CellSense: An Accurate Energy-Efficient GSM Positioning System

Mohamed Ibrahim, *Student Member, IEEE*, and Moustafa Youssef, *Senior Member, IEEE*

Abstract—**Context-aware applications have been gaining huge interest in the last few years. With cell phones becoming ubiquitous computing devices, cell phone localization has become an important research problem. In this paper, we present CellSense, which is a probabilistic received signal strength indicator (RSSI)-based fingerprinting location determination system for Global System for Mobile Communications (GSM) phones. We discuss the challenges of implementing a probabilistic fingerprinting localization technique in GSM networks and present the details of the CellSense system and how it addresses these challenges. We then extend the proposed system using a hybrid technique that combines probabilistic and deterministic estimations to achieve both high accuracy and low computational overhead. Moreover, the accuracy of the hybrid technique is robust to changes in its parameter values. To evaluate our proposed system, we implemented CellSense on Android-based phones. Results from two different testbeds, representing urban and rural environments, for three different cellular providers show that CellSense provides at least 108.57% enhancement in accuracy in rural areas and at least 89.03% in urban areas compared with current state-of-the-art RSSI-based GSM localization systems. In additional, the proposed hybrid technique provides more than 6 and 5.4 times reduction in computational requirements compared with state-of-the-art RSSI-based GSM localization systems for rural and urban testbeds, respectively. We also evaluate the effect of changing the different system parameters on the accuracy–complexity tradeoff and how the cell tower and fingerprint densities affect system performance.**

Index Terms—**GSM Localization, location fingerprinting, wireless networks.**

I. INTRODUCTION

AS CELL PHONES become more ubiquitous in our daily lives, the need for context-aware applications increases. One of the main context information is location, which enables a wide set of cell phone applications including navigation, location-aware social networking, and security. Although the Global Positioning System (GPS) [2] is considered to be the most well-known localization technique, it is not available in

many cell phones, requires direct line of sight to the satellites, and consumes a lot of energy. Therefore, research for other techniques for obtaining cell phone location has gained momentum fueled by both the users' need for location-aware applications and government requirements, e.g., Federal Communications Commission [3]. City-wide WiFi-based localization for cellular phones has been investigated in [4] and [5], and commercial products are currently available [6]. However, WiFi chips, similar to GPS, are not available in many cell phones, and not all cities in the world contain sufficient WiFi coverage to obtain ubiquitous localization. Similarly, using augmented sensors in cell phones, e.g., accelerometers and compasses, for localization has been proposed in [7]–[9]. However, these sensors are still not widely used in many phones. On the other hand, Global System for Mobile Communications (GSM)-based localization, by definition, is available on all GSM-based cell phones, which is present in 80–85% of today's cell phones [10], works all over the world, and consumes minimal energy in addition to the standard cell phone operation. Many research works have addressed the problem of GSM localization [3], [5], [11], [12], including time-based systems, angle-of-arrival (AOA)-based systems, and received signal strength indicator (RSSI)-based systems. Only recently, with the advances in cell phones, GSM-based localization systems have been implemented [5], [11], [12]. These systems are mainly RSSI based as RSSI information is easily available to the user's applications. Since RSSI is a complex function of distance, due to the noisy wireless channel, RSSI-based systems usually require building an RF fingerprint of the area of interest [5], [11], [12]. A fingerprint stores information about the RSSI received from different base stations at different locations in the area of interest. This is usually constructed once in an offline phase. During the tracking phase, the received RSSI at an unknown location is compared with the RSSI signatures in the fingerprint, and the closest location in the fingerprint is returned as the estimated location. Constructing the fingerprint is a time consuming process. However, this is typically done in a process called war driving, where cars scan the streets of a city to map it. Current commercial systems, such as Skyhook, Google MyLocation, and StreeView services, already perform scanning for other purposes. Therefore, constructing the fingerprint for GSM localization can be piggybacked on these systems without extra overhead.

In this paper, we propose CellSense, which is a probabilistic fingerprinting-based technique for GSM localization. Unlike current fingerprinting techniques for GSM phones that use a deterministic approach for estimating the location of cell phones [11], [12], the CellSense probabilistic technique provides more

M. Ibrahim was with the Wireless Intelligent Networks Center, Nile University, Cairo 12677, Egypt. He is now with the University of Technology of Troyes, Troyes, France (e-mail: m.ibrahim@nileu.edu.eg).

M. Youssef is with the Department of Computer Science and Engineering, Alexandria University and Egypt–Japan University of Science and Technology, Alexandria 21934, Egypt (e-mail: moustafa.youssef@ejust.edu.eg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

accurate localization. However, constructing a probabilistic fingerprint is challenging as we need to stand at each fingerprint location for a certain amount of time to construct the signal strength histogram. This adds significantly to the overhead of the fingerprint construction process. CellSense addresses this challenge by using gridding, where the area of interest is divided into a grid, and the histogram is constructed for each grid cell. This not only removes the extra overhead of standing at each location for a certain time but also helps in increasing the scalability of the technique as the fingerprint size can be reduced arbitrarily by increasing the grid cell length. To further reduce the computational overhead of CellSense, we propose a hybrid technique, i.e., CellSense-Hybrid, that combines a probabilistic estimation phase with a deterministic refinement phase. The CellSense-Hybrid technique also has the added advantage of its accuracy being robust to changes in its parameter values.

To evaluate CellSense, we implement it on Android-enabled cell phones and compare its performance to other deterministic fingerprinting techniques, model-based techniques, and Google MyLocation service under two different testbeds representing rural and urban environments for three different cellular providers. We also study the effect of different parameters on the performance of CellSense. Our results show that CellSense outperforms other systems, with at least 108.57% and 89.03% enhancement in accuracy for urban and rural testbeds, respectively. In addition, it has significant savings in terms of energy consumption with five to six times saving in running time. Moreover, the CellSense-Hybrid technique accuracy is robust to changes in parameter values.

To summarize, the contribution of this paper is threefold:

1) We introduce the CellSense probabilistic GSM localization system. CellSense provides high localization accuracy and depends on a novel gridding technique to reduce the fingerprint construction overhead.

2) We further extend CellSense through a hybrid technique that adds a deterministic refinement phase to the basic CellSense technique. The accuracy of the CellSense-Hybrid technique is robust to changes in its parameters values. Therefore, the CellSense-Hybrid technique parameters can be selected to achieve a low computational overhead while maintaining the same accuracy.

3) We thoroughly evaluate the performance of CellSense and CellSense-Hybrid techniques both through analysis and under two different testbeds and show their significant advantage compared with other state-of-the-art GSM localization systems.

The rest of this paper is organized as follows. In Section II, we discuss relevant related work. In Section III, we present our CellSense system. Section IV presents the performance evaluation of our system. Finally, Section V concludes this paper and gives directions for future work.

II. Related Work

In this section, we discuss the different techniques for cell phone localization and how they differ from the proposed work. We categorize these techniques as time based, AOA based, cell id based, city-wide WiFi localization, augmented sensor based, and signal strength based.

A. Time-of-Arrival (ToA)-Based Localization

In ToA-based systems, the cell phone estimates its distance to a reference point based on the time a signal takes to travel from the reference point to it. Similarly, time-difference-of-arrival-based systems use the principle that the emitter location can be estimated by the intersection of the hyperbolae of constant differential ToA of the signal at two or more pairs of base stations [3].

The most well-known localization technique, i.e., the GPS [2], can be categorized as a ToA-based system. Time-based systems require special hardware and therefore are usually deployed on high-end phones. In addition, GPS suffers from two other main problems: 1) availability and 2) power consumption. It requires line of sight to the satellites; therefore, it does not work indoors, and it consumes a lot of power of the energy-limited cell phones.

B. AOA-Based Systems

AOA-based systems use triangulation based on the estimated AOA of a signal at two or more base stations to estimate the location of the desired transmitter [3], [13]–[16]. Antenna arrays are usually used to estimate the AOA. Similar to TOA-based systems, AOA-based systems require specialized hardware, which makes them less attractive for a large deployment on cell phones.

C. Cell-ID-Based Techniques

Cell-ID-based techniques, e.g., Google MyLocation [17], do not explicitly use RSSI but rather estimate the cell phone location as the location of the cell tower the phone is currently associated with. This is usually the cell tower with the strongest RSSI. Such techniques require a database of cell tower locations and provide an efficient, although coarse-grained, localization method.

D. City-Wide WiFi-Based Localization

City-wide WiFi-based localization has been proposed in [4] and [5], and commercial products are currently available, e.g., [6]. However, WiFi chips, similar to GPS, are not available in the majority of cell phones, and not all cities in the world contain sufficient WiFi coverage to obtain ubiquitous localization.

E. Augmented Sensor-Based Localization

Using augmented sensors in cell phones, e.g., accelerometers and compasses, for localization has been proposed in [7]–[9] and [18]. For example, in [18], the accelerometer and compass are used as an energy-efficient way for estimating the phone displacement and direction. Due to the accumulation of error, they synchronize with the GPS as needed.

The main issue with augmented sensor-based localization systems is that these sensors are still not widely used in cell phones.

F. RSSI-Based Systems

Recently, RSSI-based systems have been introduced and implemented for cell phone localization. Since RSSI information is readily available to the user's applications on almost all GSM phones, such systems have the potential of localizing 80–85% of today's cell phones [10], work all over the world, and consume minimal energy, in addition to standard cell phone operation.

However, since RSSI is a complex function of distance [19], RSSI-based systems usually require building an RF fingerprint of the area of interest [5], [11], [12], [20]. A fingerprint stores information about the RSSI received from different base stations at different locations in the area of interest. This is usually constructed once in an offline phase. During the tracking phase, the received RSSI at an unknown location is compared with the RSSI signatures in the fingerprint, and the closest location in the fingerprint is returned as the estimated location. Constructing the fingerprint is a time consuming process. However, this is typically done in a process called war driving, where cars drive the area of interest continuously scanning for cell towers and recording the cell tower ID, RSSI, and GPS location. Current commercial systems, such as Skyhook, Google MyLocation, and StreeView services, already perform scanning for other purposes. Therefore, constructing the fingerprint for GSM localization can be piggybacked on these systems without extra overhead.

In the rest of this section, we summarize the current work in fingerprint-based RSSI localization systems for GSM phones, which is the closest to the proposed work.

1) Deterministic Fingerprinting Techniques: Current fingerprinting techniques for GSM localization use only deterministic techniques [11], [12]. For example, each location in the fingerprint of [11] stores a vector representing the RSSI value from each cell tower heard at this location. During the tracking phase, the K-nearest neighbor (KNN) classification algorithm is used, where the RSSI vector at an unknown location is compared with the vectors stored in the fingerprint, and the K-closest fingerprint locations, in terms of Euclidian distance in the RSSI space, to the unknown vector are averaged as the estimated location. Deterministic fingerprinting techniques require searching a larger database than cell-ID-based techniques but provide higher accuracy. Note that the overhead of constructing the fingerprint is the same as constructing the cell-ID database as both require war driving.

2) Modeling-Based Techniques: Modeling-based techniques try to capture the relation between signal strength and distance using a model. For example, the work in [11] uses a Gaussian process (GP) to capture this relation assuming that the received signal strength y_i at location x_i is $y_i = f(x_i) + \epsilon_i$, where ϵ_i is zero-mean additive Gaussian noise with known variance σ_n^2.

A GP estimates posterior distributions over functions f from a training data D (fingerprint). These distributions are represented nonparametrically in terms of training points. A key idea underlying GPs is the requirement that the function values at different points are correlated, where the covariance between two function values $f(x_p)$ and $f(x_q)$ depends on the input locations x_p and x_q. This dependency can be specified via an arbitrary covariance function or kernel $k(x_p, x_q)$. The most widely used kernel function is the squared exponential or Gaussian kernel: $k(x_p, x_q) = \sigma_f^2 exp((-1/2l^2)|x_p - x_q|^2)$, where l is a length scale that determines how strongly the correlation between points drops off.

Building a GP estimator still requires constructing a fingerprint, although a less sparse one. This fingerprint is used to estimate the model parameters (l, σ_n^2, and σ_f^2) and to compute $f(x_*)$ for any location x_*.

This reduces the size of the fingerprint and provides a way for extending a sparse fingerprint to a more dense one as it gives the fingerprint values at any arbitrary location based on the assumed model. However, this comes at the cost of substantial increase in computational requirements, as we quantify in Section IV, and there is no actual saving of fingerprinting overhead as war driving has to be done to collect the training samples D anyway. Moreover, the assumed model may not fit the real environment, thus reducing the accuracy of the returned location.

G. Summary

Compared with TOA, AOA, city-wide WiFi, and augmented sensor-based systems, our proposed system CellSense requires no specialized hardware and is more ubiquitous in terms of the number of cell phones it runs on and the coverage area.

Compared with cell-ID based systems and the current fingerprinting techniques, our technique is probabilistic. Using a probabilistic approach should enhance the accuracy of localization compared with a deterministic approach. However, it comes with its own challenges, such as constructing the RSSI probability distribution with minimal overhead. Our proposed technique addresses these challenges and provides accuracy better than all of the current techniques with minimal computational requirements, as we quantify in Section IV.

III. CELLSENSE SYSTEM

In this section, we describe our CellSense system for GSM phone localization. We start by an overview of the system followed by the details of the offline training and online tracking phases. Finally, we propose a hybrid approach that combines the basic CellSense and a deterministic approach to achieve both accurate localization and low computational overhead.

A. Overview

Fig. 1 shows our CellSense system architecture. CellSense works in two phases: 1) an offline fingerprint construction phase and 2) an online tracking phase. During the offline phase, a probabilistic fingerprint is constructed, where the RSSI histogram for each cell tower at given locations in the area of interest is estimated. This is performed in the Radio Map Builder module.
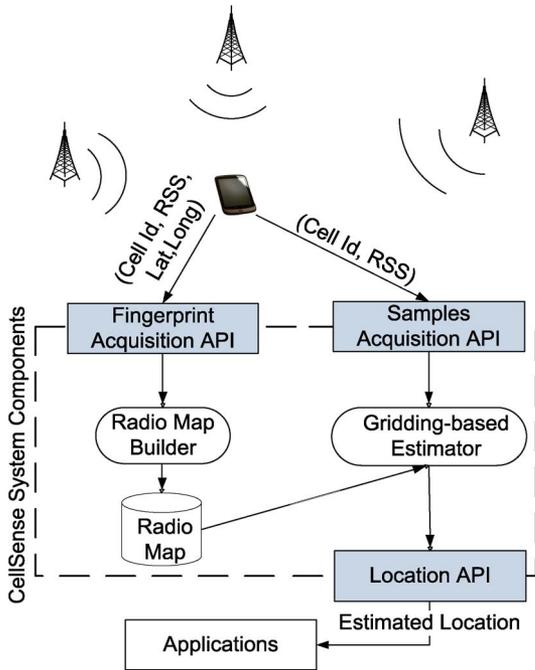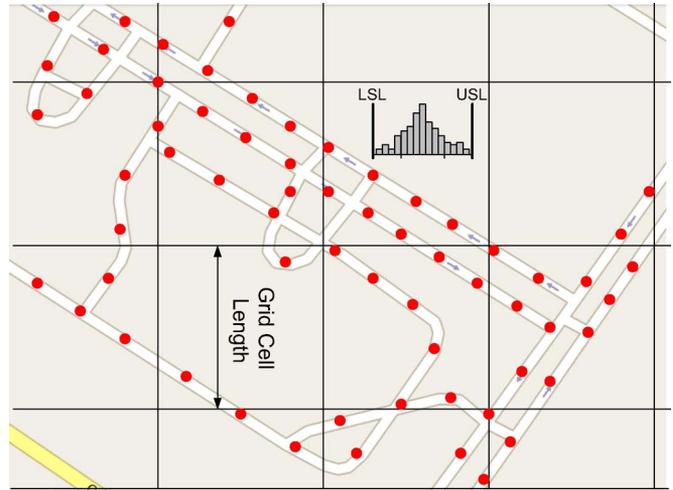
Fig. 2. CellSense approach for fingerprint construction. The area of interest is divided into grids, and the histogram is constructed using the fingerprint locations inside the grid cell. No extra overhead is required for fingerprint construction. The grid cell length parameter can be used to trade off accuracy and scalability.
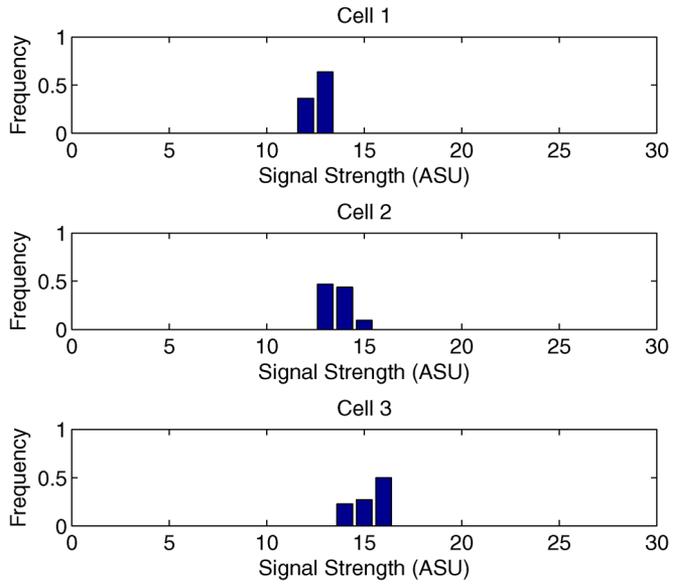

Fig. 3. Example of the histograms from three adjacent cells (grid length = 70 m) from a certain cell tower. The active set update (ASU) is an integer value returned by the phone API $(dBm = 2.ASU - 113)$.

Fig. 1. CellSense components. The arrows show information flow in the system.

During the online tracking phase, the location estimation module uses the fingerprint to calculate the most probable fingerprint location at which the user may be standing.

The RSSI samples are collected with the Fingerprint Acquisition API that interacts with the phone GSM modem to obtain RSSI information from up to seven neighboring cell towers, as indicated by the GSM standard. Finally, the Location API is used by the user's applications to query the current estimated user's location.

B. Mathematical Model

Without loss of generality, let $\mathbb{L}$ be a 2-D physical space. Let q represent the total number of cell towers in the system. We denote the q-dimensional signal strength space as $\mathbb{Q}$. Each element in this space is a q-dimensional vector whose entries represent the RSSI readings from a different cell tower. We refer to this vector as s. We also assume that the samples from different towers are independent. Therefore, the problem becomes, given an RSSI vector, $s = (s_1, \ldots, s_q)$, and we want to find the location $l \in \mathbb{L}$ that maximizes the probability $P(l|s)$.

C. Offline Phase

The purpose of this phase is to construct the signal strength histogram for the RSSI received from each cell tower at each location in the fingerprint. Typically, this requires the user to stand at each location in the fingerprint for a certain period of time to collect enough samples to construct the RSSI histogram. This will increase the fingerprint construction overhead significantly as the war-driving car has to stop at each location in the fingerprint for a certain time.

To avoid this overhead, we use a gridding approach, where the war-driving process is performed normally, and the area of interest is divided into cells. The histogram is then constructed for each cell tower in a given cell using all fingerprint points inside the cell rather than for each individual fingerprint point (see Fig. 2). Note that this gridding approach reduces the resolution of the fingerprint from individual points to cells with a certain size. The center of mass of all fingerprint points inside a grid cell is used to represent the cell.[1] Fig. 3 shows the histograms for a certain cell tower in three adjacent cells. The figure shows that the shape of the histogram changes over

---

[1]We use the term "fingerprint point" to refer to an individual point collected by the wardriving car and use the term "fingerprint cell" to denote the fingerprint collected using all points inside a given cell.

the different grid cells and, hence, could be used to distinguish between them.

The gridding approach not only removes the extra overhead of war driving but increases the scalability of CellSense as well as the fingerprint size can be arbitrarily reduced by increasing the cell size. We quantify the effect of the grid cell length parameter on performance in Section IV.

D. Online Phase

During the online phase, the user is standing at an unknown location l receiving a signal strength vector $s = (s_1, \ldots, s_q)$ containing one entry for each cell tower. We want to find the location in the fingerprint ($l \in \mathbb{L}$) that has the maximum probability given the received signal strength vector s. That is, we want to find

$$\text{argmax}_l \left[P(l|s) \right]. \tag{1}$$

Using Bayes' theorem and assuming that all the locations are equally probable,[2] this can be written as

$$\text{argmax}_l \left[P(l|s) \right] = \text{argmax}_l \left[P(s|l) \right]. \tag{2}$$

$P(s|l)$ can be calculated using the histograms constructed during the offline phase as

$$P(s|l) = \prod_{i=1}^{q} P(s_i|l). \tag{3}$$

The foregoing equation considers only one sample from each stream for a location estimate. In general, a number of successive samples N_s from each stream can be used to improve performance.

In this case, $P(s|l)$ can then be expressed as follows:

$$P(s|l) = \prod_{i=1}^{q} \prod_{j=1}^{N_s} P(s_{i,j}|l) \tag{4}$$

where $s_{i,j}$ represents the jth sample from the ith stream. Thus, given the signal strength vector s, the gridding-based estimator applies (4) to calculate $P(s|l)$ for each location l and returns the location that has the maximum probability.

Similarly, instead of returning just the most probable location, a weighted average of the K most probable fingerprint cells, weighted by the probability of each location, can be used to obtain a better estimate of location. We study the effect of the parameter K on performance in Section IV.

E. CellSense-Hybrid Technique

For the described CellSense technique, the grid cell length parameter allows us to trade accuracy and computational complexity: Larger cells lead to lower accuracy, but they reduce the computational complexity due to the reduced number of cells. The CellSense-Hybrid technique targets maintenance of the

accuracy at lower grid sizes while reducing the computational requirements. To achieve both accuracy and low complexity, the CellSense-Hybrid technique runs in two phases: 1) rough estimation phase and 2) refinement phase.

1) In the first phase (rough estimation phase), it uses the standard probabilistic fingerprint estimation technique to obtain the most probable cell in which a user may be located. However, instead of returning the center of mass of the fingerprint points inside this cell as the estimated location as in the standard CellSense, it refines this estimate in the second step.

2) In the second phase (estimation refinement phase), a KNN approach is used to estimate the closest fingerprint point, in the signal strength space, to the current user location inside the cell estimated in phase one. Note that since the histograms are constructed for an entire cell, we do not use a probabilistic technique in the second phase.

To achieve a low computational cost at low values of the grid cell length parameter, the CellSense-Hybrid technique uses only one sample to estimate the most probable cell rather than N_s samples in its first phase. The refinement phase allows it to compensate for the lost accuracy. Note that the CellSense-Hybrid technique does not have an advantage, in terms of computational complexity, for higher grid cell lengths as the number of fingerprint points involved in the second phase will dominate the computational cost.

In summary, the low computational requirement of the CellSense-Hybrid technique is achieved by using fewer number of samples in the estimation process as compared with CellSense. To compensate for the reduced accuracy, CellSense-Hybrid uses an estimation refinement phase. This allows CellSense-Hybrid to achieve both high accuracy and low computational requirements for low values of the grid cell length parameter. We quantify the performance of the hybrid technique in Section IV.

IV. PERFORMANCE EVALUATION

In this section, we study the effect of different parameters on CellSense and compare its performance with other RSSI-based GSM localization systems in terms of localization accuracy and running time. For the running time estimation, all the techniques have been implemented on a Dell Inspiron 6400 with a 1.83-GHz Intel Core 2 processor running Windows XP.

A. Data Collection

We collected data for two different testbeds. The first testbed covers the Smart Village in Cairo, Egypt, which represents a typical rural area. The second testbed covers 5.45 km^2 in Alexandria, Egypt, representing a typical urban area. Data were collected using T-Mobile G1 phones that have a GPS receiver (used as the ground truth for location) and running the Android 1.6 operating system. The experiment was performed using three phones, each with a subscriber identity module (SIM) card for a different cellular provider in Egypt.

We implemented the scanning program using the Android SDK. The program records the cell ID, signal strength, GPS

---

[2]If the probability of being at each location is known, this can be used in the equation.

TABLE I
COMPARISON BETWEEN THE TWO TESTBEDS. THE TRAINING SET SIZE REFERS TO THE NUMBER OF SAMPLES COLLECTED BY THE WAR DRIVING PROCESS. THE AVERAGE FINGERPRINT DENSITY IS THE AVERAGE NUMBER OF FINGERPRINT POINTS INSIDE A CELL FOR GRID CELL LENGTH = 70 m

Testbed	Area covered (Km^2)	Trace length (Km)	Network provider	Average calibration time(min.)	Avg. fgrprt dens/ cell	Total num. of cell towers	Training set size	Test set size	Avg. num. towers / loc.	Avg. num. of towers / Km^2	Avg. num. of towers /Km
One (Rural)	1.958	13.64	Provider 1	26.65	10.89	51	1599	573	5.63	26.05	3.05
			Provider 2	17.88	9.17	11	1073	594	1.00	5.62	1.04
			Provider 3	22.51	12.79	59	1351	592	4.83	30.13	4.33
Two (Urban)	5.450	18.27	Provider 1	51.5	11.22	137	3090	1239	5.35	25.13	6.97
			Provider 2	48.9	11.07	121	2934	1560	6.01	22.20	6.56
			Provider 3	45.06	12.12	155	2704	1564	5.32	28.44	9.28

TABLE II
DEFAULT VALUES FOR THE PARAMETERS. THESE VALUES ACHIEVE THE BEST PERFORMANCE

Parameter	CellSense (Best accuracy)	CellSense-Hybrid (**Best timing**)	Deterministic (Best accuracy)	Gaussian Processes (Best accuracy)
Rural testbed	Grid size=70, $N = 14$, $K = 2$	Grid size=70, $K = 1$	Grid size=70, $K = 8$	$N_p = 1019$
Urban testbed	Grid size=70, $N = 8$, $K = 2$	Grid size=70, $K = 1$	Grid size=90, $K = 6$	$N_p = 573$

location, and timestamp for the cell tower to which the mobile is connected, as well as the other six neighboring cell towers information as dedicated by the GSM specifications. The scanning rate was set to 1/s. Two independent data sets were collected for each testbed: one for training and the other for testing. Table I summarizes the two testbeds.

The calibration process took on average 22.34 min for the rural area and 48.48 min for the urban area. The war-driving process involved visiting each point only once. Our experience shows that visiting the same point more than one time does not lead to enhancement in accuracy.

B. Effect of Changing Parameters

In this section, we explore the effect of changing the different parameters on the performance of CellSense, mainly, grid cell length, number of samples used in estimation N_s, and number of most probable locations averaged to obtain the final location K. We also study the effect of changing the network provider, cell tower density, and effect of using a sparse radio map. Table II summarizes the parameters and their default values, which are the values that achieve the best performance.

1) Effect of Grid Cell Length: Fig. 4 shows the effect of changing the grid cell length on the median localization error. Each cell is a square with size as indicated on the x-axis. The figure shows that as the cell size increases, the accuracy decreases. This is because as the grid cell length increases, the points inside a cell become further away from its centroid, increasing the estimation error.

The figure also shows that a grid cell length of up to $200 \, m^2$ gives comparable accuracy to very small cell sizes for both testbeds. This indicates that CellSense can lead to good scalability with minimal reduction in accuracy. Moreover, the figure shows that the accuracy in urban areas is better than the accuracy in rural areas for grid cell length of up to 450 m due to the increased cell tower density. Increasing the grid cell length beyond this value leads to a significant drop in performance for the urban testbed. We believe that this is due to the fact that cell towers are configured to have a smaller range in urban
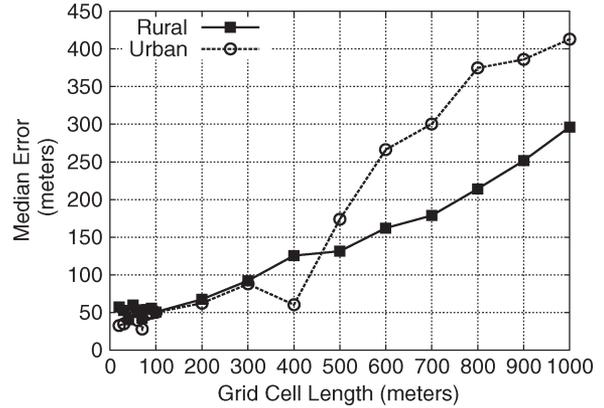

Fig. 4. Effect of changing the grid cell length on CellSense's median error.

areas. Increasing the cell length size beyond a certain value makes some cell towers not cover an entire cell, increasing the ambiguity between cells and reducing accuracy.

2) Effect of the Number of Samples Used N_s: Fig. 5 shows the effect of changing the number of samples used in estimation N_s on the median localization error. The figure shows that as the number of samples used in estimation increases, the accuracy increases until it reaches an optimal value ($N_s = 8$ and $N_s = 14$ for rural and urban testbeds, respectively) and then decreases. This is due to two opposing factors. 1) As we increase the number of samples, we have more information to estimate the user location, and hence, we should get better accuracy. 2) However, as we increase the number of samples, the time to collect these samples increases, which leads to crossing the boundary of one cell when using a large number of samples. This has a negative effect on accuracy.

The optimal point in rural areas occurs at lower N_s compared with urban areas due to the fact that the user speed is higher in rural areas than in urban areas.

3) Effect of the Number of Averaged Fingerprint Locations K: Fig. 6 shows the effect of changing the number of the most probable locations averaged K on the median localization error. The figure shows that, in general, the performance is enhanced
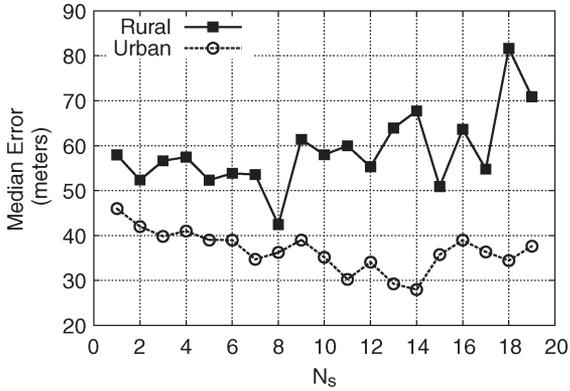
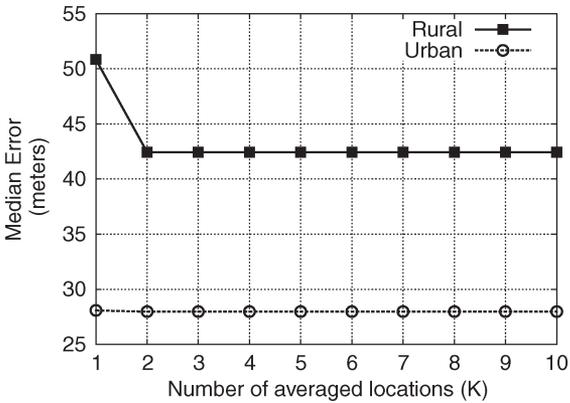Fig. 5. Effect of changing the number of samples N_s on CellSense's median error.

Fig. 6. Effect of changing the number of most probable locations averaged K on CellSense's median error.
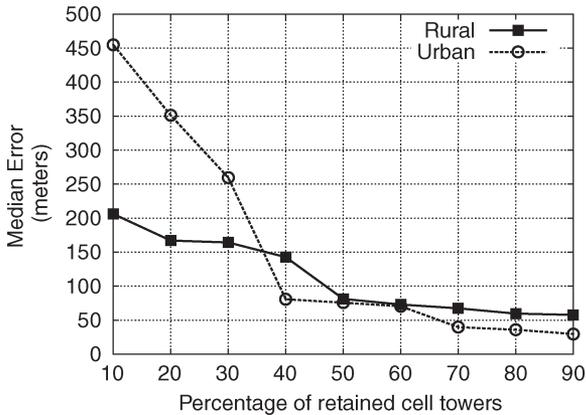

Fig. 7. Effect of changing the cell tower density on CellSense's median error.

Fig. 8. Effect of reducing the average number of data point per cell CellSense's median error.

Fig. 9. Effect of using different network providers on CellSense's median error. (a) Testbed 1 (Rural). (b) Testbed 2 (Urban).

as K increases until it saturates. This also highlights that the most probable location estimate has good accuracy.

4) Effect of Changing the Cell Towers Density: Fig. 7 shows the effect of changing the cell tower density on the median localization error. This was achieved by dropping a certain percentage of the cell towers, as indicated in the figure. The figure shows that as the cell tower density increases, the accuracy increases.

Effect of Decreasing the Radio Map Density: Fig. 8 shows the effect of decreasing the fingerprint density on the median localization error. The figure shows that as the percentage of
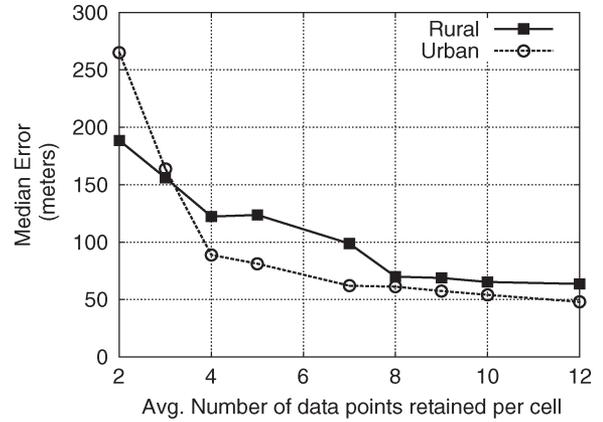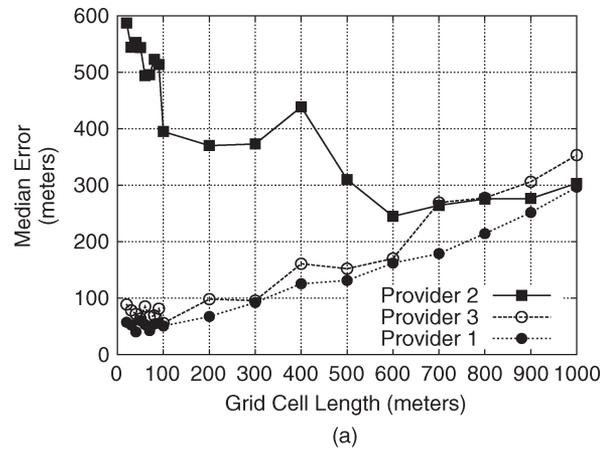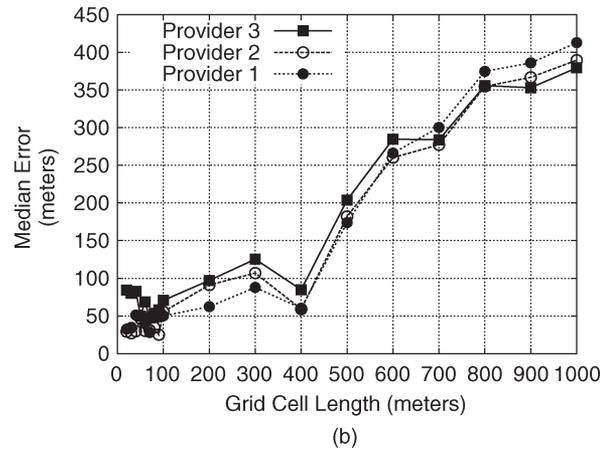
retained samples increases, the accuracy increases. The figure also shows that collecting only eight points per cell is enough to obtain good accuracy for both testbeds. In addition, the effect of reducing the fingerprint density is less than the effect of reducing the cell tower density.

Effect of Using Different Network Providers: Fig. 9 shows the effect of using different network providers in rural and urban areas. The figure shows that the accuracy of the provider is proportional to its cell tower density, as reported in Table I.
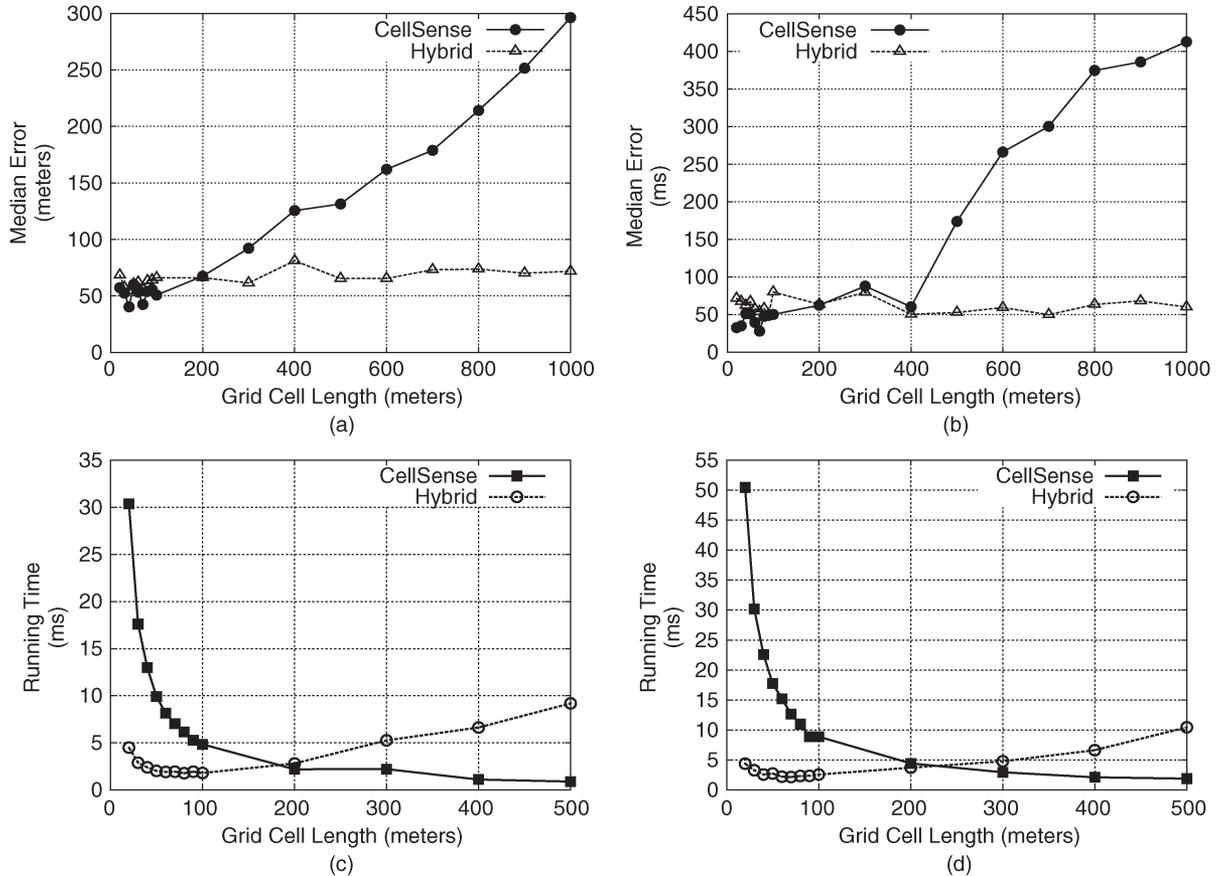
Fig. 10. Comparison between CellSense-Hybrid and CellSense techniques under the two testbeds. (a) Testbed 1 (rural): median error. (b) Testbed 2 (urban): median error. (c) Testbed 1 (rural): average running time/location estimation. (d) Testbed 2 (urban): average running time/location estimation.

The noticeable difference between Provider 2 and the other two providers in the rural testbed is due to its significantly lower cell tower density per location (1.00 as compared with 4.83 and 5.63). In addition, Provider 2's performance increases with the increase of the grid cell length until it reaches an optimal point at 600 m and then decreases again. This is due to two opposing factors. 1) As the grid cell length increases, we have more samples to construct the histogram, leading to better histograms and accuracy. 2) As the grid cell length increases, the fingerprint density decreases, and accuracy decreases. This behavior is not noticed with the other providers as they have a higher cell tower density that makes the second factor the dominating factor.

C. Results for the Hybrid Technique

In this section, we compare the performance of the CellSense-Hybrid technique described in Section III-E to the basic CellSense technique. The CellSense-Hybrid technique mixes both CellSense and a deterministic technique in its two phases. Fig. 10 shows that the accuracy of CellSense degrades as the grid cell length increases since the points inside a cell become further away from its centroid, increasing the estimation error and reducing accuracy. On the other hand, the CellSense-Hybrid technique has a robust performance, in terms of accuracy, for different grid sizes under the two testbeds. This is due to the estimation refinement phase.

The figure also shows that the running time of the CellSense technique decreases quadratically with the cell size. On the other hand, there are two factors affecting the running time of the CellSense-Hybrid technique. 1) As the grid size increases, the number of cells decreases, and hence, the running time of the first phase of the algorithm decreases. 2) However, as the grid size increases, the number of fingerprint points inside a cell increases and consequently the time for the second phase of the algorithm. This leads to the minimum point for the running time at $G = 70$ in the figure.

D. Comparison With Other Techniques

In this section, we compare the performance of the CellSense and CellSense-Hybrid techniques in terms of running time, localization error, and complexity to other RSSI-based GSM localization techniques described in Section II-F. Table II summarizes the parameters that achieve the best performance for all techniques. For the percentage enhancement numbers, our reference is the technique that achieves the best value. Therefore, we used CellSense as the reference in accuracy and CellSense-Hybrid as the reference in running time.

1) Localization Error: Fig. 11 shows the cumulative distribution function (CDF) of distance error for the different algorithms for the two testbeds. Table III summarizes the results. The table shows that our proposed techniques are better than any other technique with at least 108.57% in rural areas and
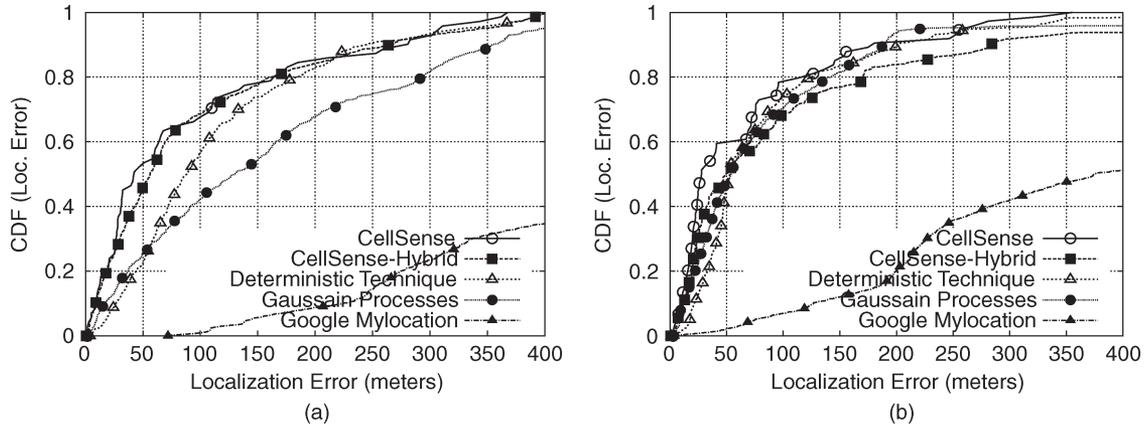
Fig. 11.   CDFs of distance error for different techniques under the two testbeds. The tails of the cdfs are truncated for clarity of presentation. (a) Testbed 1 (rural). (b) Testbed 2 (urban).

TABLE III
COMPARISON BETWEEN DIFFERENT TECHNIQUES USING THE TWO TESTBEDS. NUMBERS BETWEEN PARENTHESES REPRESENT PERCENTAGE DEGRADATION COMPARED WITH THE REFERENCE TECHNIQUE (THE BEST TECHNIQUE). q IS THE NUMBER OF CELL TOWERS. N_s IS THE NUMBER OF SUCCESSIVE SAMPLES USED IN ESTIMATION. N_p IS THE NUMBER OF PRECOMPUTED POINTS IN THE GP TECHNIQUE. N_c IS THE NUMBER OF GRID CELLS. N_0 IS THE NUMBER OF SAMPLES IN THE MOST PROBABLE CELL

Algorithms	Google MyLocation	Deterministic	Gaussian Processes	*CellSense*	*CellSense-Hybrid*
Testbed 1 (Rural) Median Error (m)	656.37 (1446.94%)	88.50 (108.57%)	130.50 (207.56%)	42.43 (**Reference**)	54.4 (32.33%)
95th percentile (m)	2767.6	328.74	401.6	308.56	354.36
Testbed 2 (Urban) Median Error (m)	374.58 (1244.51%)	52.74 (89.30%)	53.92 (93.54%)	27.86 (**Reference**)	56.19 (101.69%)
95th percentile (m)	3927.8	280.06	227.68	273.14	290.45
Testbed 1 (Rural) Avg. time/loc. (ms)	12.04 (580.22%)	10.73 (506.21%)	2873.41 (162240%)	7.04 (297.74%)	1.77 (**Reference**)
Testbed 2 (Urban) Avg. time/loc. (ms)	12.04 (462.62%)	11.56 (440.19%)	3148.35 (147019%)	12.66 (491.59%)	2.14 (**Reference**)
Complexity	$O(1)$	$O((qN_s + K)N_c)$	$O(qN_p)$	$O((qN_s + K)N_c)$	$O(qN_c + (q + K)N_0)$

at least 89.30% in urban areas. All techniques perform better in urban areas than rural areas due to the higher density of cell towers and the more differentiation between fingerprint locations due to the dense urban area structures. This is excluding the CellSense-Hybrid technique, whose accuracy is consistent between the two testbeds. The loss of accuracy of the CellSense-Hybrid technique, as compared with CellSense, comes at significant gains in running time, as quantified in the next section.

2) Running Time: Fig. 12 compares all the algorithms in terms of the average time required for one location estimate. Table III summarizes the results. The results show that the proposed techniques significantly outperform the other techniques by at least 506.21% in rural areas and at least 440.9% in urban areas. All the techniques take more time on average in the urban testbed than in rural testbed due to the increase in the number of cell towers. The cell-ID-based technique, i.e., Google MyLocation, has a consistent running time as it depends on the associated cell tower ID only. Although its time involves communicating with Google servers over the network, the average running time is much less than a typical network delay. We believe that this is due to the fact that the Location API on the phone returns a cached location, as long as the associated cell tower does not change. The GP approach is the

most demanding technique in terms of the running time. The CellSense-Hybrid technique provides about three to five times enhancement in running time compared with the CellSense technique.

3) Complexity Analysis: In this section, we analyze the algorithmic complexity of all the techniques. Table III summarizes the results.

1) Google MyLocation: Google MyLocation is a cell-ID-based technique. It has $O(1)$ complexity as it is probably a hash table lookup for the location of the cell tower the phone is connected to. However, we have no more details from Google to confirm our hypothesis.

2) CellSense: To compute the probability of each grid cell, we need $O(qN_sN_c)$ operations. Computing the weighted average of the most probable K locations, using an order statistics algorithm, requires $O(KN_c)$ for small K. Therefore, we need $O((qN_s + K)N_c)$ operations in total for each location estimate.

3) Deterministic technique: Similar to the CellSense technique, it requires $O((qN_s + K)N_c)$.

4) GPs: To compute the probability of each precomputed point, we need $O(tN_p)$. Computing the weighted average of all the precomputed locations requires $O(N_p)$
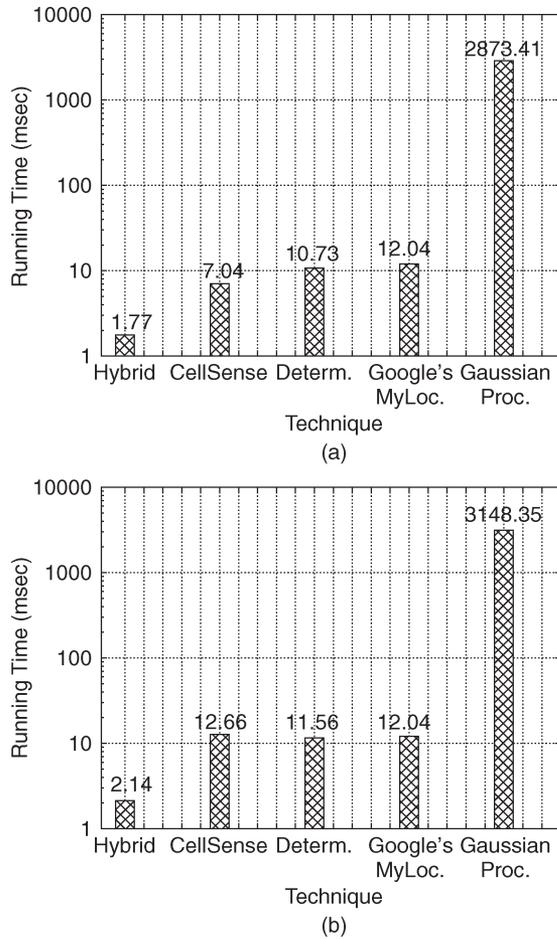
Fig. 12. Running time for different techniques under the two testbeds (log scale). (a) Testbed 1 (rural). (b) Testbed 2 (urban).

operations. Therefore, the overall all algorithm requires $O(tN_p)$ per location estimate.

5) CellSense-Hybrid: Calculating the probability of each grid cell in the first phase takes $O(qN_c)$. To apply the KNN algorithm in the second phase inside the most probable cell, we need $O((q + K)N_0)$. Therefore, we need $O(qN_c + (q + K)N_0)$ operations in total for each location estimate.

Since N_p is typically $\gg N_c$ (see Table II), to achieve reasonable accuracy, the GP approach is very slow compared with other non-cell-ID-based techniques (see Fig. 12).

Comparing CellSense to CellSense-Hybrid, we note that for a low grid cell length (the typical operation scenario for CellSense-Hybrid), N_c (number of grid cells) is $\gg N_o$ (number of samples inside a cell). Therefore, the CellSense-Hybrid computational overhead is much lower. The opposite is true for high grid cell sizes.

E. Summary

In this section, we evaluated the performance of the proposed CellSense and CellSense-Hybrid techniques. Our results show that the CellSense-Hybrid technique has comparable accuracy to the CellSense technique with significantly lower computational complexity compared with other techniques.

For the CellSense technique, as the grid cell size increases, the performance degrades. Increasing the number of samples used in the estimation or the number of averaged most probable locations has a positive effect on accuracy. Increasing the cell tower density has a more positive effect on accuracy than increasing the density of the fingerprint. The good news is that, although we do not have control on the cell tower density, reducing the fingerprint density by up to 60% still gives good accuracy.

The performance of the CellSense-Hybrid technique is consistent over different grid sizes and testbeds. This is due to the estimation refinement phase.

The accuracy of the localization technique under a certain cellular provider is correlated with the provider cell tower density.

Typically, there is always a tradeoff between computational overhead and accuracy. However, the CellSense-Hybrid technique provides a good balance between both accuracy and complexity. Its high accuracy comes from its ability of returning one of the original fingerprint points rather than the center of mass of all the locations inside the most probable cell. Its computational advantage at small grid sizes, compared with the CellSense technique, comes from using only one sample in the first phase as compared with N_s samples.

V. CONCLUSION

We have proposed CellSense, which is a probabilistic RSSI-based fingerprinting approach for GSM cell phone localization. We presented the details of the system and how it constructs the probabilistic fingerprint without incurring any additional overhead. We also proposed a hybrid approach that combines probabilistic and deterministic techniques to achieve both high accuracy and low computational requirements.

We implemented our system on Android-based phones and compared it to other GSM-localization systems under two different testbeds. Our results show that the CellSense-Hybrid technique accuracy is better than other techniques with at least 108.57% in rural areas and at least 89.03% in urban areas with more than 5.4 times savings in running time compared with state-of-the-art RSSI-based GSM localization techniques. We also studied the effect of different parameters on the performance of the system and how the cell tower density and fingerprint density affect accuracy.

Currently, we are working on extending our system in different directions, including using parametric distributions, clustering of fingerprint locations, experimenting with larger data sets, comparison with other city-wide commercial systems, targeting low-end phones [21], among others.

REFERENCES

[1] M. Ibrahim and M. Youssef, "CellSense: A probabilistic RSSI-based GSM positioning system," in *Proc. GLOBECOM*, 2010, pp. 1–5.
[2] P. Enge and P. Misra, "Special issue on GPS: The global positioning system," *Proc. IEEE*, vol. 87, no. 1, pp. 3–15, Jan. 1999.
[3] S. Tekinay, "Special issue on wireless geolocation systems and services," *IEEE Commun. Mag.*, vol. 36, no. 4, Apr. 1998.
[4] Y.-C. Cheng, Y. Chawathe, A. LaMarca, and J. Krumm, "Accuracy characterization for metropolitan-scale wi-fi localization," in *Proc. 3rd Int. Conf. Mobile Syst., Appl. Services MobiSys*, New York, 2005, pp. 233–245.

[5] I. Smith, J. Tabert, A. Lamarca, Y. Chawathe, S. Consolvo, J. Hightower, J. Scott, T. Sohn, J. Howard, J. Hughes, F. Potter, P. Powledge, G. Borriello, and B. Schilit, "Place lab: Device positioning using radio beacons in the wild," in *Proc. 3rd Int. Conf. Pervasive Comput.*, 2005, pp. 116–133.

[6] Skyhook Wireless. [Online]. Available: http://www.skyhookwireless.com

[7] R. R. C. Ionut Constandache and I. Rhee, "Towards mobile phone localization without war-driving," in *Proc. IEEE Infocom*, 2010, pp. 1–9.

[8] R. S. A. Offstad, E. Nicholas, and R. R. Choudhury, "AAMPL: Accelerometer augmented mobile phone localization," in *Proc. ACM MELT Workshop (With Mobicom)*, 2008, pp. 13–18.

[9] I. C. Martin Azizyan and R. R. Choudhury, "SurroundSense: Mobile phone localization via ambience fingerprinting," in *Proc. ACM MobiCom*, 2009, pp. 261–272.

[10] "Wikipedia," Comparison of Mobile Phone Standards—Wikipedia, the Free Encyclopedia, 2010. [Online]. Available: http://en.wikipedia.org/wiki/Comparison_of_mobile_phone_standards

[11] M. Y. Chen, T. Sohn, D. Chmelev, D. Haehnel, J. Hightower, J. Hughes, A. Lamarca, F. Potter, I. Smith, and A. Varshavsky, "Practical metropolitan-scale positioning for GSM phones," in *Proc. 18th Int. Conf. Ubiquitous Comput.*, 2006, pp. 225–242.

[12] A. Varshavsky, M. Y. Chen, E. de Lara, J. Froehlich, D. Haehnel, J. Hightower, A. LaMarca, F. Potter, T. Sohn, K. Tang, and I. Smith, "Are GSM phones THE solution for localization?" in *Proc. 7th IEEE WMCSA*, Washington, DC, 2006, pp. 34–42.

[13] E. Elnahrawy, J. Austen-Francisco, and R. P. Martin, "Adding angle of arrival modality to basic RSS location management techniques," in *Proc. IEEE ISWPC*, 2007, pp. 464–469.

[14] E. Elnahrawy, J.-A. Francisco, and R. P. Martin, "Poster abstract: Bayesian localization in wireless networks using angle of arrival," in *Proc. 3rd ACM Conf. Embedded Netw. SenSys*, 2005, pp. 272–273.

[15] P. Biswas, H. Aghajan, and Y. Ye, "Integration of angle of arrival information for multimodal sensor network localization using semidefinite programming," in *Proc. 39th Asilomar Conf. Signals, Syst. Comput.*, 2005.

[16] M. Li and Y. Lu, "Angle-of-arrival estimation for localization and communication in wireless networks," in *Proc. 16th EUSIPCO*, 2008.

[17] Google Maps for Mobile. [Online]. Available: http://www.google.com/mobile/maps/

[18] M. Youssef, M. A. Yosef, and M. N. El-Derini, "GAC: Energy-efficient hybrid gps-accelerometer-compass gsm localization," in *Proc. GLOBECOM*, 2010, pp. 1–5.

[19] M. Youssef and A. Agrawala, "Small-scale compensation for WLAN location determination systems," in *Proc. IEEE WCNC*, 2003, vol. 3, pp. 1974–1978.

[20] M. A. Youssef, A. Agrawala, and A. U. Shankar, "WLAN location determination via clustering and probability distributions," in *Proc. 1st IEEE Int. Conf. PERCOM*, 2003, p. 143.

[21] M. Ibrahim and M. Youssef, "A hidden Markov model for localization using low-end GSM cell phones," in *Proc. ICC*, 2011.

Mohamed Ibrahim (S'11) received the B.Sc. degree in computer science from Alexandria University, Alexandria, Egypt, in 2009 and the M.Sc. degree in wireless technology from Nile University, Cairo, Egypt, in 2011. He is currently working toward the Ph.D. degree with University of Technology of Troyes, Troyes, France.

His research interests include location determination technologies, sensor networks, and pattern recognition.

Moustafa Youssef (S'01–M'04–SM'09) received the Ph.D. degree in computer science from the University of Maryland, College Park, in 2004 and the B.Sc. and M.Sc. degrees in computer science from Alexandria University, Alexandria, Egypt, in 1997 and 1999, respectively.

He is currently an Assistant Professor with Alexandria University and Egypt-Japan University of Science and Technology, Alexandria. His research interests include location determination technologies, pervasive computing, sensor networks, and network security. He has eight issued and pending patents. He is an area editor of the Association for Computing Machinery Mobile Computing and Communications Review (MC2R) has served on the organizing and technical committees of numerous conferences, and has published over 70 technical papers in refereed conferences and journals.

Dr. Youssef received the 2003 University of Maryland Invention of the Year award for his Horus location determination technology and the 2010 Academy of Sciences for the Developing World (TWAS) African Academy of Sciences (AAS) Microsoft Award for Young Scientists, among others.