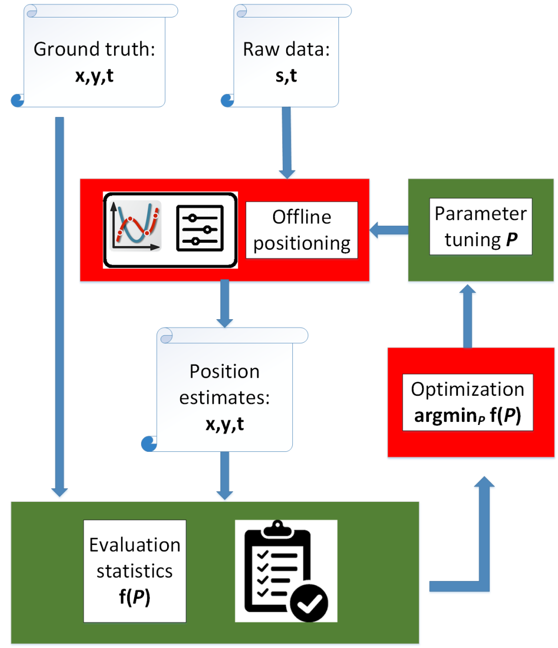

Fig. 2. The workflow for the optimization of the parameter tuning P that determines the positioning algorithm's performance.

With the proposed method, the search for the optimal setting P^*, can be done offsite either manually or automatically.

Manual parameter tuning: In the manual approach, the tester can run the offline algorithms, trying several candidate settings, one by one. For each setting, all the evaluation metrics characterising the performance of the estimation (as discussed in Section III-D) can be calculated. Also, the tester can visualize the estimated path in order to obtain a visual feeling, similar to the one a user would have while using the system. In this way, a tester can find the most appropriate tuning for a specific deployment. Moreover, evaluating without any effort several parameter settings, may offer to the tester an insight of the practical effect of each parameter to the algorithm. A thorough analysis like this, may enrich the intuition of the tester concerning the algorithm under test or even inspire him about how to improve the algorithm itself.

Automatic parameter optimization: The procedure of the automatic optimization approach is straightforward. An optimization algorithm that runs offline, tries to solve the optimization problem defined in Equation 2. The tester needs to choose four elements: (i) the recorded data over which the optimization will take place, (ii) the evaluation metric $f(P)$ (the objective function, as mentioned in the optimization terminology), (iii) the optimization algorithm, and (iv) the search space. The elements (i) and (ii) that concern the recorded data and the evaluation metric, have been discussed earlier in this work. The optimization algorithm (iii) can be chosen by the tester according to the problem's characteristics. For example, a full search algorithm would evaluate all valid candidate solutions P, in order to provide the optimal one. This approach though is computationally expensive. Therefore, heuristic optimization algorithms could be useful in case of significant computational complexity. The search space (iv) is the area of all valid values of the parameter tuning P, in the context of the optimization problem. Each parameter of the algorithm that needs to be properly tuned can be one of the dimensions of the multidimensional search space. The set of valid values for each parameter/dimension should be defined. The search space will have as many dimensions, as the number of the parameters to be optimized.

IV. Implementation

In this section, we present the software tools which allow the execution of the presented workflow. Firstly, we introduce the mobile application GpmLab, with which we perform the raw data recording along with the ground truth gathering. An early version of this application, without the raw data recording option, is also discussed here [3]. Furthermore, we present the GpmStudio platform, which runs the offline positioning, and offers evaluation and tuning capabilities. Both GpmLab and GpmStudio tools assist the main positioning framework of our team, the Global Positioning Module (GPM) [17].

A. GpmLab: Recording application

As it has been explained during the presentation of the method, the recording procedure has one prerequisite, that is the definition of the path to be followed by the tester. This path consists of a list of positions, that will be followed one after the other, and will serve as checkpoints for the tester. The minimum required fields for a checkpoint are the 2D coordinates of its position. Apart from these, the checkpoint can be enriched by adding other fields that describe with more detail a position, such as altitude, floor, room, etc. This predefined path is the only input required by the GpmLab application. An example of a predefined path is shown in Figure 3.

Before proceeding to the recording, the tester should clearly mark in the physical world the predefined path with clear landmarks or added signs. If it is not possible to precisely determine the checkpoints' positions with landmarks, the tester should measure the distances in the reference maps and place in the physical world signs for every checkpoint, such as stickers on the ground with clear numbered labels.

Fig. 3. An example of a predefined path.

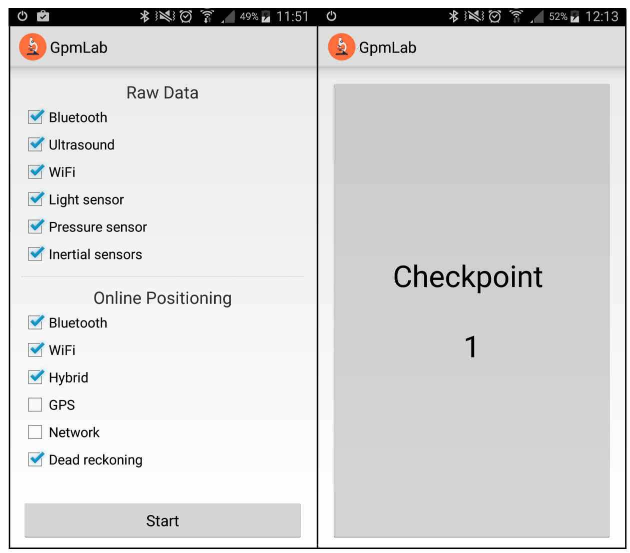

Fig. 4. The two main screens of the recording application GpmLab.

The GpmLab application has a very simple and straightforward interface with two simple screens (Figure 4). Initially, the user has to select the kind of signal readings that he wishes to record, such as Bluetooth, Ultrasound, WiFi, sensor data, etc., as seen in the left screen of Figure 4. Apart from the raw data, recording the estimations provided by online positioning algorithms is also possible. It is noted though that in the context of this work, the online positioning does not participate in the proposed methodology. As mentioned earlier, recording the estimations of the online position providers could potentially be useful for verifying the consistency between the online and the offline providers. After selecting the information to be recorded, by pressing start, the second screen appears (right side of Figure 4), which facilitates the synchronisation of the the raw data recording with the ground truth gathering. In this screen there is a single big button indicating the number of the next checkpoint to be reached.

In order to start recording the data, the tester must be placed over the first checkpoint. Then, he should indicate it to the application by clicking the unique button of the interface and start walking towards the second checkpoint. The system logs the time the user began the path and starts recording the raw signals of all selected sources, marking them with the timestamp of each reception. The application will continuously log these data throughout the whole path, and not only at the checkpoints. Every time the tester steps over a checkpoint, he should indicate it again by pressing the unique button (on which the number of the checkpoint will be indicated), in order to mark the checkpoint with a timestamp, until he reaches the end of the path. At the moment of arriving at the last checkpoint, the application will stop gathering information.

B. *GpmStudio: Offline evaluation and tunning platform*

The GpmStudio platform, is an offline positioning tool suite, that utilizes the recordings obtained by the application GpmLab. The interface of GpmStudio is shown in Figure 5.

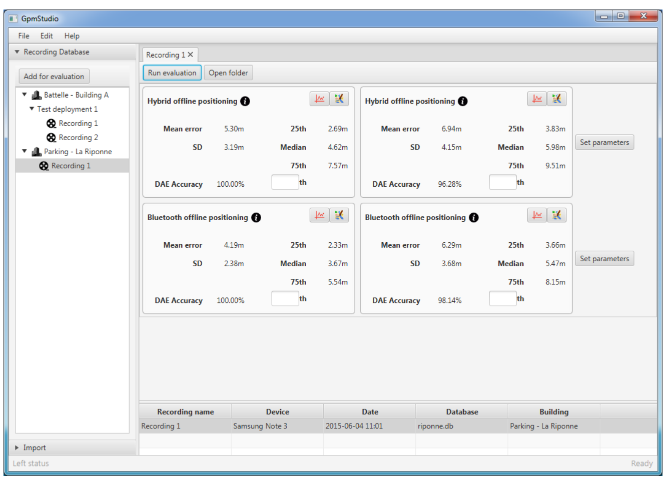

Fig. 5. The interface of the GpmStudio tool suite.

On the left panel, there is a menu where all available recordings are listed in a structured way. The tester can select the recording which he wishes to process. According to the source type of the recorded raw data, the corresponding offline position providers that can utilize these data are used, running their offline positioning algorithms with their default parameter settings. At the same time, the evaluation of the performance of each provider is done, as the spatio-temporal ground truth information for each recording is available. The default interpolation method is the *interpolation per estimation update* (presented in Section III-C1), though the user can choose the desired one.

The results of the evaluations are presented in the big central panel of the interface. The evaluation of each position provider is included in a box, highlighting the main statistical metrics of the error (mean, standard deviation, 25th, 50th and 75th percentile). Moreover, links to the visualization of the estimated path on a map (Figure 6), and plotting options of the cumulative distribution function of the estimation's error (Figure 7) are offered.

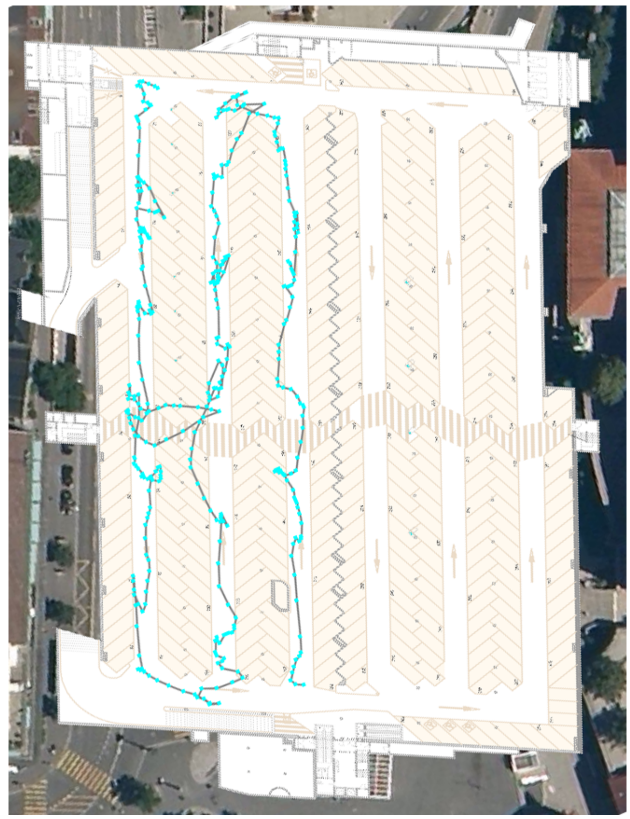

Fig. 6. The visualisation of an estimated path, produced by the offline positioning functionality of the GpmStudio.

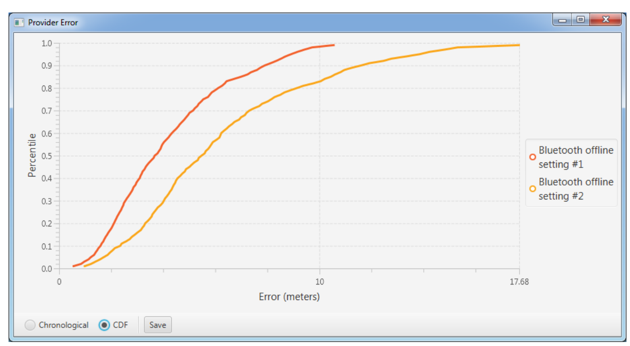

Fig. 7. The CDF of the positioning error with two different parameter tunings, produced by GpmStudio.

The presented scenario intends to exemplify the features of the evaluation platform. For clarity, we specify that a broad area (120 * 40 m) of an underground parking was used as the test environment (left half part of Figures 3 and 6). In this area, 40 BLE beacons were placed at a rhombus grid pattern, and a filtered weighted centroid algorithm was used [18]. As explained in Section III-B, the distance between the checkpoints is a designer's choice. In this example, the distance between two consecutive checkpoints is approximately 8-10m.

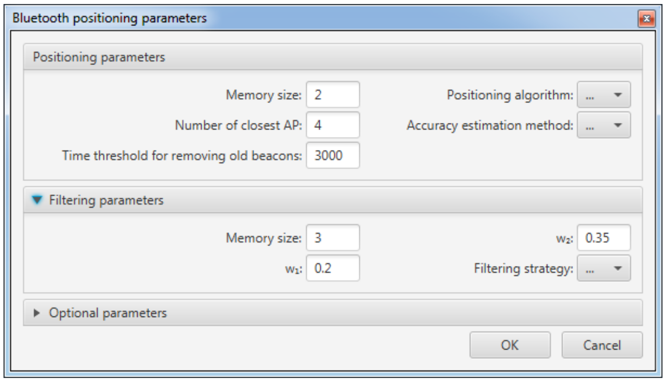

Fig. 8. A parameter selection window of the GpmStudio tool suite.

1) Manual parameter tuning: So far, the evaluating capabilities of GpmStudio platform have been presented. The main power of this platform though lays on its tuning features. The tester who has an insight of the system under test, is offered the possibility to adjust all tunable parameters of each positioning algorithm, re-evaluate each time, and compare the change in performance. This difference in performance can be evaluated by comparing the statistical metrics, comparing the respective CDF's, and by examining the estimated paths on a map. Figure 7 exemplifies the difference in performance between two different parameter settings of the same algorithm, while in Figure 5 the corresponding statistical metrics are reported.

In Figure 8, an example of a parameter selection window that is offered to the tester is shown. Using a tool like this, the designer (or the tester) of a positioning algorithm, has the possibility of testing the immediate impact of each parameter on his algorithm's performance.

This tuning option has proven to significantly facilitate the deployment procedure for our team. Following this methodology in cases where we had a new deployment at a new environment, has simplified the task of tuning the positioning system. An experienced person was able to quickly tune the algorithm, adjusting its parameters accordingly.

2) Automatic parameter optimization: The GpmStudio tool suite contains a module for automatic parameter optimization. The goal of this module is to find those parameter settings that optimize the performance of positioning algorithms, over the recorded data. To proceed with the optimization, the tester needs to choose four elements:

i. the recorded data
ii. the objective function $f(P)$
iii. the optimization algorithm
iv. the search space

The user can select (i) the recorded data of the path that he wishes to process, selecting from the available recorded files. The selection of (ii) the objective function and (iii) the optimization algorithm is done by choosing from a list of available options. Defining (iv) the search space is the most delicate action to be performed. The tester should select the dimensions of the search space. Each parameter chosen to be included in the optimization problem, is one of the dimensions of the search space. For each dimension, the tester needs to define its set of valid values. Assuming numerical parameters, the tester should define the minimum and maximum value of the parameter, as well the granularity of the search, that is the distance between two consecutive values of the parameter.

In Figure 9, an exemplification of the form of an optimization problem is given. The search space of this problem has two dimensions, and the objective function is the mean error. The first parameter is the number of the closest access points used for ranging methods. The selected range of values is $[1, 10]$, with a step of 1 which is an intuitive selection for an integer parameter. The second parameter is the number of the latest RSSI receptions from each access point that are used in the calculations (the memory size of a list), which similarly has a range of valid values $[1, 10]$, with a step of 1. The best value of the objective function (5.7 meters) is achieved at the point corresponding to 4 access points, and to a memory size of 2.

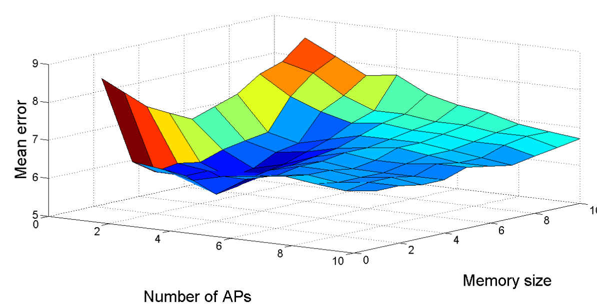

Fig. 9. Mean positional error for 100 combinations of parameter settings, exemplifying the perspective of an optimization problem.

Evidently, the problem presented above is a simple case, with a very limited search space. A full search algorithm can quickly evaluate all the points of the search space (100 points in this case), and produce the data for the plot of Figure 9. For practical needs of our group, the full search algorithm has been used, providing results in most cases within a time-frame of a few minutes, or in rare occasions where many parameters were used, a few hours.

The complexity of the problem increases, as more dimensions are added, or with a high granularity of non integer parameters. By imagining a hybrid system, utilizing information from many sources, that each has several parameters, while adding also filtering or other processes with their own parameters, it becomes evident how enormous the search space can become. For cases like this, meta-heuristic algorithms can be useful. When the search of the whole search space is not possible, local minima could provide a near optimal solution. In this context, the hill climbing algorithm (which is a local search method) has been used in this optimization module.

The hill climbing algorithm requires a starting point in the search space. Naturally, the default values of all parameters are used to define the starting point. The intuition behind this selection is that a local minimum close the default values is more likely to be near-optimal, than other local minima close to a random starting point in the search space. Nevertheless, the tester has the freedom to define the starting point.

The outcome of the optimization is a list with the best solutions, sorted by their objective value. Each solution, is a point in the search space, thus it is a set of values for the parameters that were chosen to participate in the optimization. Since many points in the search space might have very similar evaluation scores, the tester can observe the values of the parameters of each point among the best ones, and select. Selecting blindly the parameter setting that gives the best evaluation score, based on a single recording, might be quite misleading, as there is the danger of overfitting. A single recording might contain particularities that do not reflect all use cases. Factors that can play a role are: the speed of the tester while he was recording the path, the noise level of the environment during the recording, the device used, the chosen path (which parts of the area of the deployment the path covers and how equally it covers them), etc. To address these issues up to a certain extent, we are currently designing a multi-objective optimization approach, that will allow the combined usage of several recordings. This multi-objective optimization approach is further discussed in the future work section (Section VI).

V. Usage Overview and Discussion

In this work, we presented a straightforward methodology for evaluating and tuning positioning systems. The methodology includes a dynamic way of recording the ground truth information and the received raw signals. The subsequent offline phase offers the possibility of tuning the positioning algorithm in order to improve the system's performance. The tuning can take place either manually by an expert, or in an automatic way, by an optimization module of the offline evaluation platform.

The proposed methodology has several applications. A usual scenario that a positioning expert has to face is deploying at a new environment, in which the need to calibrate the positioning algorithm appears. A simple approach is using a test-and-set method. With this first approach, for every different tuning, the tester has to revisit the deployment area, and either visually evaluate the performance or actually record each time the position estimates and the ground truth. This tedious task is completely overcome with the proposed methodology. In this way, the time needed at the deployment area is minimized. Also, the consistency of the test with different parameter settings is guaranteed, as the test utilises the same data.

Moreover, each recording of a path is an addition enriching the bank of collected data. Increasing the number of collected recorded paths, offers the possibility of making consistent comparisons through time, over a variety of environments. The possibility of crowd-sourcing recorded paths, creates the perspective of a testbed based on real data.

Furthermore, having recorded all the data needed for an offline positioning algorithm, by following a precisely described methodology, can significantly facilitate the reproducibility of experiments for indoor positioning publications. Making the recorded data publicly available (raw data and ground truth), and mentioning the exact methodology of collecting these data, removes any ambiguity over the presented experiments of a publication. In this way, issues such as the ones presented in [1], and discussed earlier in this work, can be efficiently overcome. As emphasized at the published conclusions from the IPSN 2014 competition [5], in order to ensure that all systems are evaluated under identical environmental conditions (i.e., number of people in the room, interference etc.), all systems should be simultaneously evaluated at a given evaluation point. Generally, using results of another work as a baseline, requires a consistent comparison, with similar conditions. Using the same recorded data, guarantees the consistency of the comparison.

VI. Future Directions

The main next direction of this work is the improvement of the optimization module of the offline evaluation platform, which provides tuning suggestions. This next step, consists of introducing the possibility of using multi-objective optimization techniques. Multi-objective optimization handles mathematical optimization problems involving more than one objective function to be optimized simultaneously. This multitude of objectives, could serve different kinds of multiplicities.

It is generally accepted [5] that it is very hard to capture the effectiveness of an indoor localization algorithm with a single metric. Thus, a first kind of multiplicity is the use of several evaluation metrics, as different objectives to be satisfied. This approach will provide a more complete evaluation of the performance of systems under test. Using a unique metric might be not representative of a system's performance, and could be unfair when used for comparisons. A more holistic evaluation would be an appealing field of research.

Multi-objective optimization could also be used to tackle another issue that was discussed in this work: the risk of overfitting lurks when selecting parameters based on a single recording. Combining several recordings (with each recording being another objective) at the same deployment could minimize this danger. These recordings could concern the same or different predefined paths, different speeds of the tester, different devices, and so on, or simply repeated recordings with the same characteristics, to minimize the effects of punctual issues in a single recording. Extracting an estimation about the optimal tuning of the algorithm, by utilizing such a variety of inputs would strengthen the robustness of the parameter setting suggestion.

Moreover, apart from the goal of fine-tuning a system for a specific deployment, other goals could be served by a multi-objective optimization approach. As a collection of recordings gets enriched with data from several environments, it earns in variety, and can be used as a database for finding the optimal default setting of a system. If requested to provide a positioning system which 'will work everywhere', without requiring any calibration, what's the best way that the provider of this solution can test if the goal is achieved? Evaluating the system using a big database, containing recordings from a plethora of deployments could be very useful. Therefore, multi-objective optimization could provide tuning suggestions so that the system manages a good performance in a multitude of environments.

It is clear that the possibility of crowd-sourcing recorded paths, creates the perspective of a testbed based on real data. We intend to investigate this perspective. In this direction, strictly defining the procedures followed and the format used for storing the data becomes an absolute necessity. This task needs to take place with special caution, so that the resulting protocol will be able to serve as many techniques, technologies and evaluation objectives as possible.

Lastly, there are some interesting additional tests that were highlighted by the anonymous reviewers of this work, which we are willing to conduct in the future. An interesting test is to characterize the performance of recordings with different densities of checkpoints, in order to experimentally observe and characterize the impact of the checkpoint density to the evaluation of a SUT. Additionally, by evaluating several positioning technologies that have diverse characteristics (indoor or outdoor, with different frequencies of position estimations, etc.), we could observe the results of the two interpolation methods discussed in this work, or even compare them with other methodologies, such as the one used at the IPIN competition.

Acknowledgements

The authors would like to thank the anonymous reviewers of this work for their very fruitful feedback, and their spot-on remarks.

References

[1] S. Adler, S. Schmitt, K. Wolter, and M. Kyas, "A survey of experimental evaluation in indoor localization research," in *Indoor Positioning and Indoor Navigation (IPIN), 2015 International Conference on*, Oct 2015, pp. 1–10.

[2] T. Schwartz, C. Stahl, A. Mahr, C. Muller, and B. Saliba, "Using natural footstep-accurate traces for indoor positioning evaluation," in *Indoor Positioning and Indoor Navigation (IPIN), 2012 International Conference on*, Nov 2012, pp. 1–8.

[3] M. de la Osa C., G. Anagnostopoulos, N. Togneri, M. Deriaz, and D. Konstantas, "Positioning evaluation and ground truth definition for real life use cases," in *Indoor Positioning and Indoor Navigation (IPIN), 2016 International Conference on*, Oct 2016, pp. 1–7.

[4] I. Sharp, K. Yu, and T. Sathyan, "Positional accuracy measurement and error modeling for mobile tracking," *IEEE Transactions on Mobile Computing*, vol. 11, no. 6, pp. 1021–1032, June 2012.

[5] D. Lymberopoulos, J. Liu, X. Yang, R. R. Choudhury, V. Handziski, and S. Sen, "A realistic evaluation and comparison of indoor location technologies: Experiences and lessons learned," in *The 14th ACM/IEEE Conference on Information Processing in Sensor Networks (IPSN '15)*. ACM Association for Computing Machinery, April 2015. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=241638

[6] P. Barsocchi, S. Chessa, F. Furfari, and F. Potortì, "Evaluating AAL solutions through competitive benchmarking: the localization competition," *IEEE Pervasive Computing Magazine*, vol. 12, no. 4, pp. 72–79, Oct.–Dec. 2013.

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

[8] T. Van Haute, E. De Poorter, J. Rossey, I. Moerman, V. Handziski, A. Behboodi, F. Lemic, A. Wolisz, N. Wiström, T. Voigt, P. Crombez, P. Verhoeve, and J. J. De Las Heras, "The EVARILOS Benchmarking Handbook: Evaluation of RF-based Indoor Localization Solutions," in *Proc. of 2nd International Workshop on Measurement-based Experimental Research, Methodology and Tools (MERMAT'13)*, May 2013, ConferenceProceedings. [Online]. Available: http://www.tkn.tu-berlin.de/fileadmin/fg112/Papers/2013/Van_Haute13the_evarilos_benchmarking.pdf

[9] D. Becker, F. Thiele, O. Sawade, and I. Radusch, "Cost-effective camera based ground truth for indoor localization," in *Advanced Intelligent Mechatronics (AIM), 2015 IEEE International Conference on*, July 2015, pp. 885–890.

[10] S. Das, C. Gleason, S. Shen, S. Goddard, and L. Perez, "2d tracking performance evaluation using the cricket location-support system," in *Electro Information Technology, 2005 IEEE International Conference on*, May 2005, pp. 6 pp.–6.

[11] 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*, Nov 2012, pp. 1–8.

[12] S. Adler, S. Schmitt, H. Will, T. Hillebrandt, and M. Kyas, "Virtual testbed for indoor localization," in *Indoor Positioning and Indoor Navigation (IPIN), 2013 International Conference on*, Oct 2013, pp. 1–8.

[13] S. Adler, S. Schmitt, and M. Kyas, "Path loss and multipath effects in a real world indoor localization scenario," in *Positioning, Navigation and Communication (WPNC), 2014 11th Workshop on*, March 2014, pp. 1–7.

[14] S. Adler, S. Schmitt, Y. Yang, Y. Zhao, and M. Kyas, "Experimental evaluation of indoor localization algorithms," in *Indoor Positioning and Indoor Navigation (IPIN), 2014 International Conference on*, Oct 2014, pp. 291–299.

[15] S. Adler, S. Schmitt, and M. Kyas, "Experimental evaluation of the spatial error distribution of indoor localization algorithms," in *Ubiquitous Positioning Indoor Navigation and Location Based Service (UPINLBS), 2014*, Nov 2014, pp. 44–53.

[16] T. Pulkkinen and J. Verwijnen, "Evaluating indoor positioning errors," in *Information and Communication Technology Convergence (ICTC), 2015 International Conference on*, Oct 2015, pp. 167–169.

[17] A. Bekkelien and M. Deriaz, "Harmonization of position providers," in *2012 International Conference on Indoor Positioning and Indoor Navigation, IPIN 2012, Sydney, Australia, November 13-15, 2012*, 2012, pp. 1–7. [Online]. Available: http://dx.doi.org/10.1109/IPIN.2012.6418855

[18] G. G. Anagnostopoulos and M. Deriaz, "Accuracy enhancements in indoor localization with the weighted average technique," in *SENSORCOMM 2014*, 2014, pp. 112 – 116.