High-Performance Indoor Localization with Full-Band GSM Fingerprints

Bruce Denby[1,2], Yacine Oussar[2], Iness Ahriz[2], Gérard Dreyfus[2]
[1]Université Pierre et Marie Curie – Paris VI,
4 place Jussieu, 75005 Paris, France
[2]Laboratoire d'Électronique, ESPCI – ParisTech,
10 rue Vauquelin, 75005 Paris, France

Abstract- **GSM trace mobile measurements are used to study indoor handset localization in an urban apartment setting. Nearest-neighbor, Support Vector Machine (SVM), and Gaussian Process classifiers are compared. A linear SVM is found to provide mean room-level classification efficiency near 100%, but only when the full set of GSM carriers is used. To our knowledge, this is the first study to use fingerprints containing all GSM carriers, and the first to suggest that GSM could be useful for very high-performance indoor localization.**

I. Introduction

The introduction of the E911 (United States) and E112 (Europe) emergency services initiatives has spurred considerable interest in Location Based Services (LBS) for cellular telephone networks [1]. Although integrated GPS receivers can provide very accurate positioning information, few handsets are so equipped, and GPS performs poorly in indoor and urban canyon environments. For these reasons, the study of radio network based localization techniques is also a very active area.

In the database correlation method [2], a mobile is localized by comparing a recent Received Signal Strength (*RSS*) measurement to a position-labelled database of such measurements called fingerprints. GSM localization schemes often rely on regularly emitted Network Measurement Reports (NMR) containing the *RSS* and Base Station Identity Code (BSIC) of the serving cell and six strongest neighboring cells. These 7-component vectors allows localization precision of several tens of meters in outdoor environments (see for example [3,4]).

Most radio-based indoor localization studies have involved WiFi networks, where workplace "corridor waveguide" scenarios are addressed, and performance, though interesting, needs to be improved [5-7]. A novel approach using the household power lines as an antenna appears in [8]. The idea of using GSM or CDMA networks for localization in indoor, and particularly domestic, environments is still rather new (see, for example, [9] and [10]). The working hypothesis here is that the *RSS* of the external base stations should be strongly correlated with a mobile's indoor position, due to the varying absorption of electromagnetic energy by different building materials and the exact placement of doors, windows, etc. There has also been evidence that going beyond the standard 7-carrier NMR fingerprint is advantageous for indoor GSM localization [9].

In this article, we present tests of indoor GSM localization using scans containing large numbers of carriers – up to the full GSM band. We show that in an urban apartment setting, the room in which a handset is located can be identified with an efficiency approaching 100%, but only when the full set of GSM carriers is included. To our knowledge, this is the first study using fingerprints containing all of the carriers in the GSM band, as well as the first to achieve very good performance on indoor GSM localization.

The data sets used in our study are detailed in section II, while a discussion of pre-processing and a description of the classifiers tested are given in section III. Results are presented in Tables I and II and discussed in section IV, while a conclusion and some perspectives are outlined in the final section.

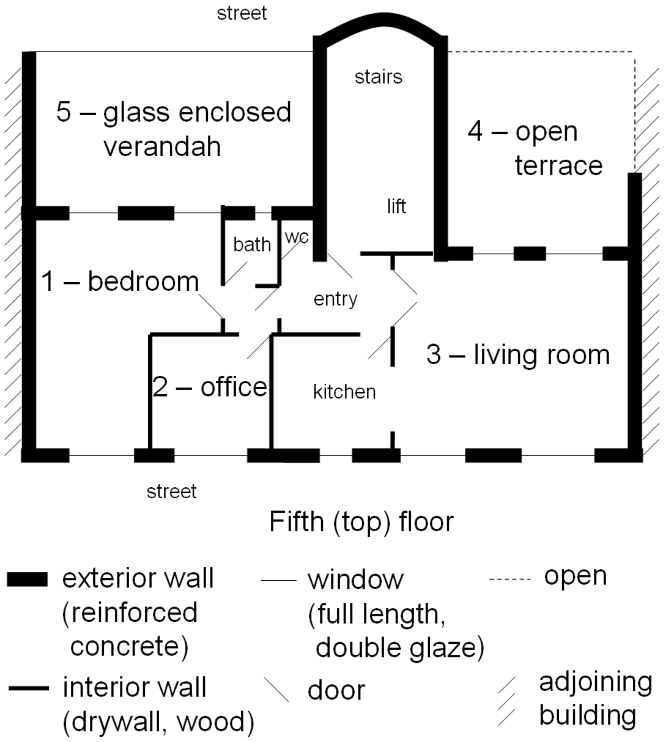

Figure 1. Schematic of apartment layout.

978-1-4244-3437-4/09/$25.00 ©2009 IEEE

II. Data Sets

Scans of the full available set of 498 GSM carriers were taken twice a day for one month in 5 rooms of a 5th (top) floor apartment in Paris, France, using the TEMS [11] trace mobile system. Both the *RSS* and the BSIC, where readable, were recorded for each carrier. A schematic of the apartment layout is given in figure 1. Acquisitions could in principle be made anywhere within a room; however, in practice, scans were typically recorded only in those areas where a laptop and cellphone could be conveniently placed and accessed.

III. Data Analysis

A. Pre-processing

Ten carriers found always to contain no energy were removed from the study. As the BSICs of the remaining 488 proved to be unreadable in many cases, a decision was made to exclude BSICs from the subsequent analysis, despite the possibility of confusing carriers at the same frequency in separate cellular motifs. The data set contained a total of 241 scans – approximately 48 scans per class, where a class is defined simply as the index of the room within the apartment, as indicated in figure 1. In order to obtain an assessment of the statistical significance of our classification results, cross validation was performed using ten independent randomly selected splits of the data, each containing 169 training examples and 72 validation examples. In any given split, the training and validation examples were randomly distributed in time over the one-month acquisition period.

B. Dimensionality Reduction and Fingerprint Types

The small size of our dataset – a reflection of the difficult and time consuming nature of obtaining labelled scan data – and its high dimensionality (488 carriers) limit the complexity of classifiers which can be effectively tested. To address this problem, signal strength based carrier selection was first carried out to define the 4 fingerprint types given below. Further dimensionality reduction on any fingerprint type can be obtained by applying Principal Component Analysis (PCA).

Three vectors are used to define the fingerprints:

$$\mathbf{g}_j^7 = \left\{ i = 1 \ldots 488, \sum_k \mathbf{1}_{RSS(i,j)<RSS(k,j)} \leq 6 \right\}$$

$$\mathbf{G}^7 = \bigcup_j \mathbf{g}_j^7$$

$$\mathbf{G}^{35} = \left\{ i = 1 \ldots 488, \sum_k \mathbf{1}_{\langle RSS(i,j)\rangle_j < \langle RSS(k,j)\rangle_j} \leq 34 \right\},$$

where $\mathbf{1}$ is the indicator function, and $\langle\ \rangle_j$ denotes the mean over the index *j*. The first, $\mathbf{g}_j^7$, contains the indices of the 7 strongest carriers, *i*, in example *j*. The vector $\mathbf{G}^7$ is composed of the indices of all carriers which were among the strongest 7 in at least one element of the training set; it contains between 36 and 40 of these "good" carriers, depending upon the random split. The last vector, $\mathbf{G}^{35}$, is made up of the indices of the 35 carriers which were the strongest, on average, over the training set. The fingerprints are then defined as follows:

1. Current Top 7

These 7 carrier fingerprints, $RSS(\mathbf{g}_j^7)$, were meant to mimic "top 7" NMRs, which were not available in our scans. Validation set fingerprints may contain fewer than 7 elements in the case of carriers which did not appear in the training set. For classifiers requiring fixed labelling of input vectors, such as KNN and SVM, the 7 $RSS(\mathbf{g}_j^7)$ values are filled in at the corresponding positions in a vector of length $\|\mathbf{G}^7\|$, the rest of whose elements are set to zero.

2. Top 7 with Memory

These fingerprints, defined as $RSS(\mathbf{G}^7)$, include the values of all 36-40 "good" carriers, and are thus "wider" than the *Current Top 7.*

3. 35 Best Overall

Another way of assessing the "goodness" of a carrier is its average *RSS* value over the whole training set. The *35 Best Overall* fingerprint, of length 35, is defined as $RSS(\mathbf{G}^{35})$.

4. All 488

All active carriers *RSS* values are included (no selection).

C. Classifiers

Three types of classifier were tested:

1. Support Vector Machines (SVM)

A 2-class SVM [12] determines the separating surface which maximises the distance (called the "margin") between this surface and the data points appearing on either side of it. The SVM may be linear, operating directly upon the data, or map the data first to a higher-dimensional space via a non-linear transformation before finding the maximum margin surface. The SVM decision rule is obtained by taking the sign of

$$f(\mathbf{x}) = \sum_{i=1}^{N_S} \alpha_i y_i K(\mathbf{s}_i, \mathbf{x}) + b$$

where $\mathbf{x}$ is the *RSS* vector to be localized, N_s is the number of support vectors $\mathbf{s}_i$ (training vectors residing on the optimal separating surface), $y_i = \pm 1$ is the class label of the vector $\mathbf{s}_i$, $K(.)$ is the selected kernel, and b and the α_i are parameters determined during the search for the optimal separating surface. It is known that for large, well behaved data sets, the SVM rule approximates the Bayes decision rule [12].

For a linear SVM the kernel function is simply the scalar product $K(\mathbf{s}_i,\mathbf{x}) = \mathbf{s}_i \cdot \mathbf{x}$. A standard Gaussian kernel was adopted in our tests of non-linear SVMs,

$$K(\mathbf{s}_i,\mathbf{x}) = e^{-|\mathbf{s}_i-\mathbf{x}|^2/\sigma^2}$$

where the variance σ^2, as well as a regularization parameter that controls the complexity of the separating surface [12], are optimized in the cross-validation stage. For m classes, it is traditional (conventional recipe [13]) to construct m binary, one-vs-rest classifiers, and identify the output class as that of the classifier with the largest output value, before thresholding. This procedure is illustrated for our case of m = 5 in figure 2. The Spider SVM modelling package [14] was used in all our analyses.

2. K-Nearest Neighbor (K-NN)

The *K*-NN classifier first ranks all training vectors according to their Euclidean distances, in *RSS*-space, from the test vector to be localized. The predicted class of the test vector is then taken to be the class which is most represented in the *K* "nearest" vectors according to the defined metric. The parameter *K* is chosen empirically to optimize performance. When the single best neighbor is used, we have *K*=1 and the classifier is denoted 1-NN.

3. Gaussian Process (GP)

As with *K*-NN, GP begins by comparing the test *RSS* vector to be localised to each vector in the training set. The probability P_1 that the two compared vectors correspond to measurements taken at (nearly) the same geographical position is assumed to be Gaussian in the Euclidean *RSS* distance between the two vectors, with a fixed variance σ^2, determined empirically. If a carrier appears in one of the compared vectors but not in the other, GP assumes that the missing value was below the threshold for reception in the deficient vector. A penalty term probability P_p is then introduced, in which the missing *RSS* value is filled in by an estimate of the reception threshold taken to be the smallest *RSS* in the vector missing the carrier. The overall GP probability *P* is given by the product of P_1 and P_p.

More precisely, let A and B be the sets of indices of carriers contained in a training set vector and a test set vector, respectively. We define the set of common carriers as C=A∩B, and the train and test non-common carrier sets as D=A-C and E=B-C, respectively. We then have

$$P_1 = \sqrt[|C|]{\prod_{i\in C} e^{-\left|RSS_i^A - RSS_i^B\right|^2/\sigma^2}}$$

$$P_p = \sqrt[|D|]{\prod_{j\in D} e^{-\left|RSS_j^D - \min_B\left(RSS^B\right)\right|^2/\sigma^2}} \times \sqrt[|E|]{\prod_{k\in E} e^{-\left|RSS_k^E - \min_A\left(RSS^A\right)\right|^2/\sigma^2}}$$

$$P = P_1 \cdot P_p$$

where RSS_i^A denotes the signal strength of the i^{th} carrier of set A, and the order of each radical normalizes the probability to the number of carriers in the corresponding term. GP is in fact the only classifier tested which is able to handle missing carriers in a natural way. When input vectors are of fixed length – a requirement for SVM and KNN – and all variables are represented, GP becomes equivalent to a 1-NN classifier. As a caveat, since we do not use the BSIC information, in some cases carriers with the same index could belong to different cellular motifs, which may degrade the performance of GP.

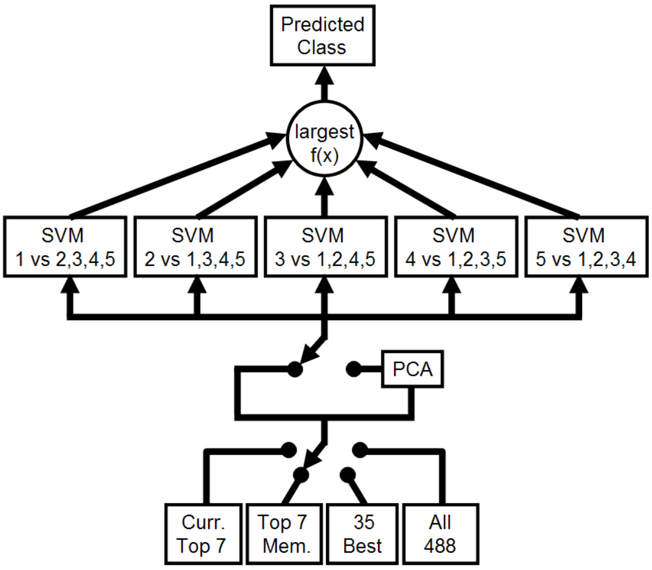

Figure 2. Architecture combining five one-vs-rest SVM classifiers to predict the class of an *RSS* vector from one of the carrier sets.

IV. Results

We define localization performance as the overall correct classification rate. Table I shows that the performance of all classifiers tested improves as more carriers are added to the fingerprint, but that very good performance – for example our best result of 97.8% in the case of the linear SVM – is only obtained on the All 488 carrier fingerprint. The implication is that indoor position can indeed be deduced from the *RSS* of GSM cell towers, but that commonly used 7-carrier NMRs and even "wide" fingerprints are insufficient – high performance requires fingerprints of very high dimensionality. Further support for this conclusion is given by Table II, in which the confusion matrices for the linear SVM classifier on 35 Best Overall and All 488 fingerprints are presented. It is again clear that the ability to sharply discriminate between rooms comes only with the inclusion of

TABLE I

Classifier			**Fingerprint Type**						
			Current Top 7 (≤7 carriers)[1]		**Top 7/Memory** (36-40 carriers)		**35 Best Overall** (35 carriers)		**All 488** (488 carriers)
Linear SVM			71.3 ± 7.2		84.6 ± 3.6		90.4 ± 3.5		**97.8 ± 1.5**
Gauss. SVM	**w/o PCA**		72.2 ± 3.6		89.2 ± 2.9		93.2 ± 3.4		–[2]
	w/PCA[3]		71.8 ± 3.2		85.6 ± 5.3		92.0 ± 3.0		96.4 ± 1.5
K_{best}	***K*-NN**	5	59.3 ± 3.5	26	85.1 ± 3.0	20	93.3 ± 2.1	20	94.9 ± 1.9
***1*-NN**			58.1 ± 5.2		74.7 ± 3.7		86.0 ± 2.9		87.2 ± 2.8
GP (σ= 5 dB)			78.8 ± 3.7		–[4]				

Percentage of correct radio fingerprint classifications on the 4 carrier sets described in the text. Figures quoted are averages and standard deviations over 10 randomly selected validation sets. All classifiers achieve their best performance when all 488 carriers are included. The most effective classifier for this case is the linear SVM.

[1] SVM and K-NN can have < 7 carriers if some did not show up in the training set.
[2] Small training set size precludes training a Gaussian SVM due to Cover's theorem [13].
[3] Best result, first 4 principal components. PCA is used exclusively in this line of the table.
[4] Gaussian process is equivalent to 1-NN for fixed input vector length.

TABLE II

Confusion Matrix	**True Class**										
	35 Best Overall						**All 488**				
Pred. Class	**1**	**2**	**3**	**4**	**5**		**1**	**2**	**3**	**4**	**5**
1	95	5.3			3.3		100		0.7		
2	1.4	93.3	3.6					100			
3	0.7	1.3	77.9	11.4					91.4	1.4	1.3
4			16.4	87.9	0.7				5.7	98.6	
5	2.9		2.1	0.7	96				2.2		98.7

Confusion Matrices for 35 Best Overall and All 488 carrier sets, using a Linear SVM classifier. Figures quoted are in percent. Using the full number of carriers tightens up the diagonal to give individual room classification efficiencies near 100%.

the full GSM carrier set. The deviation of our global result from 100% is in fact dominated by the confusion between class 3 and class 4, which appears to be the most difficult case. A non-linear SVM might provide better results, but our limited training set size precludes training such a classifier on the All 488 dataset due to Cover's theorem [15], which states that a training set is always linearly separable when the number of input variables exceeds the number of examples. In any case, the ability to achieve good performance on a small training set is an interesting result in itself, as it suggests that large, difficult to obtain sets of labelled data might not be necessary in a final application.

V. Conclusions and Perspectives

We believe this to be the first instance of including the full set of GSM carriers in an *RSS* fingerprint for a localization study. Although confirmation at additional sites will clearly be required, our results here suggest that high-performance room-level localization is possible through the use of such fingerprints. It is also interesting to note that our result appears to be robust against time dependent effects, such as network modifications, propagation channel changes, meteorological effects, etc., since our dataset was acquired over a period of one month.

Performances might be further improved by including the BSIC information on those carriers for which it is readable, using more sophisticated classification techniques, or extending the fingerprints to include other locally available information.

For larger indoor areas, a regression approach based on x-y position may be more appropriate than the room-by-room classification used here. It will also be interesting to examine semi-supervised learning algorithms in order to address the problem of the difficulty of obtaining labelled training data [7,16].

A subsequent article, implementing one-vs-one classifiers, as well as other improvements, is currently in preparation. In order to obtain a more statistically significant performance assessment, the acquisition of a new, larger data set is also in progress.

References

[1] Axel Küpper, Location-Based Services: Fundamentals and Operation, John Wiley & Sons, 2005.

[2] D. Zimmerman, J. Baumann, M. Layh, F. Landstorfer, R. Hoppe, G. Wölfle, Database correlation for positioning of mobile terminals in cellular networks using wave propagation models, in Proc. IEEE 60th Vehicular Technology Conference 26-29 September 2004; 7, pp. 4682-4686.

[3] M. Chen, T. Sohn, D. Chmelev, D. Haehnel, J. Hightower, J. Hughes, A. LaMarca, F. Potter, I. Smith, A. Varshavsky, Practical metropolitan-scale positioning for GSM phones, P. Dourish, A. Friday, Eds., in Proc. 8th International Conference on Ubiquitous Computing, Orange County, CA, USA, September 17-21, 2006, Lecture Notes in Computer Science 2006; 4206, pp. 225-242, Springer.

[4] B. Denby, Y. Oussar, I. Ahriz, Geolocalisation in Cellular Telephone Networks, proceedings of NATO 2007 Advanced Study Institute on Mining Massive Data Sets for Security, IOS Press, Amsterdam, The Netherlands, F. Fogelman-Soulié, D. Perrotta, J. Piskorski & R. Steinberger, Eds., IOS Press, Amsterdam, Netherlands, in press.

[5] M. Brunato, R. Battiti, Statistical learning theory for location fingerprinting in wireless LANs, Computer Networks and ISDN Systems April 2005; 47, Issue 6, pp. 825-845, Elsevier Science Publishers, Amsterdam.

[6] A. M. Ladd, K. E. Bekris, A. Rudys, L. E. Kavraki, D. S. Wallach, On the Feasibility of Using Wireless Ethernet for Indoor Localization, IEEE Transactions on Robotics and Automation, June 2004, Volume 20, Issue 3, Number 3, pp. 555-559.

[7] Qiang Yang, Sinno Jialin Pan, Vincent Wenchen Zheng, Estimating Location Using Wi-Fi, IEEE Intelligent Systems, vol. 23, no. 1, pp. 8-13, Jan/Feb, 2008.

[8] S.N. Patel, K.N. Truong, G.D. Abowd, Powerline positioning: A practical sub-room-level indoor location system for domestic use, proceedings of the 8th international conference on ubiquitous computing, UbiComp 2006, Orange County, CA, USA, September 17-21, 2006.

[9] V. Otsason, A. Varshavsky, A. LaMarca, E. de Lara, Accurate GSM indoor localization, in Proc. UbiComp 2005, M. Beigl et al, Eds., pp. 141-158, Springer-Verlag, Berlin, Heidelberg.

[10] W. ur Rehman, E. de Lara, S. Saroiu, CILoS: A CDMA indoor localization system, proceedings of the 10th international conference on ubiquitous computing, UbiComp 2008, September 21-24, 2008, Seoul, Korea.

[11] Test Mobile System. [Online]: http://www.ericsson.com/solutions/tems/

[12] N. Cristianini, J. Shawe-Taylor. Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, 2000.

[13] Y. Lee, Y. Lin, G. Wahba, Multicategory support vector machines: theory and application to classification of microarray data and satellite radiance data, Journal of the American Statistical Association March, 2004; 99, no. 465, pp. 67-81.

[14] The Spider. [Online]: http://www.kyb.tuebingen.mpg.de/bs/people/spider/

[15] T.M. Cover, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, IEEE Trans. Electronic Computers 1965; 14, p. 326-334.

[16] O. Chapelle, B. Schölkopf and A. Zien, Semi-Supervised Learning, MIT Press, Cambridge, Massachusetts, 2006.