

Parsimonious Mobility Classification using GSM and WiFi Traces

Min Y. Mun*, Deborah Estrin*, Jeff Burke*†, Mark Hansen*

*Center for Embedded Networked Sensing,

†Center for Research in Engineering, Media and Performance

University of California, Los Angeles

{bobbymun, jburke}@ucla.edu, destrin@cs.ucla.edu, cocteau@stat.ucla.edu

ABSTRACT

Human mobility states, such as dwelling, walking or driving, are a valuable primary and meta data type for transportation studies, urban planning, health monitoring and epidemiology. Previous work focuses on fine-grained location-based mobility inference using global positioning system (GPS) data and external geo-indexes such as map information. GPS-based mobility characterization raises practical issues related to spotty coverage and battery drain, but the more fundamental concern addressed in this paper is that of privacy. For some applications and usage models we contend that it is desirable to adopt a more *parsimonious* approach to mobility characterization; one that avoids the collection and use of fine-grained location information by relying instead on GSM and WiFi connectivity data. Building upon previous work that demonstrated the utility of using GSM and WiFi beacons for localization applications, we demonstrate that this approach to mobility classification achieves promising results (with accuracy of 88%) using a sample data set collected across several populated areas in Los Angeles and is worthy of further research.

Categories and Subject Descriptors

I.5.1 [Pattern Recognition]: Models—deterministic, statistical

General Terms

Algorithms, Experimentation.

1. INTRODUCTION

Human mobility states, such as dwelling, walking or driving, are valuable primary and meta data types for transportation studies, urban planning, health monitoring and epidemiology. Substantial work exists on techniques to infer mobility state from time-series sensor data, with and without contextual models [2,11,12,17]. Much of this work uses data from GPS devices [2,11,12]. Recording such “fine-grained” location traces poses a privacy risk that may be unnecessary if mobility state can be satisfactorily inferred from “coarser-grained”

location. In this paper, we explore how visibility and signal strength of GSM cell towers and WiFi beacons, that is already available on standard mobile handsets, can be used to generate mobility profiles. Other practical limitations of continuous GPS sampling are also avoided, including reduced phone battery life, inconsistent coverage for typical users[10], and limited availability of integrated GPS in current mobiles phones.

Excellent results have been obtained in tracking and predicting a person’s location and movement using WiFi and GSM data[3,5,8,13-15,17]. However most prior work targets indoor environments with known access point and tower locations for localization or classifying a user as still or moving. In [17], a user is classified as stationary, walking or driving using GSM data, but does not attempt to leverage smaller cell-size data such as WiFi to supplement the limitations of large cell size data from GSM[9]. We focus on profiling unconstrained mobility throughout a typical day without a priori knowledge nor estimated location of access points that will be encountered(usually cell tower locations are known by cellular carriers and fairly static. This information can be used to roughly indicate a user’s location). Not only is this demanded by the target applications, it takes full advantage of how modern mobiles phones can collect data throughout the day in a collaborative participatory sensing usage model[4]. Exploring how mobility classification can be performed using only GSM and WiFi observations contributes to our ongoing work in participatory privacy regulation for urban sensing. For example, our Personal Environmental Impact Report(PEIR) project uses mobility-annotated location traces as input into models of hazard exposure and environmental impact[4]. The work described in this paper could enable PEIR to offer its participants the option of sharing only coarse-grained location and still receiving useful feedback.

We demonstrate using one urban data set how we can achieve an overall 88% classification accuracy in distinguishing between pedestrian movement (with accuracy of 90.17%) and vehicle movement (with accuracy of 87.73%). The basis of our success is the combination of data sources with complementary cell footprint sizes: GSM and WiFi beacons data(accuracy of using only GSM: 79%, accuracy of using only WiFi: 75%). We are not aware of any modeling approach that uses a combination of two coarse-grained traces to achieve mobility characterization.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

HotEmNets'08, June 2-3, 2008, Charlottesville, Virginia, USA
Copyright 2008 ACM ISBN 978-1-60558-209-2/08/0006 ...-\$5.00

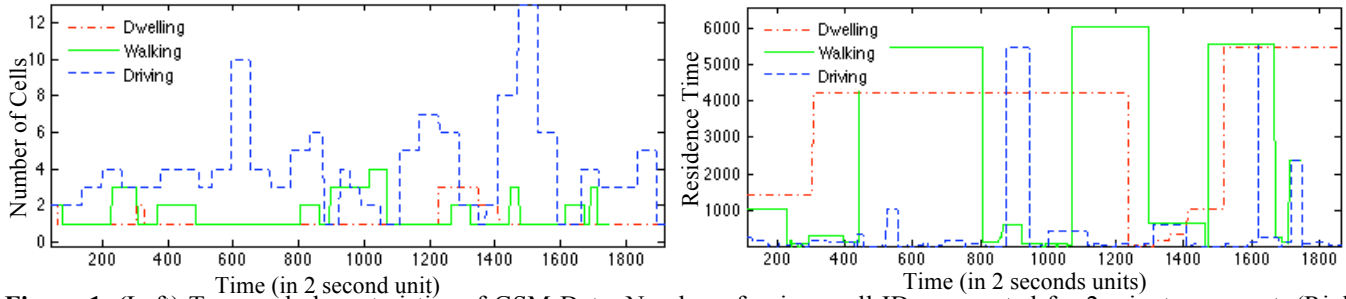


Figure 1. (Left) Temporal characteristics of GSM Data, Number of unique cell IDs connected for 2 minute segment, (Right) Spatial characteristics of GSM Data, Residence time (seconds) in each cell footprint for different states. Only partial data points for dwelling are shown to be compared with data for other states. Dwelling, walking and driving have different time and space features.

2. DATA COLLECTION

We use GSM and WiFi beacon data to build and evaluate a mobility classification scheme. Below we detail the nature of our data collection and useful features for mobility state differentiation.

2.1 Experimental Design

Our data set was collected by one of the authors for two days. We used the Nokia N95 “smart phone” as the mobile handset platform[21] and T-Mobile was our cellular carrier. A custom application was written in Python to collect data autonomously. Every two seconds¹, the application captures and records the cell tower ID, surrounding WiFi beacons and GPS location.

To obtain *in situ* ground truth mobility annotation, users recorded their mobility state, i.e., dwelling (remaining at one place for more than 5 minutes), walking or driving, whenever they change states. We observed in previous tests that users easily forgot to annotate mobility, especially if they collect data for an extended period of time. To help *post facto* annotation correction, we captured images every ten-seconds. Using this extra context information, we were able to fill in missing annotations. This image collection is not a part of the target application, rather a technique used to improve the ground truth annotation of our test data sets.

GSM and WiFi beacon densities are correlated with population density [13]. To cover areas having different GSM cell and WiFi beacon densities, we picked four differently-populated areas from the U.S. Census Bureau data in 2000. We collected data in LA Downtown (27,845 persons/Sq mile), Santa Monica promenade (16,647 persons/Sq mile), West Hollywood (10831 persons/Sq mile) and Howard Hughes Shopping Mall (3636 persons/Sq mile). For realistic data collection, participants went to common places within the designated areas such as retail stores and dining places. In general the specific model that was built works better in urban areas rather than in less populated or rural areas - we might need to capture additional data to handle these cases. This will be studied further as future work.

In total, we gathered 12.5 hours of data. 59% of data is annotated as dwelling, 20% walking and 21% driving.

2.2 GSM Data

GSM(Global System for Mobile communications) is the most popular standard for mobile phones in the world. As of January 2007, 82% of the global mobile market uses the standard and GSM is used by over 2 billion people with deployment in more than 212 countries [6]. A GSM base station is typically equipped with a number of directional antennas that define sectors of coverage or cells, each of which has uniquely identifiable cell ID by combination of ‘country code’, ‘network code’, ‘area code’ and ‘cell id.’ Information from the cell ID can provide a rough indication of a person’s position. The approach described in [17] uses information about multiple in-range cell tower to infer stationary, walking and driving. Unfortunately a large portion of current mobile devices do not have access to multiple cell tower information. In particular, mobiles using Symbian OS, which account for 67% of the ‘smart phone device’ market, only provide the primary cell id information[20]. Therefore, in our studies, we use single connected cell ID data.

2.2.1.1 Temporal Characteristics of GSM Data: Number of Unique Cell IDs

Speed recorded from GPS data has been used to infer mobility information [2,11,12]. The same principle can be applied to GSM data. As users move faster, they see more cells for a certain duration.

We divided data into 2 minute segments and counted the number of unique cell IDs to which the phone was connected. We chose a 2 minute segment size after running experiments with different segment sizes; 10, 20, 40, 60 seconds and 2, 3 minutes. The segment size may need to be adjusted in areas having different cell densities. For example, in rural areas, larger segment sizes would be more suitable. Figure 1(Left) shows the number of unique cell IDs associated with a phone within a segment for each of the ground truth mobility states: Dwelling state is associated with one cell over the course of 2 minutes, walking typically 1-2, and driving state has larger ranges and is associated with more than 3.

2.2.1.2 Spatial Characteristics of GSM Data: Residence Time in a Cell Footprint

We now consider the spatial characteristics of GSM data. Again the same principle applies: users pass through more cells when moving faster. We divided data into segments by cell IDs so that all the data points in a segment belong to the same cell ID. Residence time is computed for each cell. Cell

¹ In future study, we plan to vary this sampling rate to examine tradeoff between accuracy and power consumption

sizes may vary from meters to a few kilometers. This is based on a heuristic that the footprint of cells in one metropolitan area are relatively similar ; this needs further validation. We expected that dwelling state would have the longest residence time while driving would have the shortest. As shown in Figure 1(Right), driving state certainly has much shorter residence times in each cell. Often walking state has longer residence times than dwelling. By looking at the data, we found that residence time for dwelling state overlapped those for walking state when a single cell contains walking and dwelling mobility states.

2.3 WiFi Beacon

Given the nature of large GSM cell sizes, identifying the individual locations and mobility states within a single cell cannot be done well from the GSM cell data alone. Therefore we explore the use of data from networks with smaller cell sizes, such as WiFi. WiFi beacons are another good data source for localization and mobility inferences. Variance of signal strength from location of WiFi access points has been widely investigated for indoor localization problems [3,5,8,13,15]. WiFi access points are ubiquitous and have shorter range signal, thus very attractive data sources. Location of WiFi APs are not fixed. In the same space, users may see different WiFi APs at different times. Techniques relying on location of the AP alone can be unreliable. In the following subsections, we try to apply commonly used features to help with mobility classification and also derive other features.

2.3.1.1 Signal Strength Variance

Signal strength variance measures have been used to infer motion, either still or moving in [8,14]. This is based on the observation that when a WiFi receiver is moving, the signal strengths it observes are noisier than when it is not moving. We want to see if this feature is useful for our application and expect to see different noise levels for the three mobility states. We divided the data into 40 second segments and computed average of signal strength variation for the 3 strongest APs. Euclidean distance is used to measure signal strength variation between two data points. Note that a smaller size of segment is chosen with WiFi data due to smaller cell size, which improves accuracy of the model only with cell ID data. Figure 2(Top) shows signal strength variance for mobility states. Driving state has higher fluctuation up to 140 and is distinguished from other mobility states. However, different from our expectation, dwelling state often shows higher variance than walking. This is because dwelling state can involve a certain degree of mobility. For example, in a shopping mall, users are in motion much of the time. Also local structures found on the inside of buildings seems to contribute.

2.3.1.2 Duration of Dominant WiFi Access Point in View

Dwelling in one place can involve a certain degree of mobility as mentioned earlier and ambiguity between dwelling and other mobility states is an issue in our study. However, Mobility in dwelling state differs from other mobility states in that it does not aim in a particular direction. That is, it is more

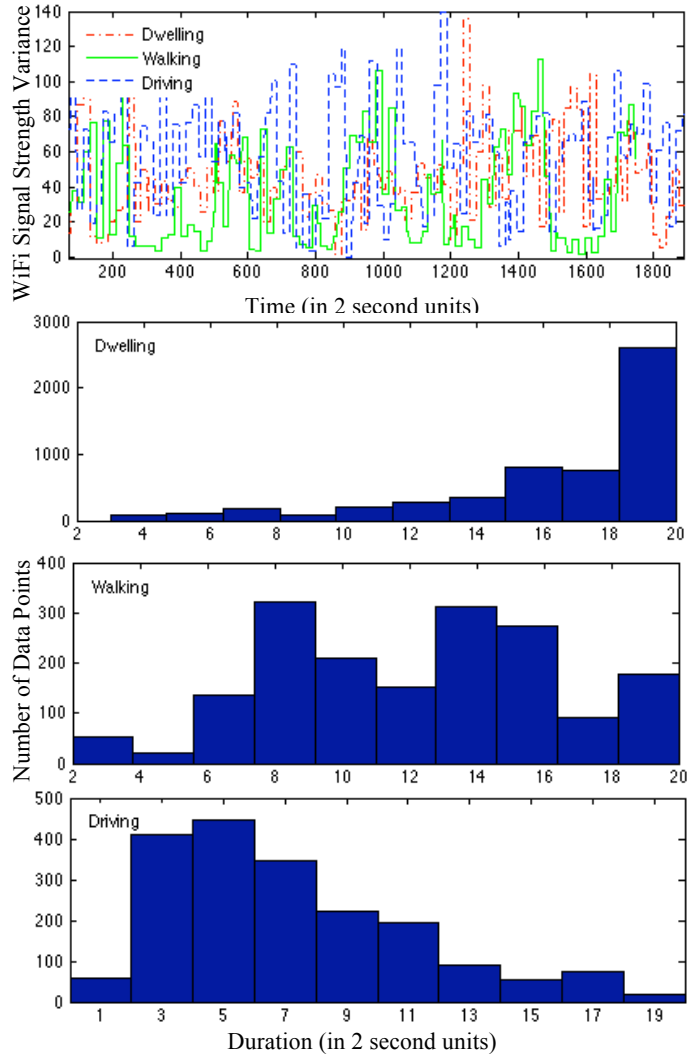


Figure 2. (Top) WiFi signal strength variance. Only partial data points for dwelling are shown to be compared with data for other states. Dwelling, walking and driving have different time and space features. (Bottom) Distribution of duration of dominant WiFi AP visibility during 40 seconds segment

likely that users dwell within the same WiFi AP footprint even though they are moving. A footprint is small and less likely to contain multiple mobility states. For dwelling state, many WiFi APs may be seen and the strongest WiFi AP may vary because of signal strength fluctuations. However, the dominant one is seen for the greatest amount of time. Walking state has “switch dominance” as users move from one place to another and the duration of the most dominant WiFi AP is relatively low. Switch dominance would happen more frequently in driving state. This is dependent on neither densities of WiFi nor on the number of available WiFi APs.

We chunk data points into 40 second segments and compute the amount of time that the most dominant WiFi AP in each segment is seen. As depicted in Figure 2(Bottom), the dominant WiFi AP is seen for the largest portion of time while dwelling. It usually appears fewer than 10 times during driving state. Walking state has a larger range, yet is

distinguishable.

3. MOBILITY CLASSIFICATION MODEL

We built a mobility classification model using four derived features from GSM and WiFi beacons data, and evaluated with a 12.5 hour sample dataset. The results of mobility classification without knowledge of the users location was promising. Moreover, we examined the primary causes to provide insight for future work.

3.1 Preprocessing

We discovered the so-called “pingpong” phenomenon in our data set with connected cell IDs, which was originally reported by [7,18] and used in [19]. That is, when a user is within the coverage of two or more cell towers, because of cell channel fluctuation, signal strength from the towers also fluctuates. This causes repetitive changes of “associated cell” even when users are stationary. It may be possible that users physically move from one cell to another. However, the intuition is that if a cell ID is repetitively and dominantly shown within a certain time of data, it is more likely that users dwell in one area. We replace alternative cell IDs with the dominant one, which boosts disambiguation of mobility. For example, a mobility classifier using only the number of unique cell tower IDs for two minute data segment characterize mobility 60.34% correctly without smoothing, but 70.18% with smoothing. Smoothing out the pingpong phenomenon boosts disambiguation of mobility. The pingpong phenomenon also appears in WiFi data, but smoothing was not necessary because available WiFi AP information was available and dominant WiFi is still visible.

Other preprocessing steps were used such as removing null data points, adding ground truth annotation and chunking into segments with different duration to compute features.

3.2 Inferring Mobility

The goal of our mobility model is to infer a user’s mobility, either dwelling, walking or driving. Each data point consists of cell ID, a list of available WiFi APs with signal strength, ground truth mobility annotation and timestamp. We smoothed out cell ID data and computed four features for each data point: number of unique cell IDs, residence time in a cell footprint with 2 minute segmentation and signal strength variance, duration of dominant WiFi AP in view with 40 second segmentation. We used four features to train a decision tree and pruned it using a standard heuristic that the decision tree has to work well across all dataset to prevent the model from over-fitting and rules of each attribute should be consistent across the tree. Figure 3 shows the structure of decision tree. All four features are useful to classify three mobility states. Note that driving state is predicted only if residence time is less than 600.5 seconds and other three features are used for further classification. Driving has greater signal strength variance and unique cell IDs than walking and dwelling, and walking does than dwelling. For evaluation, we randomly divided our data set into three and used three fold cross validation method.

In addition, we built a model to distinguish between pedestrian and vehicle movement using a reduced dataset

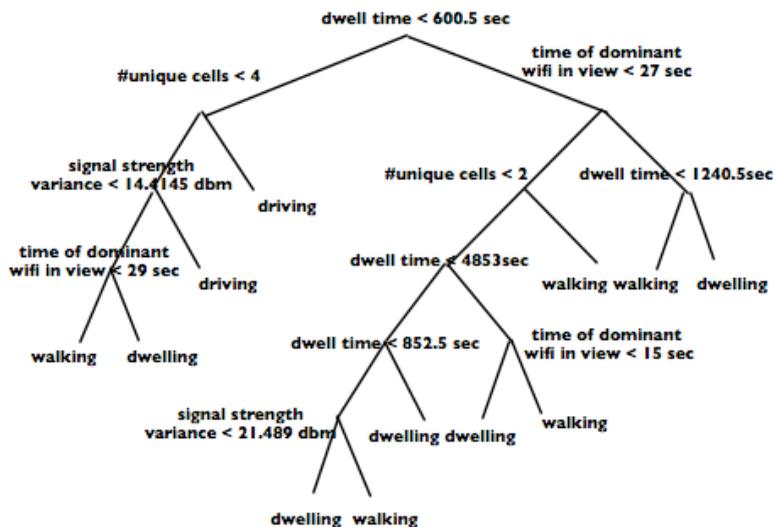


Figure 3. Structure of decision tree for dwelling, walking and driving classification

excluding indoor dwelling state and re-annotating outdoor dwelling as pedestrian. We applied the same procedures. A data point having greater duration of dominant WiFi AP in view, signal strength variance and more unique cell IDs was classified as vehicle movement.

3.3 Results and Discussion

Table 1 shows precision (true positive/(true positive + false positive)) and recall (true positive/(true positive + false negative)) confusion matrices. Precision is the percentage of correct predictions and recall is the percentage of correctly identified ground truth cases. Our overall accuracy (true positive/total number of data points) is the percentage of correctly classified data points, 83%.

Our classifier identified most of dwelling states with 90.26% recall and prediction is made 88.41% correctly. Walking states have the lowest accuracy with 55.40% recall and 65.45% precision, which is the major cause of errors of the model. This occurred because dwelling states are often incorrectly inferred as walking. The classifier performs quite well for characterizing driving states with 90.73% recall. Yet driving precision is relatively low. Mainly because walking states (12.65%) and dwelling states (11.61%) are incorrectly classified as driving.

We examine performances of the classification model on the walking state in more detail by comparing ground truth mobility states and classification results in Figure 4. In Figure 4-A, user dwelled at gas station and walked to nearby shop. These two mobility states happened within a single cell footprint, and WiFi features had to be used to distinguish further. But because of mobility during dwelling, signal strength variance failed to differentiate the two states. Furthermore, the gas station has weak WiFi access points available and short duration of dominant WiFi. Thus, dwelling states are classified as walking. In Figure 5-B, the user shopped at the mall, and periodically stopped and moved. We found that dwelling states mis-classified as walking for the same reason found earlier. The user walked during Figure 5-C and the dominant WiFi was also seen for a short duration.

Table 1. Precision and recall confusion Matrices

recall		prediction		
ground truth		dwelling	walking	driving
	dwelling	88.41%	0%	7.23%
	walking	6.67%	55.40%	2.03%
	driving	4.89%	44.60%	90.73%

precision		prediction		
ground truth		dwelling	walking	driving
	dwelling	90.26%	24.67%	11.61%
	walking	0%	65.45%	12.65%
	driving	9.74%	9.87%	75.73%

Walking was incorrectly inferred as driving.

Since movement during dwelling states often causes incorrect classification. We explored excluding indoor dwelling mobility states from our data set and building a classifier to infer either pedestrian or vehicle movement. As expected, overall performance of the model improved up to 88% accuracy. It identified pedestrian mode 90.17% and vehicle movement 87.73% correctly.

4. CONCLUSIONS AND FUTURE WORK

We proposed approach to parsimonious mobility classification. We demonstrated one such technique that successfully distinguishes meaningful differences in mobility states using only coarse grained location traces, GSM and WiFi beacons. This work is just a first exploration of this area. Much work is needed to evaluate its effectiveness using a larger set of data from multiple individuals and in a wider range of physical settings. The techniques applied here are just a first attempt. In future works we will study improvements to the base classifier, elaborations to the space of features, and conduct various experiments involving sampling rates.

5. ACKNOWLEDGMENTS

We acknowledge Joseph Kim, Sasank Reddy, Thomas Schoellhammer and Mohommad Ramini for their insightful comments and contributions.

6. REFERENCES

[1]E. Agapie, G. Chen, D. Houston, E. Howard, J. Kim, M. Y. Mun, A. Mondschein, S. Reddy, R. Rosaio, J. Ryder, A. Steiner, J. Burke, D. Estrin, M Hansen, M. Rahimi, Seeing Our Signals: Combining location traces and web-based models for personal discovery, *Hotmobile* 2008

[2]D. Ashbrook, T. Starner, Using GPS to learn significant locations and predict movement across multiple users, *Personal and Ubiquitous Computing* 2003

[3]P. Bahl, V.N.Padmanabhan, Radar: An in-building rf-based user location and tracking system, *In Proceedings of the IEEE Infocom*, 2000

[4]J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, MB Srivastava, Participatory Sensing, *World Wide Sensor Web Workshop at Sensys*, 2006

[5]Exahau Positioning System, <http://www.ekahau.com>

[6]GSM World Statistics, GSM Association,

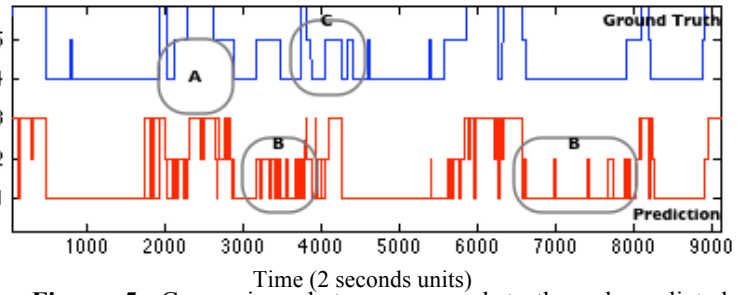


Figure 5. Comparison between ground truth and predicted mobility states. 1 and 4 on y axis represent dwelling, 2 and 5 walking and 3 and 6 driving. Notable misclassification examples are circled.

<http://www.gsmworld.com/gsmastats.shtml>

[7]T. Henderson, D. Kotz, and I. Abysoz. The change usage of a mature campus-wide wireless network. *In Proceedings of the Tenth Annual International conference on Mobile Computing and Networking (MobiCom)*, 2004

[8]J. Krumm, E. Horvitz, Locadio: Inferring Motion and Location from Wi-Fi Signal Strengths, *MobiQuitous*, 2004

[9]K. Laasonen, M. Raento, H. Toivonen, Adaptive On-Device Location Recognition, *Pervasive* 2004

[10]A. LaMarca, Y. Chawathe, S. Consolvo, J. Hightower, I. Smith, J. Scott, T. Sohn, J. Howard, J. Hughes, F. Potter, J. Tabert, P. Powledge, G. Borriello, B. Schilit, Place Lab: Device Positioning using Radio Beacons in the Wild, *Pervasive Computing*, 2005

[11]L. Liao, D. Fox, H. Kautz, Extracting Places and activities from GPS Traces Using Hierarchical Conditional Random Fields, *The International Journal of Robotics Research*, 2007

[12]L. Liao, D. Patterson, D. Fox, H. Kautz, Learning and inferring transportation routines, *Artificial Intelligence*, 2004

[13]K. Muthukrishnan, N. Meratnia, M. Lijding, G. Koprnikov, P. Havinga, WLAN location sharing through a privacy observant architecture, *COMSWARE by IEEE Communication Society Press*, 2006

[14]K. Muthukrishnan, M. Lijding, N. Meratnia, P. Havinga, Sensing Motion using Spectral and Spatial Analysis of WLAN RSSI, *Smart Sensing and Context*, 2007

[15]V. Otsason, A. Varshavsky, A. LaMarca, E. de Lara, Accurate GSM Indoor Localization, *Ubicomp*, 2005

[16]Schapire, R.E. The boosting approach to machine learning: An overview. In: D.D. Denison, M.H.Hansen, C.Holmes, B. Mallick, B. Yu, editors, *Nonlinear Estimation and Classification*. Springer, 2003

[17]T. Sohn, A. Varshavsky, A. LaMarca, M.Y. Chen, T. Choudhury, I. Smith, S. Consolvo, J. Hightower, W. Griswold, G. E. de Lara, Mobility Detection Using Everyday GSM Traces, *Ubicomp*, 2006

[18]L. Song, D. kotz, R. Jain, and X.He. Evaluating location predictors with extensive Wi-Fi mobility data. *In Proceedings of the 23rd Annual Joint Conference of the IEEE Computer and Communications Societies(INFOCOM)*, 2004

[19]J. Yoon, B. Noble, M. Liu, M. Kim, Building Realistic Mobility Models from Coarse-grained Traces, *Mobisys*, 2006

[20] <http://www.canalys.com/pr/2007/r2007024.htm>

[21] www.sensorplanet.org