Evaluating AAL Solutions Through Competitive Benchmarking: The Localization Competition

Paolo Barsocchi, Stefano Chessa, Francesco Furfari, and Francesco Potortì

Abstract— **Evaluation of Ambient Assisted Living (AAL) systems is particularly challenging due to the complexity of such systems and to the variety of solutions adopted and services offered. This problem is related to the evaluation of pervasive and ubiquitous systems that has been the focus of many researchers in the recent years and that still awaits for solutions. On the other hand, analyzing and comparing AAL solutions is paramount for the assessment of the research results in this area. EvAAL (Evaluating AAL Systems Through Competitive Benchmarking) is a recently established international competition that aims to address this problem in order to let benchmarking and comparison methodologies of AAL systems emerge from experience. This work describes the first EvAAL competition which was devoted to localization and tracking; proposed evaluation criteria, benchmarks, and achieved results are reported.**

I. Introduction

Ambient Assisted Living (AAL) [1], an innovation funding program issued by the European Commission, aims at improving the quality of life of the elderly, by increasing their autonomy, assisting them in their daily activities, and by enabling them to feel secure, protected and supported. AAL spaces are physical places, where users live or work, that integrate a number of Ambient Intelligence (AmI) technologies [2], ranging from environmental sensors and actuators to services and intelligence that supports the integration of such services. A typical AAL system involves a number of activities that include sensing, acting, reasoning, interacting etc. These activities are generally implemented by a number of software components (such as context managers, profile and service managers, reasoners, user interfaces, security managers, etc) which, in turn, are incorporated into a number of devices spread in the environment, such as sensors and actuators, gateways, appliances, domotic devices, communication devices, smartphones. As a result AAL systems are typically complex distributed systems that make use of middleware platforms to support communication and integration of the different components.

In this scenario, recognized evaluation methodologies are essential to compare different AAL solutions. AAL systems need such methodologies to enable researchers to objectively compare new state of art contributions. Evaluating AAL solutions is difficult since these systems are complex, therefore the approaches tend to be subjective, piecemeal, or both. To ensure the validity and usability of the proposed systems researchers must reach consensus on a set of standard evaluation methods for AAL systems, otherwise the scientific advantages on the state of art will remain unclear. However, as a consequence of their intrinsic complexity, full AAL systems are hardly comparable among themselves, and, in fact, evaluation and comparison of such systems is a challenging problem that is still far from being solved [3].

Driven by this objective we organized an annual international competition promoted by the AALOA association [4]. It aims at advancing the state of the art in the evaluation and comparison of AAL platforms and architectures, by creating an environment in which the researchers, students and industries can compare their solutions and exchange ideas, and where the comparison of AAL systems may become feasible. In particular, EvAAL adopts a step by step approach, by dividing the problem into smaller pieces. In an initial phase it promotes competitions on specific, small scale topics in order to create publicly accessible data sets and to evolve benchmarks and evaluation methodologies. In a second phase, when methodologies and tools of EvAAL become more mature, complex services and even complete systems can be evaluated and compared (figure 1). One major specific topic to be explored is indoor localization, since it is a key component of many AAL services. Recent years have witnessed an increasing attention on location-based services and applications. In most cases, however, location information is limited by the accessibility to Global Navigation Satellite Systems (GNSS), largely unavailable for indoor environments. The main scope of this paper is to describe the criteria used in EvAAL for identifying the best indoor localization system from the point of view of Ambient Assisted Living (AAL) applications.

II. Purpose of EvAAL

EvAAL aims at contributing to AAL disciplines in the same way as other competitions have contributed to their respective areas. Under this respect the idea of initiating EvAAL was inspired by successful competitions such as the Trading Agent Competition [5] and DARPA Grand Challenge [6]. Beyond supporting the growth of the AAL community, the main technical objectives of the competitions organized by EvAAL are to:

- Enable the comparison of different AAL solutions
- Experiment with benchmarking and evaluation methods
- Identify relevant AAL problems, requirements and issues
- Identification of new, original solutions for AAL

EVAAL aims at enabling the comparison of different AAL solutions, by establishing suitable benchmarks and evaluation metrics that will be progressively refined and improved with time. In particular, EvAAL focuses not only on comparison

P. Barsocchi, S. Chessa, F. Furfari and F. Potortì are with the ISTI-CNR, Pisa Research Area, Via G.Moruzzi 1, 56124 Pisa, Italy.

S. Chessa is with the Computer Science Department, University of Pisa, Largo B. Pontecorvo 3, 56127 Pisa, Italy.

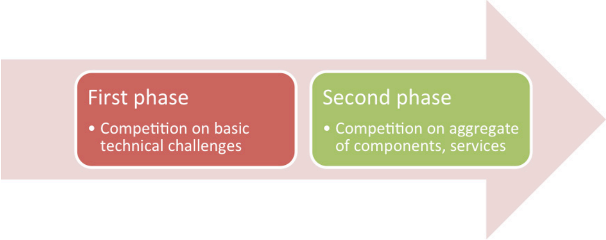

First phase
• Competition on basic technical challenges
Second phase
• Competition on aggregate of components, services

Fig. 1. The EvAAL focus shift.

of hard data such as accuracy of positioning and system reliability, but also on soft data like compatibility with existing standards, deployment effort and user acceptance.

EvAAL's objective is to fill the gap by tackling the evaluation issues and by offering researchers an arena where to try, test and experiment not only AAL solutions but also benchmarks and evaluation methods. To this purpose, EvAAL is open to all issues related to the test environment (living laboratories vs. into the wild), to the benchmarking (automatic vs. based on users' evaluations), to the tools supporting the competition etc. The intended outcome is a toolkit of techniques from which system developers can draw. Making these techniques open, available, and easy to use will enable comparative evaluation between similar components across systems and, in the end, of whole AAL systems.

In front of this grand objective, EvAAL recognizes that facing the full complexity of the evaluation of AAL systems is not feasible with the current knowledge and, in fact, a clear vision on the methods for the evaluation of full, complex AAL systems is still to be reached. For this reason EvAAL initially focuses on specific technological challenges related to AAL, and it uses the results achieved from the initial competitions to set up tools and methodologies that will support the evaluation of aggregate of components, subsystems, or even entire systems in a subsequent phase (figure 1). In order to keep the pace with technological evolution, the decision about the yearly topics of the competition is reached by organizing a public discussion, which is stimulated and initiated by a Call for Ideas. The call is published yearly soon after the previous competition is completed. Its aim is to collect ideas and suggestion for the next year competition, and to stimulate the involvement of other researchers, institutions and industries in the decision processes of EvAAL. With this view on the long term, the first EvAAL edition consisted of a competition on localization and tracking. This competition is the subject of the rest of this paper.

III. The 2011 EvAAL competition: Localization and Tracking

The first EvAAL competition was held in the CIAMI living lab in Valencia [7], which is an open space environment composed by a kitchen, a dining room, a bedroom and a bathroom, as shown in the next figures starting from figure 3. In order to define the basic specifications for a localization system we make reference to the *personal activity management scenario* in the AAL road map [1]. In this particular scenario, the main point is that the home knows what the user is doing; where the user is, if he is standing or sitting, whether appliances are in use, and what object, if any, the user is handling.

TABLE I
Metrics and their weights.

Metric	Weight
Accuracy	5
Availability	3
Installation complexity	4
User acceptance	4
Integrability in AAL systems	2

A. Choosing the metrics

Metrics for the competition have been chosen in accordance with the above scenario. Five metrics were identified. Two of them are objectively measurable (*hard*) quantities: *accuracy* and *availability*. They are based on the assumption that each localization system provides, each half a second, the coordinates of an *actor*'s position. Only one actor was considered for the first edition, without any other intervening person in the environment. Maybe the next edition will ask the systems to locate more than one actor, or to locate one actor among several people in the environment.

Accuracy is the classical measurement of the goodness of a localization system, based on samples of the distance between the point where the system locates the user and the point where the user really is. Availability is a measure of how well the system performs at providing regularly spaced measurements: this is especially significant for experimental or prototypal systems.

Besides *hard* quantities, some *soft* ones were considered, namely *installation complexity*, *user acceptance* and *integrability in AAL systems*. While scores for the hard quantities were obtained through an automated process, scores for the last two quantities were obtained from a jury decision, and installation complexity was simply linked to the needed installation time. The weights of the metrics were established as shown in table III-A.

The rationale behind the choice of these metrics and their relative weights is multi-faceted. First of all, we wanted *accuracy* to be the relatively most important of metrics, because we are assessing the performance of a localization and tracking system. Yet we did not want it to be prominent with respect to the rest of metrics. AAL systems often do not need a high precision, and giving accuracy too much importance would have lowered the significance of scores with respect to real-life systems. At the same time, we needed availability to be important, because an unresponsive system can be as difficult to manage as an inaccurate one. The weight of the *hard* metrics together should not be more than the weight of metrics related to interaction with the main stakeholders for an AAL system: system integrators, installers and final users.

We would have very much liked to include cost as one additional metric, but after debating possible ways of doing it, we gave up. While cost is essential for a finished product, it is very difficult to predict what the cost will be when the system is at the prototype or experimental stage, mostly because cost of devices mainly depends on how many of them are produced, which, in turn, depends on developments that are outside of the AAL field and which are largely unpredictable.

The following subsections illustrate the choices behind each metric in detail.

B. Hard metrics: accuracy and availability

Accuracy is measured by taking the data sent by the competing system and comparing it with reference data. *Availability* is a measure of how regularly the systems produce the expected data.

1) Choosing a reference system: The first issue is to define a references system and the extent to which the competing systems should conform to it. Accuracy is defined as some statistics on the error, which in turn is defined as the distance between a reference and a data sample given by the system under measure. We should then decide to which resolution this error should be measured and the time rate at which samples should be produced.

By making reference to the already mentioned *personal activity management scenario*, where persons are localized during their home activities, a resolution finer than 30 cm looks useless: this is about the size of a foot, is less than the body diameter and is definitely less than the extent to which a stretched arm or leg can go. With similar reasoning, a time resolution of half a second was chosen, because in the considered scenario no high-speed activity is requested or needs to be evaluated.

Our reference system should be then *accurate* to about 30 cm in space an 0.5 s in time. It should be *transparent*, that is, its outcomes should be easy to verify. It should be *realistic*, that is, should be applicable to a real person. And it should allow measurements to be *repeatable* for equity among competitors, for checking and for debugging.

Transparency as defined above depends mostly on implementation. As an example, a proprietary system without any known information on its inner working is considered minimally transparent.

Repeatability is implemented by requiring that the actor moves along a path that is drawn on the floor, and to move according to a predefined *script*, that is, put his feet on precise spots the instant a clock chimes. This way, we obtain a path that is repeatable in time as well as in space.

One possibility we considered was to use a very accurate localization system to be taken as a reference, which takes measurements of the position of an actor moving along a predefined path. If the system is reliable and accurate enough, that would be a nice choice because it is easily automated. The main drawbacks are cost and possible lack of transparency. In order to make repeatable measurements, the actor needs to move along a path that is drawn on the floor, and to move according to a predefined *script*, that is, put his feet on precise spots the instant a clock chimes. This makes the path repeatable in time as well as in space.

One other possibility was to avoid managing an actor and use a small robot instead, which would move along one or more programmed paths. The path is reproducible both in space and time, with accuracy depending on the robot movement accuracy. The problems here are to find a robot reliable and accurate enough, the work needed to program it and the fact that, from the point of view of the competing systems, a robot may be quite different from a person. Differences may arise because of size, reflection properties of light or electromagnetic waves, infrared emission and possibly others.

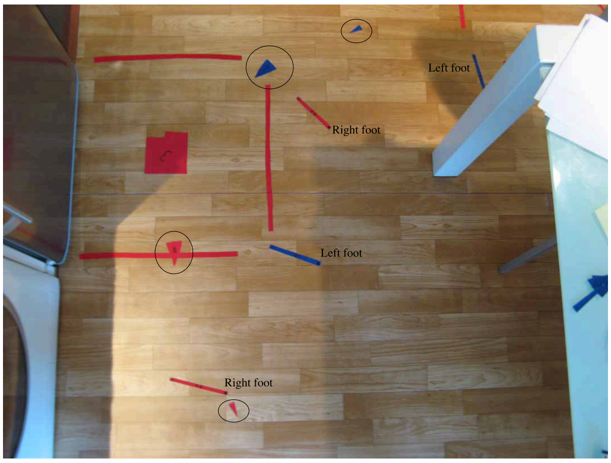


Fig. 2. The reference localization system: the blue marks are related to the left foot while the red ones are related to the right foot. The triangles (highlighted with a circle in the figure) are related to the path followed by the actor during the AoI test.

2) The chosen reference system: In the end, we decided for a trained actor following a predefined path, basing the reproducibility of the path on the actor's movement regularity and using as a reference the measured coordinates of marks put on the floor. As shown in figure 2, the Living Lab's floor was covered with red and blue marks (for the right and left foot, respectively) that show where the actor had to step on. A software metronome connected with the data collection system gave the time. Once a second, a chime indicated the time that the actor should put its foot on the mark. The reference position at each chime (whole-second positions) was taken as the midpoint between the feet, while the reference position for the half-second instants between chimes was taken as the midpoint between two subsequent whole-second positions.

An example of the checks that we made for testing this method can be found in the short movie at http://evaal.aaloa.org/2011-competition/reference-system, together with a high-resolution picture of the floor where the details of the marks that we set are clearly visible.

Four predefined routes were defined: three *paths* and one route traversing five *AoI* (Areas of Interest). The competitors were expected to identify the moving actor's position along the three paths in real time (figure 4). The paths were not previously disclosed to competitors. The first path was 36 s long, the second path 52, and the last one 48. Each path included some waiting points, where the actor stood still for 5 seconds. In addition to the three paths, each team was expected to identify the AoI—along the fourth route—where the actor was, or to state that the actor was outside of all AoIs.

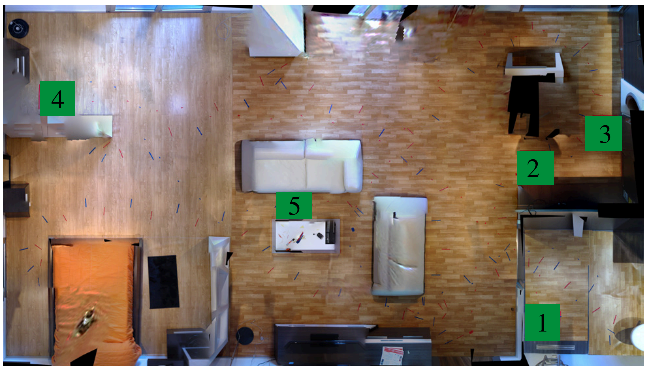


Fig. 3. The Areas of Interest deployed in the Living Lab.

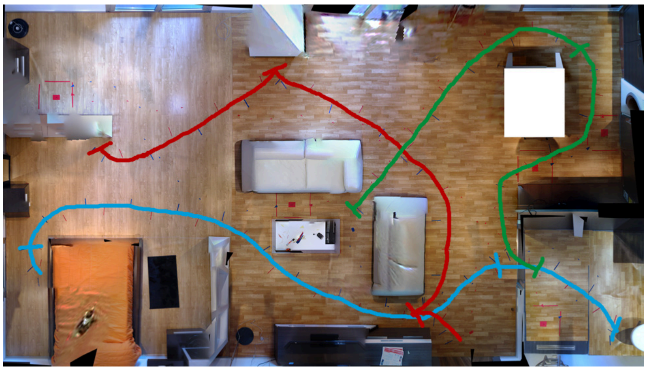


Fig. 4. The three different paths: path 1 (green line), path 2 (blue line), and path 3 (red line).

Each AoI was a square of sides 50 cm. The actor moved along the predefined route, following the triangle marks highlighted in figure 2, stopping in each AoI for 30 seconds. The total duration of the route was about 5 minutes.

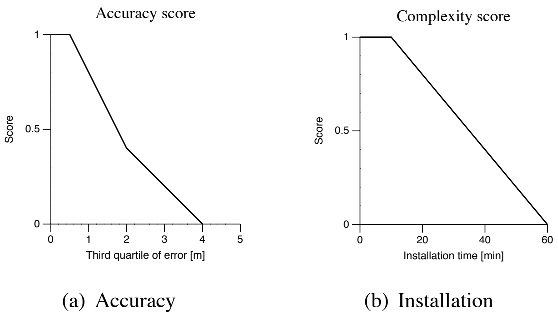
3) Availability: Availability measures the capacity of the systems to produce fresh data continuously. As such, it is simply computed as the ratio between the number of received samples and the number of expected samples (one every half a second). The algorithm used allows for a jitter not wider than ±250 ms. Samples exhibiting a higher jitter are discarded (if in advance) or start a new time base (if delayed).

The same algorithm applies for the paths and for the AoI route, and the value of availability is linearly scaled into a score ranging from 0 to 10. The final score is the mean of the scores obtained for each benchmark.

4) Accuracy: Accuracy is a measure of how good the system is at doing its main work: telling where the actor is as precisely as possible.

As far as the AoIs are concerned, the score is the ratio of correct answers given by the competing system to the total answers given.

As far as the three paths are concerned, we define the *error* as the distance between the real point where the actor is and the coordinates estimated by the competing system. Distances are computed in two dimensions. Error is computed for each answer given by the competing system. The error series should be reduced to a scalar score, and the literature is rich in methods to reach this result.



(a) Accuracy

(b) Installation

Fig. 5. Accuracy score as function of the third quartile of error and Installation complexity score as a function of installation time.

By analyzing 195 papers of the first edition of the Indoor Positioning and Indoor Navigation (IPIN 2010) Conference, we verified that 115 works describe real or simulated systems that are amenable to being evaluated by measuring some kind of metrics. The metrics taken into account in these works are visual path comparison, usually as a graph that shows the real and the estimated path (32% of cases), mean error (31%), cumulative distribution function CDF (20%), a quantile value (11%) and finally error variance (5%).

Since we are aiming at removing subjectivity as much as possible, we discarded both CDF and visual path comparison. We chose the third quartile (75^{th} percentile) as the base statistic to rank the accuracies of the competitors' localization systems.

The score for the three paths is a piecewise-linear function of the third quartile of the error, with a flat maximum between 0 and 50 cm, a null minimum at 4 m and above, and a corner at 2 m where the score is 40% of maximum, as shown in figure 5. The flat maximum ensures that the score is not impaired by inaccuracies smaller than 50 cm; the long tail up to 4 m is intended to discriminate among competing systems that give a completely wrong or random output and those that, while inaccurate, are able to give an idea of the area where the actor is. The overall score is evaluated as the mean between the AoI and the paths scores.

C. Soft metrics

Soft metrics are those that are not measured using a mechanical method, but are the outcome of a jury decision.

1) Installation complexity: Of the soft metrics, this is the simplest: we just consider the installation time used by the competitors, multiplied by the number of people engaged in the installation. The installation time is defined as the time elapsing from when the competitors enter the living lab to the moment they declare the installation completed—no further operations/configurations of the system are allowed afterwards. The score is a piecewise-linear function of the installation time, with no penalty for times less than 10 minutes, as shown in figure 5.

2) User acceptance: It expresses how much the localization system is invasive in the user's daily life and thereby the impact perceived by the user. This criterion is qualitative and was evaluated by taking the mean of scores given by individual evaluation committee members. A predefined list of questions were posed to the competing team.

3) Integrability in AAL systems: The score relative to the integrability was chosen by the jury member after discussion and reaching consensus. It was based on the following predefined criteria:

2 points: availability of libraries for integration;
2 points: use of open source libraries;
2 points: use of standards;
2 points: tools for testing/monitoring the system;
1 point: sample applications;
1 point: documentation.

IV. Organization and results

Of the many practical details involved in organizing a public competition, we are going to touch on those that are relevant to the scientific soundness of the procedure and consequently of the results.

We started by issuing a call for competition and by peer-reviewing the ten submissions we got. The submission were not required to be necessarily novel in concept, but to provide detail enough that the reviewers could judge whether the proposed system could work and was useful for AAL purposes. Six out of the ten proposed localization systems were accepted, plus one that was accepted out of contest because details were not disclosed within the allowed deadline because of a pending patenting procedure.

The competitors were invited to the living lab at a fixed hours, and each of them was given a three-hours time slot. We managed three competitors per day. Once a competitor arrived, a series of steps were followed, of which the most significant were the following.

1) The floor of the Living Lab was covered with carpets, so that the competitor could not see the paths during the system deployment.
2) Position of devices deployed in the Living Lab was measured.
3) Installation time was measured in order to assess the *installation complexity* score.
4) Integration between the competitor's software and the evaluation tool was performed.
5) The carpets are removed and the actor entered the living lab. Any other person had to exit at this time.
6) The evaluation phase for assessing *accuracy* and *availability* was videorecorded, while the position indicated by the system was shown in real time on a display.
7) Interviews were done for assessing the scores for *integrability* and *user acceptance*.

Eventually, papers describing the competing systems and the competition setup were presented during the EvAAL workshop, as part of the AAL Forum 2011 in Lecce (IT). This meeting gave the competitors the opportunity to meet together and exchange ideas. It was also the setting where the winners were announced and the prizes awarded. The AAL Forum was chosen as hosting conference for the EvAAL Workshop because it is a major, annual conference of the Ambient Assisted Living Joint Programme, it has a large audience interested in AAL, and it gives a considerable attention to the most recent EU initiatives.

The six teams competing in Valencia at the CIAMI Living lab were n-core Polaris (University of Salamanca) [8], AIT (Austrian Institute of Technology) [9], iLoc (Stuttgart University of Applied Sciences and iHomeLab at Lucerne University of Applied Sciences) [10], OwlPS (University of Franche-Comt) [11], GEDES-UGR (University of Granada) [12], and SNTUmicro (Sevastopol National Technical University) [13]. Table II summarizes the scores.

Figure 6 shows the predefined path and the estimated user position for the best accuracy performers: the third quartile of error was 62 cm for AIT in the first path, 81 cm and 82 cm for iLoc in the second and the third paths, respectively.

The timestamped localization data logged by the competing systems, which were used to compute the accuracy and reliability of the systems, are publicly available on the EvAAL web site [14].

The web site also reports the timestamped paths followed by the actor and the data that some of the competitors were able to provide from the internals of their systems, together with associated metadata that describe them and the system itself. We are confident that these will be useful for other researchers and practitioners in the indoor and localization tracking field.

V. Conclusions

Feedback from the competitors was encouraging, and we are currently organizing the second edition of the EvAAL competition. Year 2012 will see two tracks: one devoted again to localization and tracking, with the addition of context information provided by the living lab infrastructure, such as opening and closing door events, or switching the light on and off, so that systems able to exploit this information can improve their localization accuracy. A second track will be devoted to indoor activity recognition.

Our aim is to gradually expand the scope of the competition to topics that can be integrated into a rich AAL environment. In the short term, beyond localization, tracking and activity recognition, other topics on which we are seeking for convergence from other AAL stakeholders are, for example, teleoperated robots, user interaction and interfaces, reasoning and possibly others. The Call for Ideas, which is published yearly, is the mean to reach this convergence, and all researchers that share this view are invited to respond and contribute.

References

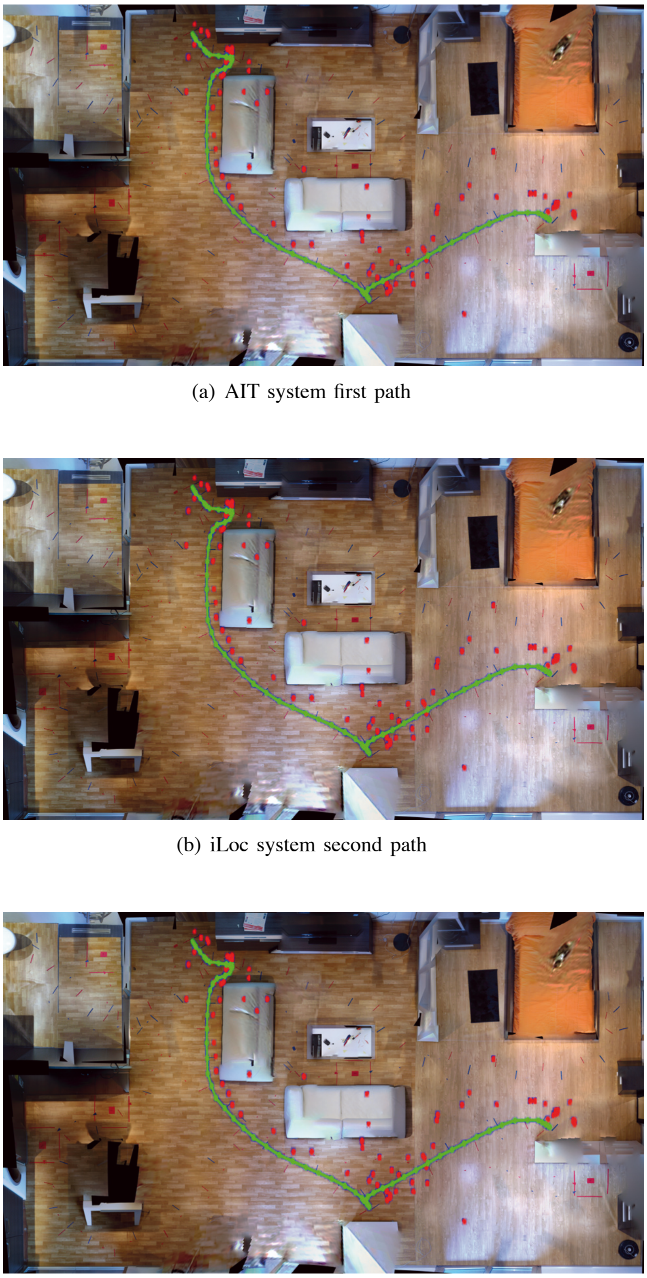
[1] "Ambient Assisted Living roadmap," 2009. [Online]. Available: http://www.aaliance.eu/public/documents/aaliance-roadmap/

[2] K. Ducatel, M. Bogdanowicz, F. Scapolo, J. Leijiten, and J. Burgelman, "Scenarios for ambient intelligence in 2010," IST Advisory Group, Tech. Rep., February 2001.

[3] K. Connelly, K. Siek, I. Mulder, S. Neely, G. Stevenson, and C. Kray, "Evaluating pervasive and ubiquitous systems," *IEEE Pervasive Computing*, vol. 7, no. 3, pp. 85–88, 2008.

TABLE II
THE FINAL SCORES OF COMPETING SYSTEMS.

Competitor	Accuracy	Availability	Installation Complexity	User Acceptance	Integrability in AAL	Final score
n-Core [8]	5,96	9,88	10	7.6	6.5	7.14
AIT [9]	8,45	1,37	6,8	6,88	8,5	5,90
iLoc [10]	7,80	9,39	0	5,88	4,5	4,98
OwlPS [11]	1,37	9,43	8,5	6,5	1	4,85
GEDES-UGR [12]	1,81	9,02	0	6	10	4,00
SNTUmicro [13]	0	0	10	4,38	3	3,17



(a) AIT system first path

(b) iLoc system second path

(c) iLoc system third path

Fig. 6. Reference paths and estimated ones. Systems with the highest accuracy score are depicted.

[4] F. Furfari, F. Potort, S. Chessa, M. Tazari, M. Hellenschmidt, R. Wichert, J. Gorman, and A. Kung, "The AAL open association manifesto," in *Smart Sensing and Context EuroSSC 2010. Platforms for AAL Applications*, K. K. Paul Lukowicz and G. Kortuem, Eds., vol. 6446. Heidelberg, Germany: Springer, 2010, pp. 190 – 194. [Online]. Available: http://www.aaloa.org/manifesto/manifesto_0_9_0

[5] "Trading Agent Competition," 2010. [Online]. Available: http://www.sics.se/tac/

[6] "DARPA Grand Challenge," 2007. [Online]. Available: http://www.darpa.mil/grandchallenge/index.asp

[7] "CIAmI Living Lab." [Online]. Available: http://www.ciami.es/valencia/

[8] "n-Core." [Online]. Available: http://n-core.info/

[9] T. Fuxreiter, C. Mayer, S. Hanke, M. Gira, M. Sili, and J. Kropf, "A modular platform for event recognition in smart homes," in *e-Health Networking Applications and Services (Healthcom), 2010 12th IEEE International Conference on*, july 2010, pp. 1 –6.

[10] S. Knauth, C. Jost, and A. Klapproth, "iLoc: a localisation system for visitor tracking and guidance," in *in Proceedings of the Embedded World Conference*, Nuremberg, Germany, Mar. 2009.

[11] M. Cypriani, F. Lassabe, P. Canalda, and F. Spies, "Wi-fi-based indoor positioning: Basic techniques, hybrid algorithms and open software platform," in *Indoor Positioning and Indoor Navigation (IPIN), 2010 International Conference on*, sept. 2010, pp. 1 –10.

[12] T. Ruiz-Lòpez, J. L. Garrido, C. Rodrìguez-Domìnguez, and M. Noguera, "Sherlock: A Hybrid, Adaptive Positioning Servicebased on Standard Technologies," Sept. 2011, to appear in AAL Forum proceedings.

[13] I. B. Shirokov, A. Ponyatenko, and O. Kulish, "The measurement of angle-of-arrival of microwave in a task of precision landing of aircraft," in *In Electromagnetics Research Symposium, Cambridge, USA*, Cambridge, USA, July 26 2008, pp. 175 – 181.

[14] "EvAAL web site." [Online]. Available: http://evaal.aaloa.org/