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

PerfLoc (Part 2): Performance Evaluation of the Smartphone Indoor Localization Apps

Nader Moayeri, Chang Li, and Lu Shi
National Institute of Standards and Technology
Gaithersburg, MD, USA
Email: {nader.moayeri, chang.li, lu.shi}@nist.gov

***Abstract*—This paper describes the structure of the PerfLoc Prize Competition organized by the US National Institute of Standards and Technology (NIST). The Competition consisted of collecting an extensive repository of smartphone data, releasing the data to researchers across the world to develop smartphone indoor localization algorithms, allowing them to evaluate the performance of their algorithms using a NIST web portal, and rigorous, live testing of the Android apps implementing the best algorithms at NIST. The paper presents detailed performance analysis of the algorithms developed by the top 10 participants as well as the Android indoor localization app that won the first prize in PerfLoc. It also provides a comparison of PerfLoc with other ongoing, annual indoor localization competitions.**

***Index Terms*—indoor localization, smartphone sensor data, smartphone apps, Android, PerfLoc**

I. Introduction

With billions in daily use around the world, the smartphone is the unrivaled king of personal mobile devices. The navigation capabilities of the smartphone enabled by the Global Positioning System (GPS) are widely used for vehicular navigation or simply looking for a place to eat while on foot in unfamiliar surroundings in a city. Even though GPS does not work indoors, many applications are envisioned for smartphone indoor localization and navigation, such as navigating to a store in a shopping mall or a work of art in a museum. The smartphone could even be used by first responders for situation-awareness and command and control purposes while responding to emergencies inside buildings.

Android phones use the Fused Location Provider (FLP) Application Programming Interface (API) [1] developed by Google for indoor/outdoor localization. iPhones use Apple's Core Location Framework [2] for indoor/outdoor localization. A comprehensive performance evaluation of Android FLP and Apple Core Location is beyond the scope of this paper, but the indoor localization accuracy they currently provide may not be adequate for some applications. Hence, it is worthwhile to explore whether more accurate smartphone indoor localization apps can be developed. With that goal in mind, the US National Institute of Standards and Technology (NIST) created and ran the PerfLoc Prize Competition [3] from March 2017 to April 2018. The preparatory steps for the Competition, however, started in August 2015. This paper describes PerfLoc from inception to conclusion.

The rest of the paper is organized as follows. Section II describes our smartphone data collection campaign. The process we used for over-the-web performance evaluation of PerfLoc "algorithms" during the Competition Testing Period that ran from March 2017 to January 2018 is described in Section III. Section IV presents a performance analysis of the top algorithms that were developed during the Competition Testing Period. Two finalist teams were invited to NIST for live tests of their "apps". The details of that evaluation and how that finalist apps performed are presented in Section V. Related work is described in Section VI. Finally, concluding remarks are provided in Section VII.

II. Smartphone Data Collection

Our data collection campaign was carried out in two phases. The original PerfLoc data was collected in early 2016. In fall 2017, i.e., about six months into the Competition Testing Period, NIST collected additional training data sets that were released to the PerfLoc user community in response to their request for such data.

A. Original Data Collection

PerfLoc was developed for the Android Operating System (OS) only, because the iPhone platform is closed and its sensor measurements and RF data are not readily accessible. NIST used four Android phones of different brands to collect data from sensors, such as accelerometer, gyroscope, magnetometer, and barometer, as well as Wi-Fi Received Signal Strength Indicator (RSSI), the strengths of signals received from cellular base stations, and GPS fixes that were occasionally available inside buildings in early 2016. The phones used for data collection were LG G4, Motorola Nexus 6, OnePlus 2, and Samsung Galaxy S6. They were strapped to the arms of the person who collected the data as shown in Figure 1. We used different smartphones to determine whether the data collected varied significantly across different brands and how those differences would affect the performance of PerfLoc indoor localization algorithms. The total space in the four buildings was more than 30,000 m^2. The buildings were a subterranean research facility with two levels below ground level, a three-story office building, a large single-story building housing a warehouse and industrial shops, and a single-story machine shop. We refer to them as Buildings 1-4, respectively, in the rest of the paper. PerfLoc competition participants

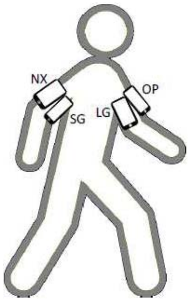

Fig. 1. Positioning of the phones on test subject's arms (LG = LG G4, NX = Motorola Nexus 6, OP = OnePlus 2, and SG = Samsung Galaxy S6)

were not told which building corresponded to which number. To the extent possible, the buildings were selected based on guidance from the international standard ISO/IEC 18305, Test and evaluation of localization and tracking systems [4], whose development NIST led during 2012-2016. Figures 2 and 3 show the interior of Buildings 1 and 4, respectively. The former shows the basement level of Building 1 that houses Heating, Ventilation, and Air Conditioning (HVAC) equipment. The building has a sub-basement level also. Building 1 did not have any Wi-Fi access points (APs) and hardly any cellular or GPS signal was ever received in that building. The relatively small number of weak Wi-Fi signals received in that building were from APs in neighboring buildings, whose locations we had not surveyed. Therefore, Building 1 was the most challenging from an RF signal availability point of view. Building 2 had a good number of Wi-Fi APs with one AP for roughly every 225 m^2 of space. The huge Building 3 had only a few Wi-Fi APs, such that in roughly half of its area no Wi-Fi signal could be received at all. Building 4 was the smallest of the four and it had a few Wi-Fi APs. One could receive Wi-Fi signals everywhere in that building, but from a couple of APs only. We installed more than 900 dots, circular floor markers of diameter 3 cm, in the four buildings and had the dot locations as well as the Wi-Fi AP locations in the buildings professionally surveyed. The dots were used as test points for the algorithms developed by PerfLoc competition participants.

We described Wi-Fi signal availability in detail, because the Wi-Fi signal is the primary information that could be used in PerfLoc to mitigate the drift inherent in localization based on an Inertial Measurement Unit (IMU). (The GPS signal was mostly unavailable in the buildings and localization based on cellular signals is notoriously inaccurate [5].) NIST did not systematically collect Wi-Fi fingerprints in the four buildings, but it provided the 3D coordinates of all Wi-Fi APs and their radio MAC addresses (BSSIDs). This is an important distinction between PerfLoc apps and Android FLP or Apple Core Location. The latter two do not have the Wi-Fi AP locations available to them, but Google and Apple collect a lot of data when a smartphone user allows them to track his/her location. Presumably, Google and Apple have large repositories of Wi-Fi data in buildings. NIST is not privy to how Google and Apple use that information.

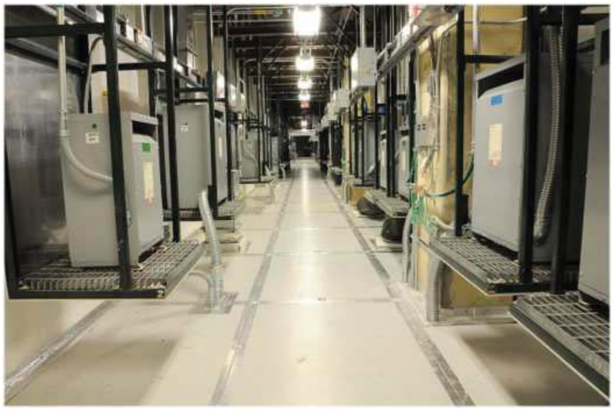

Fig. 2. Interior of basement level in Building 1

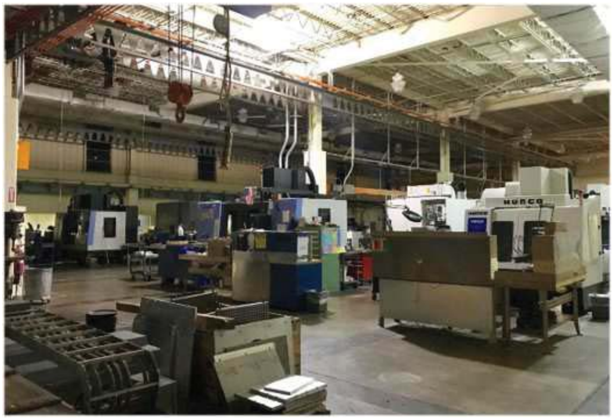

Fig. 3. Interior of Building 4

We collected timestamped training and test data sets. The former comes with ground truth location at each timestamp so that PerfLoc competition participants could assess the accuracy of their indoor localization algorithms. We collected data over 34 test & evaluation (T&E) scenarios in the four buildings, including one training set for each building. This resulted in roughly 15.6 hours of data collected with each phone, which makes PerfLoc a very comprehensive repository of smartphone data. Each T&E scenario involves following a pre-determined course in a building using one or more mobility modes. The mobility modes used were walking normally, walking normally but pausing for 3 s at each dot visited, running, walking backwards, walking sideways (sidestepping), crawling on the floor, using an elevator instead of stairs to change floors, and placing the four phones on the bed of a cart that we pushed around in buildings. For each T&E scenario, we provided the 3D coordinates of the starting point and the initial direction of motion (E, N, W, or S).

We also provided the footprint of each building to allow PerfLoc algorithms to reject location estimates that fell outside the footprint. Specifically, we provided the coordinates of all corners of each building in counterclockwise order.

More details about our original data collection campaign can be found in [6], where we presented statistical analysis of the collected data such as how fast various sensors could be sampled, how periodic the sensor samples were (statistics of inter-sample times), whether different phones saw the same number of Wi-Fi APs, and the correlation of accelerometer data to the motion the phone experienced.

B. Supplemental Data Collection

We had deliberately not specified in the original PerfLoc data when a transition takes place from one mobility mode to another. The PerfLoc competition participants asked NIST to provide additional training data sets that they could use to develop models for various mobility modes as well as smartphone attitude estimation, i.e. in which 3D direction the phone is facing. NIST deployed an additional 386 dots on the floor in a building, one yard (3 floor tiles) apart from each other, and collected several training data sets using the original four phones mentioned above as well as a Google Pixel XL phone that we later used for live tests of the PerfLoc finalist apps. When collecting data with the four original phones, we attached them to the arms of the person who collected the data as shown in Figure 1 or put them on a cart. In case of the Google Pixel XL phone, the person collecting data held it in one hand in front of his/her chest, just like how most people interact with a smartphone. The cart scenario was the same as before. Just to provide some detail, in the scenario that involved crawling on the floor, the person walked for a while, then started crawling on the floor at a time that we specified in the meta data for the scenario, and finally started to walk again at a time specified in the meta data. Whether walking or crawling on the floor, we provided the timestamps for each dot on the course, ground truth location of each dot, and of course all sensor measurements made by the phone(s) on a continuous basis. In the data set for attitude estimation, we put the phone on the floor in one of six ways that we specified in our meta data and the timestamp for visiting each dot and its ground truth location. Attitude estimation plays an important role in indoor localization, because the accelerometer, gyroscope, and magnetometer in the phone measure respective sensor values with respect to the reference coordinate system of the phone, but the phone's orientation changes continuously as a result of the body motion of the person who collected the data. Just like the original PerfLoc data, the supplemental data is available on the PerfLoc website [3] along with detailed descriptions of the data.

III. Offline Evaluation of PerfLoc Algorithms

NIST created an over-the-web performance evaluation capability for PerfLoc algorithms to provide instant objective feedback to PerfLoc competition participants on how good their algorithms were and to give them the opportunity to make their algorithms better. We developed a web portal where a PerfLoc competition participant could upload the location estimates generated by his/her algorithm for the timestamps specified in each of the 30 test data sets. The web portal would then compute the spherical error 95% (SE95) [4] for each of the four buildings as well as an overall SE95 for all test data sets and report those figures instantaneously back to the PerfLoc competition participant. Competing PerfLoc algorithms were ranked and listed in a leaderboard published on the PerfLoc website in the ascending order of overall SE95. Location estimates generated by all four phones had to be uploaded to allow assessing the performance of a given algorithm on different phones, even though in reality most PerfLoc competition participants used the measurements made by "all" four phones to estimate the location at time instances of interest and then uploaded the same location estimates for all four phones! That aspect was unfortunate, because it prevented NIST from quantifying the differences between the four phones in terms of localization accuracy achieved by the same algorithm.

A key aspect of designing the performance evaluation portal was to incorporate mechanisms to protect the integrity of the method used to rank competing algorithms and to prevent any PerfLoc competition participant from gaming the system and achieving a bogus performance. Specifically, NIST had to prevent PerfLoc competition participants from algorithmically computing the locations of our dots (test points) based on the numerical feedback they were getting from the PerfLoc performance evaluation portal. We realized that the system would be vulnerable to such abuse if we reported the mean of magnitude of 3D localization error, i.e. the difference between the ground truth location of a dot and the estimate for that location provided by the PerfLoc algorithm under test. It is possible to precisely compute a dot location by three uses of the performance evaluation portal and then solving a not-so-difficult nonlinear system of equations. Consequently, we chose not to use the mean of magnitude of 3D error vector as our performance metric, even though all other indoor localization competitions use that metric. Instead, we selected SE95 as our performance metric. We realize that even SE95 is vulnerable to abuse, but the algorithm to compute dot locations based on SE95 feedback is much more complicated than the corresponding algorithm for the mean of magnitude of 3D error vector.

Other steps we took to make it harder to abuse the performance evaluation portal was to ask each team to pledge they would create only one user account. We then limited the number of times a team could use the portal to three uploads per day and a total of 150 uploads during the Competition Testing Period. In reality, we did not have an assured way of guaranteeing that a team would not create more than one user account and there was one instance of abuse of that clause in the competition rules that led to disqualification of one team late into the Competition Testing Period. Overall, the mechanisms NIST developed and implemented in the PerfLoc Prize Competition were a compromise between preventing abuse of the system and ease of joining the competition and using the system.

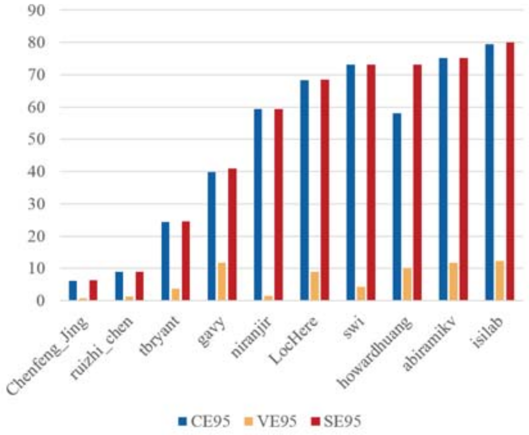

Fig. 4. CE95, VE95 and SE95 of the top 10 teams

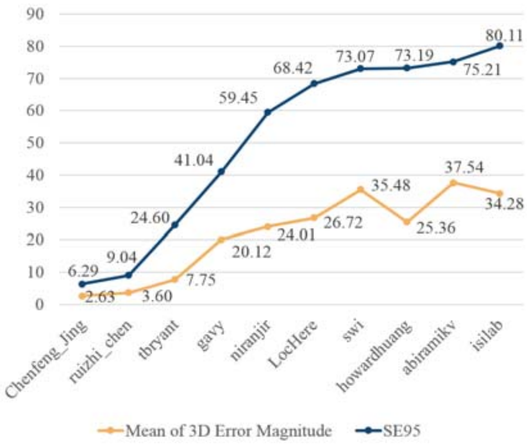

Fig. 5. Comparison of mean of 3D error magnitude and SE95 for the top 10 teams

IV. Offline Performance of PerfLoc Algorithms

A total of 152 teams registered on the PerfLoc website, but only 16 teams uploaded location estimates on the website. We attribute this to PerfLoc's steep learning curve. There were a lot of details in the PerfLoc User Guide [3] that the participants had to master before decoding the data and beginning to use it. The Competition Testing Period, during which participants were allowed to upload location estimates, lasted for almost ten months and it closed in January 2018. The overall SE95, circular error 95% (CE95), which characterizes horizontal accuracy, and vertical error 95% (VE95) of best performance achieved by the top 10 teams computed over all 30 T&E scenarios have been depicted in Figure 4. These are the 95 percentile points on the Cumulative Distribution Functions (CDFs) of the respective error magnitudes. We observe that the vertical error is much smaller than the horizontal error, and hence the 3D error is dominated by the horizontal error. The figure also shows the usernames of the top 10 teams.

Figure 5 compares the overall SE95 and mean of magnitude of 3D error achieved by various teams. It shows that the means are considerably smaller than the SE95s. The top team achieved an overall SE95 of 6.29 m and mean of magnitude of 3D error of 2.63 m.

Table I shows the SE95 and mean of magnitude of 3D error achieved by the top 10 teams in different buildings. Aside from the top two teams who have achieved roughly the same localization accuracy in all four buildings, most of the remaining teams have achieved better performance in Buildings 2 and 4 than in Buildings 1 and 3. This is explained by the fact that Buildings 2 and 4 have better Wi-Fi coverage than Buildings 1 and 3.

Table II shows the VE95 and mean of magnitude of vertical error achieved by the top 10 teams in different buildings. It shows that Buildings 3-4 results are generally better than Buildings 1-2 results. This is due to the fact that while 3-4 are single-story buildings, 1-2 have multiple floors.

V. Live Testing of PerfLoc Apps

At the conclusion of the Competition Testing Period in January 2018 and according to the published PerfLoc Competition Rules, NIST invited teams ranked 2-4 on the PerfLoc Competition Leaderboard for live testing of their apps at NIST. (The top-ranked team was not invited, because they were not eligible to receive cash prizes.) Team #4 declined the invitation. Teams #2 and #3 accepted the invitation and traveled to NIST for live testing of their apps on April 26-27, 2018. Unfortunately, Team #3 was not able to get its app to function at all and dropped out of the race. Therefore, Team #2 (with username ruizhi_chen) from Wuhan University in China, being the only team that managed to go through the suite of tests administered by NIST, was declared the winner of the PerfLoc Prize Competition.

A. Purpose and Structure of Live Tests

The purpose of the live tests were threefold. First, NIST wished to ascertain that any finalist PerfLoc algorithm could be implemented as an Android app. NIST could not determine from the offline performance results whether an algorithm would need to be run on a super computer and/or would take days to generate location estimates. Second, it was important to find out how a finalist app would fare in a blind, live test and how that would compare to the offline performance of the same app/algorithm. By blind test we mean testing an app in a building that the finalist knew nothing about until the test days. In other words, the finalists did not know how large the building was, how many Wi-Fi APs it had, and they did not have any training data for the building. NIST assumed that PerfLoc finalists had perfected their algorithms/apps during the Competition Testing Period and they were supposed to be ready for testing in "any" building by the live test days. This was indeed a tall order. Third, NIST wished to measure the latency of each finalist app, i.e. the time it takes for an app to provide a location estimate. Note that there is a tradeoff between latency and localization accuracy. If an algorithm uses a bit of lookahead, i.e. some future sensor and RF measurements, before generating a location estimate for the present time, its location estimates would be more accurate

TABLE I
SE95 AND MEAN OF 3D ERROR MAGNITUDE IN DIFFERENT BUILDINGS FOR THE TOP 10 TEAMS

Rank	Participant	SE95 Performance (m)				Mean of 3D Error Magnitude (m)			
		Bldg. 1	Bldg. 2	Bldg. 3	Bldg. 4	Bldg. 1	Bldg. 2	Bldg. 3	Bldg. 4
1	Chenfeng_Jing	5.44	7.04	6.26	5.38	2.14	3.19	2.89	2.23
2	ruizhi_chen	12.74	8.60	7.34	8.99	3.84	3.61	3.27	3.48
3	tbryant	31.92	17.78	20.93	17.91	8.44	8.17	6.83	6.30
4	gavy	42.71	24.56	37.28	26.33	27.34	14.23	17.97	13.90
5	niranjir	41.88	24.81	71.25	52.27	24.91	10.45	37.62	20.63
6	LocHere	73.12	21.83	74.55	40.29	36.88	10.42	32.56	15.74
7	swi	42.79	57.68	82.99	49.62	25.09	33.17	55.96	26.07
8	howardhuang	126.87	20.20	72.77	39.83	34.71	9.38	29.92	20.38
9	abiramikv	42.71	60.14	84.95	65.39	27.34	33.57	56.30	36.95
10	isilab	73.59	17.98	339.29	25.03	42.96	10.07	54.07	12.44

TABLE II
VE95 AND MEAN OF VERTICAL ERROR MAGNITUDE IN DIFFERENT BUILDINGS FOR THE TOP 10 TEAMS

Rank	Participant	VE95 Performance (m)				Mean of Vertical Error Magnitude (m)			
		Bldg. 1	Bldg. 2	Bldg. 3	Bldg. 4	Bldg. 1	Bldg. 2	Bldg. 3	Bldg. 4
1	Chenfeng_Jing	0.04	3.25	0.05	0.02	0.07	0.58	0.03	0.01
2	ruizhi_chen	0.58	2.81	0.03	0.18	0.09	0.72	0.01	0.17
3	tbryant	3.78	4.85	3.40	3.40	1.68	1.70	1.69	1.31
4	gavy	11.84	5.97	1.22	6.48	9.56	3.22	0.39	1.54
5	niranjir	0.05	2.29	0.03	0.02	0.10	0.98	0.01	0.01
6	LocHere	9.00	3.55	0.03	0.02	3.67	0.44	0.01	0.01
7	swi	2.39	4.30	0.02	0.01	2.41	2.66	0.01	0.01
8	howardhuang	122.45	5.98	0.61	1.04	15.64	3.73	0.51	1.02
9	abiramikv	11.84	7.30	0.02	0.24	9.56	4.47	0.01	0.23
10	isilab	12.40	3.41	9.13	0.02	6.26	0.72	0.69	0.01

Fig. 6. The building used for the live tests

compared to not using any lookahead. Perhaps a lookahead of ~2 s would be acceptable.

The finalist app was tested in a very large, tall building with about 30,000 m^2 of space and 131 Wi-Fi APs. A picture of this building is shown in Figure 6. Note that the first floor and the basement of this building are much larger than the tower part. This building by itself is as large as the four buildings used during the offline performance evaluation phase put together. The 3D coordinates and BSSIDs of the Wi-Fi APs, the building footprint, and the 3D coordinates and initial direction of motion for each of the 8 T&E scenarios used were provided to the finalist app. Unlike the offline phase, where NIST had provided smartphone sensor and RF data at sampling rates of NIST's choosing, the decision of how fast to sample various data was left to the developers of the finalist app. The T&E scenarios were (i) normal non-stop walking, (ii) normal walking with 3 s stops at each test point visited, (iii) transporting an asset on a push cart, (iv) normal walking and use of elevators, (v) normal walking with instances of leaving and reentering the building, (vi) sidestepping, (vii) walking backwards, and (viii) crawling on the floor.

B. Live Test Performance Results

The latency of the Wuhan University Team app turned out to be ~5 milliseconds.

Table III shows the performance of the app. The first observation we make is that the live test results are not nearly as good as the offline performance results. This discrepancy is due to (i) unavailability of training data sets in the live tests, (ii) real-time operation requirement of the live tests that prevented the app from post-processing and revising/improving location estimates for past test points, (iii) the building used in the live tests being much larger than the four used during offline performance evaluation, and (iv) the app being forced to use a single algorithm during the live tests as opposed to possibly using building-specific algorithms during the offline phase.

A few other observations can be made. The first two rows of Table III show that the horizontal error is much larger than the vertical error. The same observation can be made by comparing CE95 and VE95 figures. Normal walking's performance is considerably better than the overall (over all mobility

TABLE III
LIVE TESTING PERFORMANCE OF WUHAN UNIVERSITY TEAM APP

	Overall	Normal Walking	Cart	Sidestepping	Walking Backwards	Crawling
Mean of Magnitude of Horizontal Error	17.66	10.76	49.67	27.79	38.07	10.31
Mean of Magnitude of Vertical Error	1.71	2.09	0.72	0.57	0.20	0.34
Mean of Magnitude of 3D Error	18.16	11.43	49.68	27.81	38.07	10.32
CE95	72.84	24.86	78.92	79.17	83.10	18.56
VE95	5.20	5.22	1.16	0.95	0.47	0.57
SE95	72.84	24.86	78.92	79.17	83.10	18.56
Floor Detection Probability	70.66%	59.35%	100%	100%	100%	100%

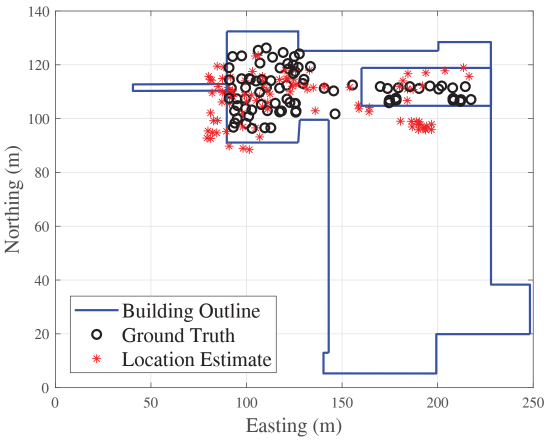

Fig. 7. Horizontal error of walking scenario with 3 s stops

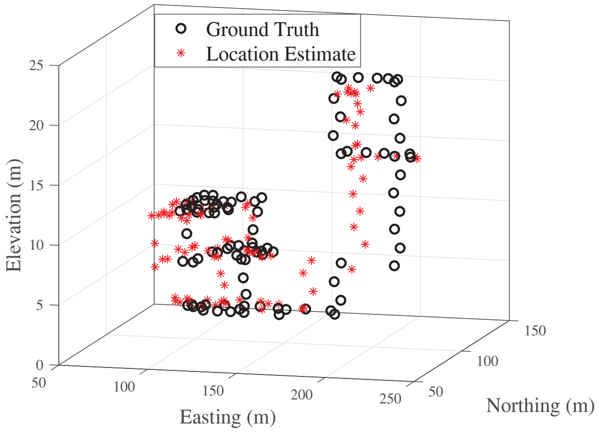

Fig. 8. Vertical error of walking scenario with 3 s stops

modes) performance. Specifically, the cart, sidestepping, and walking backwards scenarios have much worse performance than normal walking.

Figures 7 and 8, respectively, show the horizontal and vertical performance of the app in the normal walking scenario with 3 s stops at test points visited. It is hard to tell that Figure 7 represents good performance. However, at least it can be seen that the black circles and the red stars do not look like two sets separated spatially as in some figures to be introduced shortly. Figure 8 does show that the app is doing a good job of tracking the test subject's elevation.

Figures 9, 10, and 11, respectively, show the horizontal performance of the app in the cart, sidestepping, and walking backwards scenarios. It is clear that the app is not doing a good job of tracking the test subject's location. Sometimes it overestimates how far the test subject has moved and sometimes it underestimates. In all three cases, one can see a good separation of the black circles and the red stars. As shown in Table III, the vertical error achieved in these scenarios, which were carried out on the same floor of the building, is small. Therefore, no figures on vertical performance are provided for these scenarios.

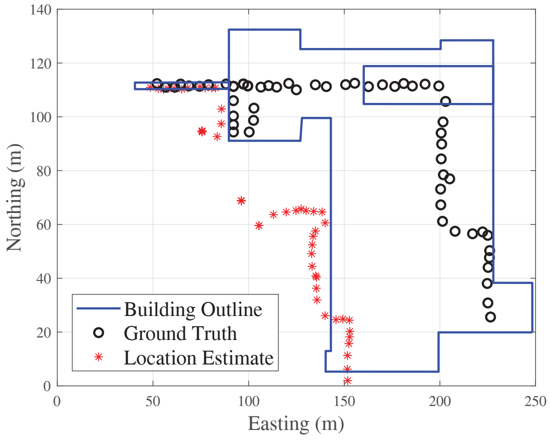

Fig. 9. Horizontal error of the cart scenario

Figure 12 is a plot of the magnitude of 3D error vs. time in the scenario that involved normal walking and four uses of elevators. In two cases of using the elevator marked in the figure, where we went up or down by several floors, a large increase in the magnitude of error is observed. After each such increase, the error drops after some time, perhaps as a result of getting good location fixes from the Wi-Fi signals. There are two other cases in this scenario, where we used the elevator to go up or down by just one floor at ~330 s and ~565 s. In these cases, there is no significant change in the magnitude of error.

Figure 13 is a plot of the magnitude of 3D error vs. time in the normal walking scenario with two instances of leaving and reentering the building. We conclude that getting GPS fixes by leaving the building does not help to mitigate the

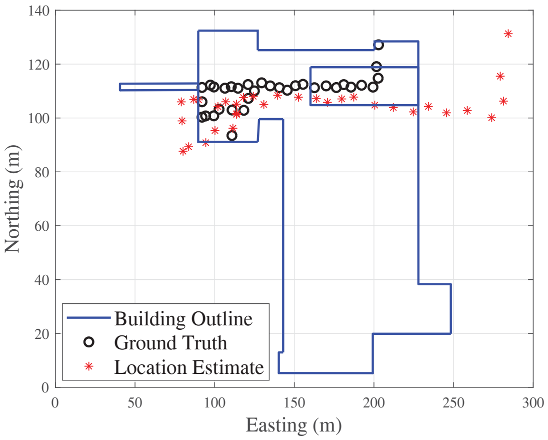

Fig. 10. Horizontal error of the sidestepping scenario

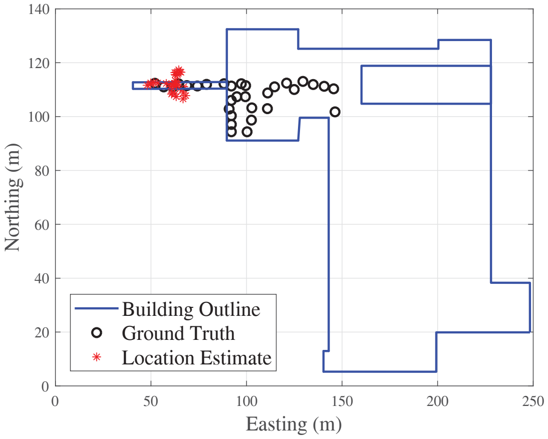

Fig. 11. Horizontal error of the walking backwards scenario

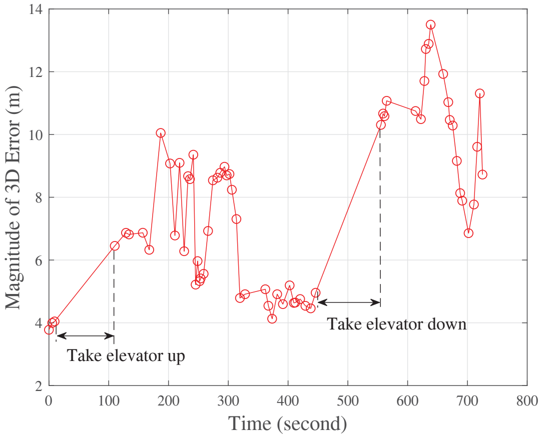

Fig. 12. Mean of 3D error magnitude vs. time for the scenario involving walking and use of elevators

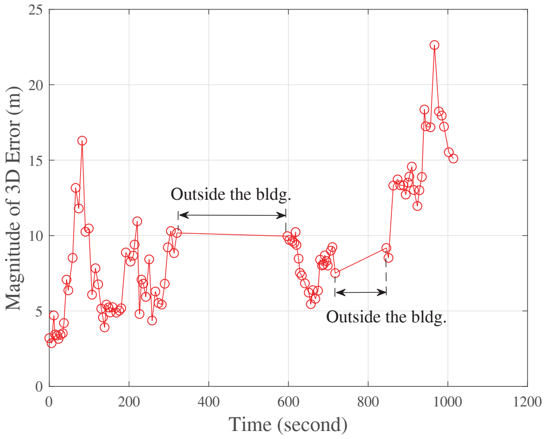

Fig. 13. Mean of 3D error magnitude vs. time for the scenario involving walking in/out of a building

drift of inertial sensors any more so than location fixes from Wi-Fi signals. Note that the error is in the range 7.5-10 m after walking outdoors and getting GPS fixes for about 4.5 minutes in the first instance of leaving the building and about 2.5 minutes in the second instance.

VI. Related Work

There are two indoor localization competitions that have been held on an annual basis for the past several years. The first one is the competition held in conjunction with the IEEE Indoor Positioning and Indoor Navigation (IPIN) Conference since 2011. These competitions are typically organized by the EvAAL (Evaluating AAL Systems through Competitive Benchmarking) Project [7], where AAL stands for Ambient Assisted Living. In addition to indoor localization, EvAAL is interested in indoor activity recognition. The second one is the Microsoft Indoor Localization Competition [8] that has been held in conjunction with the IEEE Information Processing in Sensor Networks (IPSN) Conference since 2014. The structure of these competitions may change somewhat from one year to the next. We describe the structure for the latest edition of each competition prior to the writing of this paper.

The 2018 Microsoft Indoor Localization Competition was held in two tracks, (i) 2D Track and (ii) 3D Track. The systems competing in the 2D Track used technologies such as Pedestrian Dead Reckoning (PDR), camera, and Wi-Fi fingerprinting. One system used Wi-Fi Time of Flight (ToF). The competing systems were evaluated based on mean horizontal localization error and the best system achieved 2.3 m mean error. In the 3D Track, the contestants were allowed to install up to 10 anchor nodes of their choice in the evaluation area, which was about 600 m^2. Different systems were compared based on the mean 3D error performance metric. In addition to the technologies used in the 2D Track, the systems competing in this track also used ultra wideband (UWB) ranging, sound, and ARKit [9]. One system used a phase-based localization technique. The best system in the 3D Track integrated UWB and ARKit to achieve a mean 3D error of 27 cm. Of course, this is far better than the 18.16 m mean 3D localization error the winning PerfLoc app achieved in the live tests described in Section V, but such a comparison would not be fair for many reasons. First, the PerfLoc app did not have the benefit of using UWB or ARKit, granted that the use of the latter would be reasonable in a smartphone app. (In case of Android phones, one would use ARCore [10].) Second, the PerfLoc app was tested over scenarios that lasted as long as 20 minutes in a huge, tall building. Had it been tested over a 2-3 minute scenario in a small area, it would have achieved better results. Third, the PerfLoc app was tested over many different modes of mobility. Its performance for normal walking is 11.43 m, as shown in Table III. Even a comparison of PerfLoc with the 2D results of the Microsoft Indoor Localization Competition would not be fair due to the differences in the evaluation area size. Wi-Fi fingerprinting was allowed in the Microsoft Indoor Localization Competition, but ironically PDR alone was better than PDR plus fingerprinting due to the use of transmit power control by Wi-Fi APs. It is fair to say that PerfLoc played a different role than the Microsoft Indoor Localization Competition does. The latter generates a lot of interest every year as a forum for emerging techniques and solutions. PerfLoc was restricted to the Android smartphones and it used more rigorous testing.

It is difficult to compare performance results from various competitions. Even if one compares similar systems, e.g. smartphone apps that use Wi-Fi fingerprinting, one still has to take into account the differences in the evaluation areas. These issues have been addressed in detail in the international standard ISO/IEC 18305 [4]. These evaluations are site-dependent. The best that can be done is to compare several systems evaluated in the same set of buildings roughly at the same time.

The 2017 IPIN Competition was held in four tracks: (i) Smartphone-Based, (ii) PDR Positioning, (iii) Smartphone-Based (Off-Site), and (iv) PDR for Warehouse Picking (Off-Site). Track 1 was similar to the live tests of the winning PerfLoc app, but it allowed Wi-Fi fingerprinting prior to the tests (as opposed to providing Wi-Fi AP locations only in PerfLoc) and it made the detailed map of the evaluation area available to the contestants (as opposed to providing the building footprint only in PerfLoc). In Track 3, training data sets with ground truth location data were made available to the contestants to develop and fine-tune their algorithms, which were subsequently tested by the competition organizers using a data set the contestants did not have access to. Unlike PerfLoc, there were no opportunity to get any feedback on the performance of an algorithms during its development phase. Wi-Fi fingerprints were also made available to the contestants in the IPIN Track 3 Competition. The evaluation area in the 2017 IPIN Competition was about 1,500 m^2 over two floors. The best CE75 (as opposed to CE95 used in PerfLoc) in Tracks 1 and 3 were 8.8 m and 3.48 m, respectively. These performance results look better than that of the winning PerfLoc app, but one has to keep in mind that IPIN apps were tested in a smaller area, had detailed maps available to them, and could use Wi-Fi fingerprinting. In addition, CE75 is a less stringent performance metric than CE95. The evaluation area used in the 2016 IPIN Competition was even larger and covered 3 floors of a building.

VII. Conclusions

PerfLoc was a prize competition for developing smartphone indoor localization apps with the minimal requirement of having access to the locations and BSSIDs of Wi-Fi APs, if the building has Wi-Fi. This is less onerous than having to make Wi-Fi fingerprints available to apps. The Android apps developed during this competition were evaluated during an offline phase and through comprehensive live tests at NIST. The winning app achieved a mean 3D error of about 10 m when the test subject walked around the building with the smartphone in his/her hand and used stairs or elevators to change floors. The winning app did a great job of estimating on which floor of the building the person carrying the smartphone was by achieving a mean vertical error of 1.71 m. However, the app did not perform well for other modes of mobility such as sidestepping, walking backwards, or tracking the movements of an object transported on a push cart in the building. The PerfLoc Prize Competition is now closed, but the problem of developing more effective indoor localization smartphone apps with better accuracy is still very much open. The extensive repository of PerfLoc smartphone data continues to be available to researchers for doing just that.

Disclaimer

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards nd Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

References

[1] Google, “Fused Location Provider API.” [Online]. Available: https://developers.google.com/location-context/fused-location-provider/

[2] Apple, “Core Location.” [Online]. Available: https://developer.apple.com/documentation/corelocation

[3] US National Institute of Standards and Technology, “PerfLoc Prize Competition.” [Online]. Available: https://perfloc.nist.gov

[4] “ISO/IEC 18305:2016, Information technology – Real time locating systems, Test and evaluation of localization and tracking systems.” [Online]. Available: https://www.iso.org/standard/62090.html

[5] P. A. Zandbergen, “Accuracy of iPhone locations: A comparison of assisted GPS, WiFi and cellular positioning,” *Trans. GIS*, vol. 13, no. s1, pp. 5–25, 2009.

[6] N. Moayeri, M. Ergin, F. Lemic, V. Handziski, and A. Wolisz, “PerfLoc (Part 1): An extensive data repository for development of smartphone indoor localization apps,” in *Proc. IEEE Int. Symp. Pers. Indoor Mobile Radio Comm. (PIMRC)*, Valencia, Spain, Sep. 2016, pp. 1–7.

[7] “Evaluating AAL Systems through Competitive Benchmarking.” [Online]. Available: http://evaal.aaloa.org/

[8] “Microsoft Indoor Localization Competition.” [Online]. Available: https://www.microsoft.com/en-us/research/event/microsoft-indoor-localization-competition-ipsn-2018/

[9] Apple Developer, “ARKit.” [Online]. Available: https://developer.apple.com/arkit/

[10] Google Developers, “ARCore.” [Online]. Available: https://developers.google.com/ar/discover/