Sensing in Rich Bluetooth Environments

Marion Hermersdorf, Heli Nyholm,
Jukka Salminen and Henry Tirri
Nokia Research Center
P.O. Box 407 FIN-00045 Finland

Jukka Perkiö and Ville Tuulos
Complex Systems Computation Group
Helsinki Institute for Information Technology
P.O.Box 68 FIN-00014 Finland

Abstract

The abundance of short-range radios, such as Bluetooth, in the current office and urban environments opens up new possibilities for data modeling and analysis. Although individual devices are constantly moving and active only sporadically, various environments have specific statistical characteristics. Mobile handsets are often personal in nature, so the statistics reflect also behavioral patterns of their users.

In this paper we first demonstrate how prototypical patterns of behavior may be found in data produced by Bluetooth scanning. It appears that cyclical nature of the patterns reflect well various daily routines of test persons. Our second model demonstrates that rich Bluetooth environments are often stable enough so that they can be used for locationing without any base stations.

Keywords

Bluetooth sensing, behavioral patterns, beaconless locationing

1 Introduction

In recent years the penetration of devices providing short range wireless connectivity has increased rapidly. The Bluetooth radio especially is largely integrated in mobile phones, PDAs and laptops. Every week 10 million Bluetooth radios are shipped, resulting in an estimated 550 million shipped Bluetooth devices in 2006 [11].

The short range radio can be used to connect to sensors distributed in the environment. Mobile phones, often providing the wireless connectivity, have the computing power as well as memory to collect, process and store various sensor data. Additionally mobile phones are perceived as personal devices mainly being at the same location as the users are.

The abundance of Bluetooth-enabled devices, plus their implicit role as personal identifiers, make Bluetooth sensing a valuable source of information. Numerous existing studies utilize the phone as part of a sensor network, for example in telemonitoring and telemedicine [8, 10], tracking and positioning [7, 1], context awareness [3] and social proximity sensing [13, 9].

In the following we demonstrate and explore the potential of rich Bluetooth environments. We will show that it is possible to derive complex behavioral patterns from the collected Bluetooth data, in the tradition of *Reality Mining* [9]. We will also present a probabilistic model for indoor locationing, related to ideas presented in [6] and [1]. In contrast to many previous approaches, our model does not require any stationary beacons. We conclude with some ideas for future work.

2 Data

The Bluetooth class 2 radio has a maximum range of 10 meters and therefore provides information of the Bluetooth devices in the proximity of the scanning device. An application was running on the mobile phone which scanned approximately every thirty seconds for other Bluetooth devices in the proximity. The scanning results were sent via GPRS to a backend server.

Fourteen mobile phones with this application were distributed to selected users working in an office building. The users were encouraged to carry this Bluetooth scanning phone, additional to their personal phone, with them while being in the building. We also asked the users to leave the phone over night in their office for recharging.

We placed fifteen passive Bluetooth beacons in the building, which provided us the ground-truth for positioning. The beacons were standard off-the-shelve USB Bluetooth dongles for PCs. Once started up, by connecting to a PC, the Bluetooth adapters were disconnected from the PC without interrupting the power supply (use of a battery pack) and placed at defined locations.

The Bluetooth scanning data of users were collected on the backend server for approximately 10 days. It was possible for us to monitor the data reception to ensure that all phones and beacons were operating properly. The data set consists of 2,867,167 rows of

$$(PhoneID, timestamp, MAC)$$

tuples. Some of detected MACs correspond to known locations (beacon MACs).

Even though our test site might not correspond to a typical office environment of today, the trend is obvious: Ur-

Table 1. Bluetooth Data Summary

Description	Quantity
Total number of BT scans	73 588
Total number of BT devices detected	854 926
Number of individual BT addresses detected	1 299
Average number of BT devices detected on one scan	11.6
Max. number of BT devices detected on one scan	52

ban areas are becoming increasingly covered by various uncontrolled short-range radios, such as Bluetooth, WiFi and ZigBee. Table 1 summarizes our data set, collected within four days by fourteen people in one building. One should note especially the remarkable number of unique BT addresses (1299) and the high average number of detected devices (11.6). We see that rich environments like this provide fruitful ground for data-intensive tasks, such as probabilistic modeling.

3 Behavioral Patterns

The data was collected using mobile phones that were constantly scanning for Bluetooth devices in vicinity. Since we are interested in detecting general trends in the device's Bluetooth surroundings, we restrict the analysis to the number of detected BT devices at each moment. Each of the devices produced a time series of Bluetooth densities which is characteristic to the person who was carrying the phone.

On average, people's daily behavior consists of repeating routines. This shows up in the recorded time series as cyclical trends: Daily, weekly, monthly and yearly cycles are recognizable and they can be seen as *prototypical patterns of behavior*. However these patterns are rather coarse and convoluted and they can be understood as mixtures of more fine-grained patterns of behavior. As one collects more data the average patterns become more reliable but also more generic, gradually losing their possible specificity.

We are interested in finding out a transformation from N observed time series $\mathbf{t} = (t_1, t_2, \ldots, t_n)$ of length n to M prototypical vectors or patterns $\hat{\mathbf{t}} = (\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_n)$ of length n that can be thought to form the original N time series. Note that $M << N$. Now this can be presented as

$$\hat{\mathbf{t}}^M = f(\mathbf{t}^N), \tag{1}$$

where f is the transformation from the observed time series to the prototypical components that constitute the time series. This task is commonly known as *Blind Signal Separation problem* [5]. We will use *Independent Component Analysis* to approximate a solution.

3.1 Independent Component Analysis

Independent Component Analysis (ICA) [5] is a statistical multivariate data-analysis method, which assumes the observed data to be a linear mixture of some latent features.

The basic ICA model assumes following mixing model for the latent features:

$$\mathbf{x} = \mathbf{A}\mathbf{s},$$

where $\mathbf{x}$ are the observed data, $\mathbf{A}$ are the linear mixing coefficients and $\mathbf{s}$ are the latent components. Now the task is to find inverse transformation so that we can see the latent features as

$$\mathbf{s} = \mathbf{A}^{-1}\mathbf{x}.$$

One has to remember that the only data that we know is the observed data $\mathbf{x}$. There are many different algorithms to estimate the ICA. We chose to use the *FastICA* [4] algorithm. A typical procedure when performing ICA is to first employ standard Principal Component Analysis (PCA) to reduce the dimensionality of data and then perform ICA in the lower dimensional space. In our case transformation f in Equation (1) assumes dimensionality reduction from N dimensions to M using PCA and the ICA transformation.

It should be noted that in our case this model is not completely realistic as it does not consider the possible noise in the data and also the latent features are assumed to be linearly mixed, which may not be the case. However in practice the use of PCA in the first stage reduces noise effectively. Alternatively one could use a discrete model, such as the Multinomial Principal Component Analysis [14], to solve the task.

3.2 Patterns in Bluetooth device density

We were looking for two kinds of patterns from our data set that is explained in Section 2 Firstly we were interested in seeing how the daily activity is explained through some latent independent features and secondly we were interested in seeing the same from a longer time period. Thus we analyzed both daily and weekly behavior from the same set of people. We were not interested in individual behavior as our data set is too small for that kind of analysis.

We used a subset of the data containing seven people from a time period Jan. 11.–Jan. 20. 2006 for analysis of daily behavior and Jan. 11.–Jan. 17. 2006 for the analysis of weekly behavior. The data contains also some anomalies like occasionally switched off phones and some participants not showing up in the workplace every day resulting the phones sitting alone in the office.

3.2.1 Daily analysis

In the analysis of daily behavior the number of latent features was chosen as five. The dimensionality was reduced using PCA and after that the FastICA algorithm was used.

The average pattern is shown in Figure 1A. That pattern is very comprehensible as the activity starts rising from around 8:00, reaches its peak between 13:00 and 16:00, and then starts to go down so that around 22:00 it is to its lowest level again. Figures 1B–1F show the independent patterns that were found. These figures interestingly show four different activity patterns and fifth pattern (Figure 1F) that probably relates to the fact that some of the participating individuals spent time in environments where there are not many people around and that has very few Bluetooth devices. That pattern may also be explained by the earlier mentioned anomalies in the data. The patterns in Figures 1B–1E imply that some of the people have a pattern of staying in the office only part of the day. That is understandable also based on the fact that many of the participating subjects have to spend big part of the day in meetings that have fewer people and BT devices present.

Increasing the number of estimated components does not change the situation much. From this data we always estimated 4 – 5 active components and the rest were low activity components. This implies that in this data the number of non-redundant components is around five.

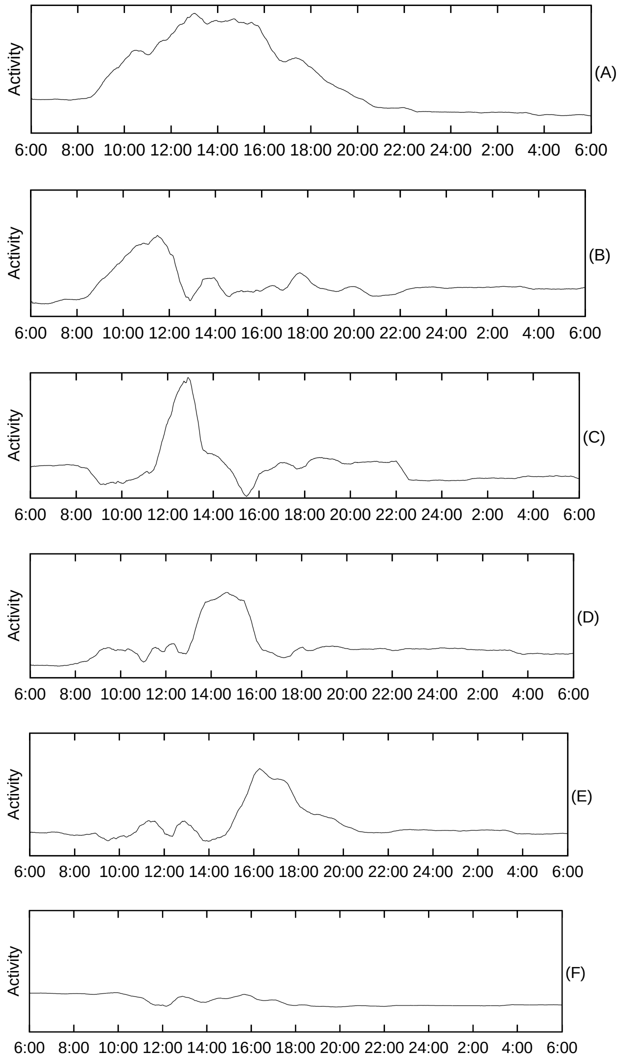

Figure 1. Daily average BT activity pattern (A) and five independent patterns (B–F) from a time period that covers eight working days between Jan. 11.–Jan. 20. 2006. Pattern A is a simple average over all the data and patterns B–F are the independent components estimated using FastICA algorithm.

3.2.2 Weekly analysis

The number of components for the analysis of weekly behavior was chosen as three since we are investigating only seven people. As we are doing ICA it is generally a good idea to reduce the dimensionality since that also reduces the noise in the data and ensures that the space is better spanned by the data.

The average prototypical weekly behavior is shown in Figure 1A. That figure is very comprehensible as the highest number of Bluetooth devices is during the working days and during the nights and the weekend the activity is lower.

Figures 2B – 2D show the estimated three independent components. As the components are estimated from only seven different people during a week's time they are most likely explained by absence of some people from the work place. E.g. Figure 2B most likely implies that a person or some persons were not present between Jan. 11. and Jan. 16. 2006. This can be further verified by looking at the original data and that turns out to be the case.

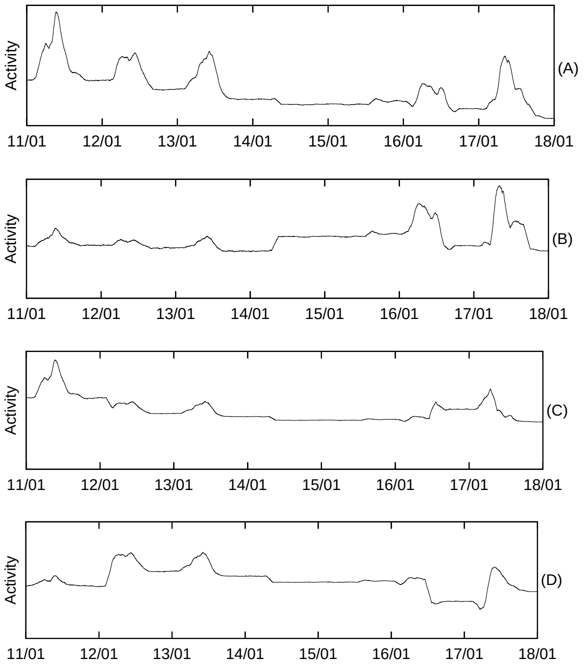

Figure 2. Weekly average BT activity pattern (A) and three independent patterns (B–D). Pattern A is simple average over all the data and patterns B–D are independent components estimated using FastICA algorithm. Each tic is set at 6:00 am.

3.2.3 Remarks

Presented results are only preliminary and we plan to do a full analysis of the data using different component models. The main result from this analysis is that it is feasible to use this kind of methods for mining Bluetooth activity data and that the patterns that we found are understandable. In this kind of corporate environment these methods can be used e.g. planning the daily functions of the environment.

4 Beaconless Locationing

A typical system for radio-based indoor locationing, e.g. [2], builds upon a set of stationary base stations or beacons, against which the current location is estimated. The system is calibrated either by careful installation of the beacons to known locations or by empirically collecting calibration data, often signal strengths, from various locations.

Following the approach presented in [6], empirical calibration assigns a probability distribution $p(o \mid l)$ for any given location l and measured signal vector o. Now given an observation we get posterior distribution of the location by applying the Bayes rule

$$p(l \mid o) = \frac{p(o \mid l)\, p(l)}{p(o)} = \frac{p(o \mid l)\, p(l)}{\sum_{l' \in \mathcal{L}} p(o \mid l')\, p(l')}, \qquad (2)$$

where $p(l)$ is the prior probability of being at location l and $\mathcal{L}$ denotes all possible location values. The posterior distribution $p(l \mid o)$ may be used to estimate the location give the current observations. In [6] the above model was used to estimate location given WiFi signal strengths. WiFi-based estimation requires careful calibration due to small number and wide coverage of base stations, making observations somewhat sparse.

Consider that instead of base stations or beacons, we would base the locationing on the observed Bluetooth neighborhood. We assume that there are some invariant features in the environment which characterize different locations. Assuming that people carrying Bluetooth-enabled devices are predictable enough and there are enough devices for robust estimation, we hypothesize that different locations have characteristic Bluetooth surroundings which serve the purpose of base stations. In contrast to WiFi base stations, there are plenty of Bluetooth devices and their range is much smaller, resulting to denser space of indicators. On the other hand, mobile Bluetooth devices are much more unreliable indicators than stationary base stations.

Let $Q_t = \{q_1, ..., q_K\}$ denote the set of Bluetooth MAC addressed detected within time-window of $[t - \delta : t]$, where δ is a free parameter. The primary purpose of this window is to mitigate the effect indeterministic Bluetooth inquiring which does not guarantee the order in which MACs of the surrounding devices are returned. If δ is made larger than a Bluetooth inquiry period (30s in our case), it acts also as a rudimentary tracking model. Quite likely a proper tracking model, as presented in [6], would improve the results when accompanied by the MAC window.

We would like to estimate $p(Q_t \mid l, t)$ as above. We make some simplifying assumptions: We ignore time in estimation, assuming that observations stay rather constant in time. Moreover we assume that MAC addresses in Q are mutually independent. This leads us to the so called Naive Bayes model

$$p(l \mid Q_t) = \frac{1}{Z} p(l) \prod_{i=0}^{|Q_t|} p(Q = q_i \mid l), \qquad (3)$$

where Z is a scaling factor as above. Using this model we may estimate a location given a set of observed MAC addresses. We considered the Maximum a Posteriori (MAP) estimate as the desired location in evaluation.

We evaluated this locationing model with our data set. Known beacons were separated from the data set and the known locations were used only for training (calibrating) the model. Testing was done with a distinct set of observation windows, without any knowledge on beacons. This was done with stratified 10-fold cross-validation [12]. Stratification is crucial, since there is much more data from some popular locations in contrast to some individual offices.

Table 2. Locationing accuracies

Location	Best estimate	2nd best estimate
a3 coffee	**81.1** a3 coffee	**11.9** a3 lab
a5 lab	**78.9** main entrance	**11.4** a5 lab
b5 lab	**55.3** b5 lab	**35.5** b525
3rd sofas	**73.7** 3rd sofas	**15.1** a3
a3 lab	**99.2** a3 lab	**0.2** 7th sofas
cafeteria	**98.1** cafeteria	**0.8** 1st tables
b525	**84.1** b525	**13.1** b5 coffee
main entrance	**96.5** main entrance	**1.7** a3 lab
b5 coffee	**70.9** b5 coffee	**18.1** 1st tables
7th sofas	**55.0** 7th sofas	**13.8** 7th offices
7th offices	**89.4** 7th offices	**9.4** 7th sofas
1st sofas	**67.6** 1st sofas	**12.4** 1st tables
5th sofas	**88.0** 5th sofas	**5.8** 7th sofas
1st tables	**40.0** 1st sofas	**30.0** 1st tables
a3 lab	**62.8** a3 lab	**19.1** 3rd sofas

The results are presented in table 2. For each known location, the table shows the most frequently estimated location in the second column and the second-most frequent estimate in the third column. The percentage tells how often the column's estimate was preferred over the others. The shown results are obtained with $\delta = 60s$. Varying δ between 10 and 120 seconds seems to have only little effect on results.

In 13 / 15 of the cases the most frequent estimate is the correct one. More importantly, in almost all cases the third column shows a place nearby the true location. Clearly the places, such as cafeteria, where lots of data was collected produce the best results. This indicates that applicability of this method depends heavily on richness of the sensed environment.

5 Future Work

Both modeling of behavioral patterns and beaconless locationing are based on the assumption that people are predictable – otherwise we could not find prototypical patterns or track locations during several days reliably. Even though routines per se are invariant, routines vary between different individuals. We could improve locationing by taking this fact into account.

There is not enough data to estimate a robust locationing model for each individual separately. However as shown by behavioral pattern analysis, many people share rather similar routines. Thus instead of individual models, we might estimate specific locationing models for groups of people. In practice, we could use methods of *collaborative filtering* for this purpose.

Another line of ongoing research is to fuse Bluetooth information with other sensor data, such as images. In one of our preliminary settings, a stationary camera produced a stream of images in which moving objects are automatically recognized. Combined with Bluetooth scans, we aim at associating the objects with their most probable Bluetooth identity.

One very important aspect, that is not discussed in this paper, are the privacy concerns that raise from the application of these methods in real life situations. We acknowledge the importance of these concerns and they are one of the most important research questions for future work.

6 Conclusions

We have presented two models for utilizing rich Bluetooth environments. We demonstrated how a model for *Blind Signal Separation* can be used to derive prototypical behaviors from Bluetooth data. This model gives a remarkably fine-grained view to people's daily routines, in contrast to straightforward summary statistics.

We hypothesized that given enough Bluetooth devices in the environment, we might use characteristic set of devices in different locations as a surrogate for stationary base stations. Stochastics of large number of detected devices would make the system robust enough for practical use. We validated the hypothesis with our Bluetooth data set, containing the ground-truth about locations. The obtained results seem to support the hypothesis.

7 References

[1] LaMarca A., Chawathe Y., and Consolov S. Placelab: Device positioning using radio beacons in the wild. In *Third International Conference on Pervasive Computing*, 2005.

[2] Pandya D., Jain R., and Lupu E. Indoor location estimation using multiple wireless technologies. In *Proceedings of the 14th IEEE Symposium on Personal, Indoor and Mobile Radio Communications*, 2003.

[3] Chen G. and Kotz D. A survey of context-aware mobile computing research. Technical report, Department of Computer Science, Dartmouth College, 2000.

[4] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9:1483–1492, 1997.

[5] Hyvärinen A., Karhunen J. and Oja E. *Independent Component Analysis*. Wiley, 2001.

[6] Kontkanen P. and Myllymäki P. and Roos T. and Tirri H. and Valtonen K. and Wettig H. Topics in probabilistic location estimation in wireless networks. In *Proceedings of the 15th IEEE Symposium on Personal, Indoor and Mobile Radio Communications*, 2004.

[7] Kotanen A. and Hännikäinen M. and Leppäkoski H. and Hämäläinen T. Experiments on local positioning with bluetooth. In *Proceedings of the International Conference on Information Techonology: Computers and Communications (ITCC)*, 2003.

[8] Galarraga M., Ucar B., Led S., and Serrano L. Gateway bluetooth - gprs for ecg signal transmission: Implementation in mobile phones. In *Proceedings of the 3rd European Medical and Biological Engineering Conference*, 2005.

[9] Eagle N. *Machine Perception and Learning of Complex Social Systems*. MIT, 2005.

[10] Pärkkä J. and Van Gils M. and Tuomisto T. and Lappalainen R. and Korhonen I. A wireless wellness monitor for personal weight management. In *Proceedings of IEEE EMBS International Conference on Information Technology Applications in Biomedicine*, 2000.

[11] ABI Research. Bluetooth: The global outlook. `http://www.abiresearch.com/products/market_research/Bluetooth`, 2006.

[12] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society (Series B)*, 36:111–147, 1974.

[13] Nicolai T., Behrens N., and Yoneki E. Wireless rope: Experiment in social proximity sensing with bluetooth. In *Fourth Annual IEEE International Conference on Pervasive Computing*, 2006.

[14] Buntine W. and Jakulin A. Discrete components analysis. In *Subspace, Latent Structure and Feature Selection Techniques*. 2006.