- ORIGINAL ARTICLE -

Accuracy of Bluetooth based Indoor Positioning using different Pattern Recognition Techniques

Precisión del posicionamiento en interiores utilizando Bluetooth con diferentes técnicas de reconocimiento de patrones

María Rodríguez-Damián, Xosé A. Vila, Leandro Rodríguez-Liñares

Department of Computer Science, University of Vigo, 36310, Vigo, Spain
{mrdamian, anton, leandro}@uvigo.es

Abstract

Object indoor location is a field that receives much research effort but that is lacking enough maturity for its integration in popular devices like mobile phones. This paper describes the results of an experiment carried out to compare different pattern recognition algorithms in order to process the information from a set of Bluetooth transmitters, located in fixed positions, with the aim of locating an object in a precise position. Our conclusion is that the best algorithms, among the five we tested, are random forests and model-based clustering, which gave accuracies around 90%. We have also conducted experiments to analyse the influence of the number of Bluetooth transmitters and to determine the sets of features with better performance. The proposed approach is simple and gives 90% of accuracy for locating objects with 1 m precision, making it suitable for a wide range of applications.

Keywords: Bluetooth, indoor location, indoor positioning.

Resumen

Actualmente en el campo de la localización de objetos en interiores está recibiendo mucha atención por parte de los investigadores, pero todavía necesita de la madurez suficiente para su integración en dispositivos tan populares como los teléfonos móviles. En este artículo se describen los resultados de un experimento realizado para comparar diferentes algoritmos de reconocimiento de patrones con el fin de procesar la información de un conjunto de transmisores Bluetooth, situados en posiciones fijas, con el objetivo de localizar un objeto en una posición precisa. Nuestra conclusión es que los mejores algoritmos, entre los cinco que hemos experimentado son: Random forest y model-based clustering, que alcanzaron una precisión cercana al 90%. También hemos llevado a cabo experimentos para analizar la influencia del número de transmisores Bluetooth y para determinar los conjuntos de características con mejor rendimiento. El enfoque propuesto es simple y proporciona un 90% de precisión para la localización de objetos con una precisión de 1 m, lo que lo hace adecuado para una amplia gama de aplicaciones.

Palabras claves: Bluetooth, localización en interiores, posicionamiento en interiores.

1. Introduction

Nowadays, the number of mobile phone lines worldwide exceeds the number of inhabitants of the planet. Most smartphones include a GPS (Global Positioning System) antenna that can be used to locate their position, and therefore their owners'. Mobile services making use of this location capability are known as location-based services. However, as performance inside buildings of GPS-based location services degrades severely, other technologies have been proposed to indoor location.

Some examples of services currently under development are guidance of people in hospitals, airports or administrative buildings, information of offers to customers in malls, "find your friend" and other gaming applications, guidance in museums, ambient assisted living, and surveillance applications.

The structure of the rest of this paper is as follows. Next section includes a review of the technologies that can be used and the algorithms they are based on to process the information obtained from the sensors. Then, section 3 (Methodology) describes the experimental testbed. Section 4 contains the results obtained from the experiments and a discussion of these results. Finally, the paper ends with some conclusions.

2. Background

Due to the great amount of attention received by indoor positioning during the last years, many proposals have appeared. In this section, a review of some of these approaches is included.

For this purpose, first the technical alternatives are presented and then the algorithms used by the systems are summarized.

2.1. Technology used for indoor location

According to the nature of the sensors the indoor location used, some authors as in [1] and [2] classified them in:

- *Optical positioning:* based on the processing of images or video. Accuracy of these systems have reached the 10^{-1} *meters*. Their main disadvantages are the high cost and the required processing capabilities.
- *Infrared positioning:* the object to locate emits an unique code (ID) at regular intervals via an infrared transmitter. Infrared receivers are located in fixed positions, allowing to localize each ID within a map. An example is described in [3] and using Active badges.
- *Ultrasonic positioning:* distance between the object to locate (target) and a point station is estimated by using ultrasounds. Two well-known examples are Active Bats in [4], and Cricket in [5].
- *Radio Frequency Identification (RFID):* targets carry tags including an identification number (ID) and readers are distributed throughout the room in fixed positions. The opposite scenario is also possible (a mobile reader on the target and fixed tags). An example developed by Ni is LANDMARC [6].
- *Inertial Navigation Systems (INS):* equipped with motion and rotation sensors capable of estimating position, orientation, and velocity of an object without external references as described in [7].
- *Wireless technologies*: the most used ones are Wi-Fi (IEEE 802.11) and Bluetooth (IEEE 802.15.1). Their main advantages are the reuse of existing infrastructure, that can be found in a broad range of devices, and their low cost. Between these two, Bluetooth offers lower power consumption and smaller size.

The most used one is wireless technologies. In these systems, transmitters and receivers interchange protocol-dependent information which contains some parameters that can be used to estimate the distance among them and to locate the target. According to [8] the parameters most frequently used are: Time of Arrival and Time of Flight, Angle of Arrival, Received Signal Strength Indication (RSSI) and Connectivity.

2.2. Algorithms for indoor location

Once the data are collected from the sensors, systems must rely on some algorithm to estimate the location of the target. These algorithms can be broadly classified into three types:

- *Proximity algorithms*: they require the deployment of a dense grid of antennas, each one on a well-known position. The target is considered to be adjacent to the antenna which receives the strongest signal [8].
- *Triangulation*: the position of the target is estimated by measuring its distances from multiple reference points. It is very sensitive to the presence of walls, furniture or even moving people as detailed in [9].
- *Fingerprinting*: signal features in some specific positions are obtained. The location of the target is then estimated by matching new signal features against the previously collected ones and choosing the best match, as explained in [10]. Fingerprinting has received more and more attention since it offers the possibility of deploying accurate and low-cost positioning systems. Complete reviews of fingerprinting can be found in [1] [9] [10].

Fingerprinting location techniques are based on two differentiated phases. The training phase, obtains the RSSI values at fixed locations from all the transmitter devices in range to build the fingerprints of these particular locations. The collection of all fingerprints is usually called a radio map. In the second or on-line phase, the location of the target is estimated by comparing the RSSI values with those already stored in the radio map. The actual position of the target is indicated by the best match.

The on-line phase relies on pattern matching techniques that can include both deterministic or probabilistic methods [11]. Deterministic methods use a similarity metric to compare the on-line signal against the fingerprints. Examples of these methods are k-Nearest Neighbour (kNN) and its variants, Kernel Density Estimation (KDE), Neural Networks (NN), Support Vector Machines (SVMs) or Linear Discriminant Analysis (LDA). Probabilistic methods

are based on the statistical inference between the on-line signal and the fingerprints, and each estimated location may be accompanied by a confidence interval. Examples of these methods are Bayesian Networks, Gaussian Processes or Maximum Likelihood Estimation (MLE) [11].

The topic of how to build the fingerprints for the radio map has been widely investigated. Proposals range from average values of RSSI or statistical parameters to Gaussian models, or other complex distributions such as RSSI histograms. In this work, we used statistical parameters obtained from the RSSI measurements; for comparing them we tested different pattern recognition techniques in order to determine which ones are the most accurate and reliable.

3. Methodology

3.1. Experimental testbed

The experimental testbed was aimed at measuring the performance of different algorithms which make use of information obtained from Bluetooth devices to locate people indoor. In figure 1 a scheme of the experimental setup is shown. In a room measuring 8 x 3.5 *m*, three Bluetooth devices (BT1, BT2 and BT3) were deployed.
The room contains a Wi-Fi access point and it is also covered by other adjacent Wi-Fi networks. The room was divided into 24 imaginary "cells" of equal size 1 x 1.17 *m*. The goal was to assess the accuracy of different algorithms to determine in which of these cells a subject carrying a mobile Bluetooth device is located.

3.2. Data collection

We used the same procedure to build two independent data sets: one for training and validation, and one for testing. For collecting data, a tablet was placed at the center of each of the 24 cells. Then, the tablet executed a procedure known as "*device discovery*", which searches for Bluetooth devices configured as visible in its coverage area. This is an asynchronous process that involves what is known as an "*inquiry scan*" with duration of approximately 12 *seconds*, followed by a "*page scan*" directed at each found device. This page scan is answered by each discovered device, and from these answers information such as MAC addresses and RSSI values can be obtained.

The radio map was built by placing the tablet at the different cells and running the "*device discovery*" 40 times at each one. Each inquiry gives as result three values per device, namely, the identification of the cell, the MAC address and the RSSI value. At the end of this process, the radio map was composed by 2880 {cell, MAC, RSSI} triplets (24 cells x 3 BT devices x 40 inquiries).

The test set was recorded in a similar way. The main difference is that the tablet completed six full tours of the room, making 10 inquiries each time a cell was visited. This set was thus composed by 4320 {cell, MAC, RSSI} triplets (24 cells x 3 BT devices x 10 inquiries x 6 tours). In the radio map, each cell is represented by 120 RSSI values (40 RSSI values for each BT device). In order to build the feature map, we will try three different approaches:

- **RSSI**: use the median of RSSI values for each BT device. In this case, each cell is characterized by three parameters.
- **RSSdec**: calculate the deciles of the RSSI distribution for each BT device. It makes a total of 30 features per cell.
- **RSSstat**: extract eleven basic statistical parameters that characterize the distribution of the RSSI values: variance, standard
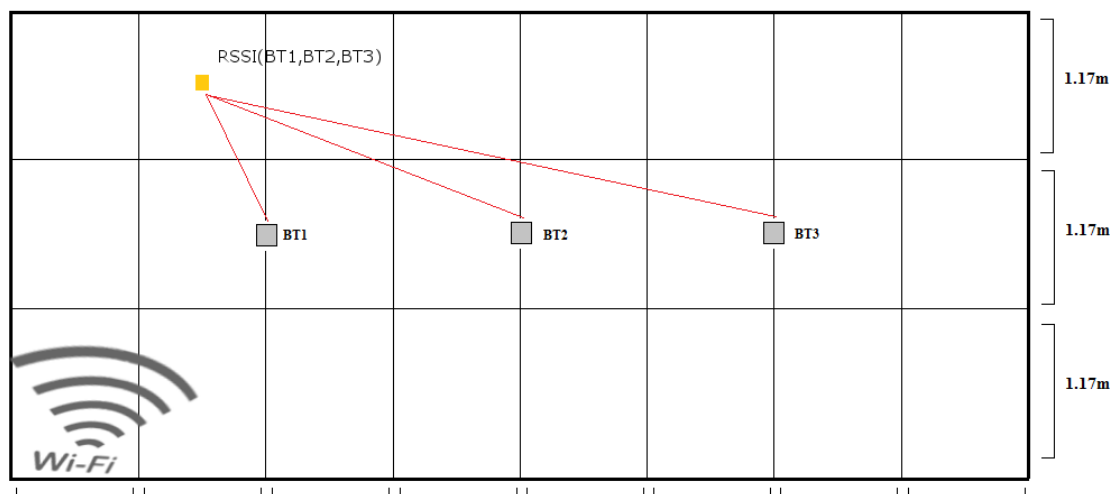

Figure 1. Experimental room showing the positions of the Bluetooth devices.

deviation, Skewness, Kurtosis, inter quartile range, trimmed mean at 10%, median, standard error, coefficient of variation (standard deviation/mean), mean absolute deviation and median absolute deviation. This gives a total of 33 features per cell.

3.3. Pattern recognition algorithms

Different algorithms have been employed for indoor location using Bluetooth devices. In this paper we explore the combination of the aforementioned features with the following classification algorithms: random forests (RF), support vector machines (SVM), K-means, hierarchical clustering, and model-based clustering.

RFs are trained by making random partitions in the training dataset and fitting a simple prediction model to each of these partitions. The models are represented as decision trees, where each node tests an attribute, each branch corresponds to an attribute value, and the leaf nodes correspond to classification decisions, as detailed in [9]. All the trees are used in the testing phase: each tree classifies new data into one class (it gives a "vote" for that class) and, finally, the forest chooses the class with more votes across all the trees.

RFs are also a tool to study the validity of the different variables for classification. Two measurements can be calculated from RFs: accuracy and gini. *Accuracy* represents how performance of the model worsens without each variable: a high decrease in accuracy would be expected for very predictive variables. *Gini* essentially measures how pure the nodes are at the end of the tree. This procedure can be useful to select subsets of data including only the most relevant variables [10] [11].

SVMs, developed by Cortes and Vapnik [12], are based on the concept of finding decision hyper-planes to separate sets of data having different class memberships. Good separation will be achieved by the hyper-plane with the largest distance to the nearest training data points from all the classes. SVMs decision functions depend on subsets of the training data (the support vectors), and kernel functions can be specified for these decision functions.

Clustering can be seen as an exploratory data analysis task, which aims to find the intrinsic structure of data by organizing data objects into similarity clusters. One of the most used is K-means, which determines a partitioning of data into k clusters that minimize the dispersion of data of each cluster and maximize the distance among cluster centroids, (see a complete review in [13]). Another similar algorithm is hierarchical clustering, which is based on building a nested sequence of partitions by merging the closest (or splitting the farthest) groups of data to create a hierarchy of clusters, as proposed in [6].

Hierarchical clustering and K-means are heuristic approaches, that is, they are not based on formal models. On the other hand, model-based clustering considers data as matching a distribution coming from a mixture of components, one for cluster [14]. Each cluster k is thus modelled by a Gaussian distribution characterized by a mean vector, a covariance matrix (which stablishes the geometric feature of the cluster), and an associated probability in the mixture. These parameters of the clusters are estimated using the Expectation-Maximization algorithm. For each new sample, the model is capable of estimating the probability of belonging to each of the clusters.

All the tests were done using R, a language and environment for statistical computing developed by R Core Team [15]. The following libraries were used: *caret* [16] to automatize the procedure for training and testing, *randomforest* [17] for RFs , *e1073* [18] for SVM, *stats* [15] for K-means and hierarchical clustering, and *mclust* [19] for model-based clustering.

4. Results

4.1. Feature selection

Three different features sets were tested: RSSI, RSSdec and RSSstat. Whereas RSSI is scalar, both RSSdec and RSSstat are vectors containing 10 and 11 parameters, respectively. RSSdec and RSSstat are both composed by features that may not have the same discriminatory power. Thus, we first tried to study the relevance of these features to remove the redundant ones and reduce the size of the feature vectors. For this purpose, a RF algorithm was applied to the training dataset and ranked the different features according to their *importance* or contribution to the *accuracy* and to the *gini* values. The process was repeated twice to reduce the number of features in both vectors: first, the 5 most relevant features were chosen, and then, among these remaining 5 features, the 3 most relevant were selected. To assess their accuracy a classifier was employed with all the vectors. The most relevant features are listed in table 1.

Table 1. RSSdec - RSSstat description and best features.

RSSdec	5 features: 10, 30, 70, 90 and 100 deciles.
	3 features: 10, 90 and 100 deciles.
RSSstat	5 features: trimmed mean, standard deviation, median, standard error and coefficient of variation.
	3 features: trimmed mean, median and coefficient of variation.

As we can see in figure 2, the shorter the vectors, the higher the accuracy of the system. The discrimination power of the sets is increased by the removal of the redundant features. This is called the Hughes phenomenon [20] and it is related with the finite size of the training dataset.
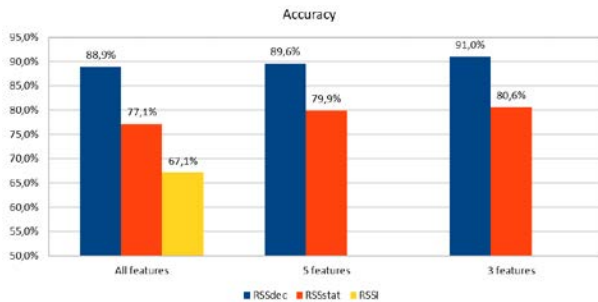

Figure 2. Accuracy of the classification algorithm.

4.2. Number of BT devices

Three Bluetooth devices (BT1, BT2 and BT3) were deployed for our experimental testbed. We decided to test the impact of the number of the BT devices in the final accuracy of the system. For this purpose, the RSSI values and the 3-parameter versions of RSSdec and RSSstat vectors which yielded the best results in the previous experiment were used as inputs to a RF classifier.
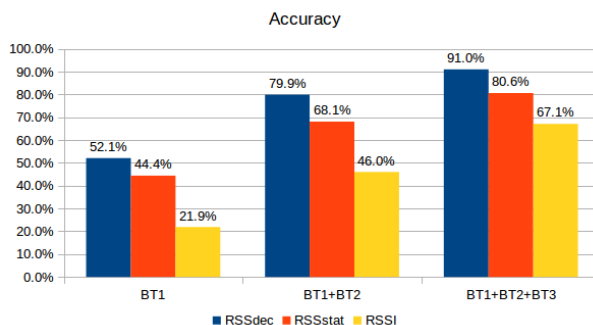

Figure 3. Accuracy of the classification algorithm (RF) using one, two and three BT devices.

As can be seen in figure 3, one BT device gives rather poor results (around 50% of accuracy). Two BT devices increase accuracy to 80% in the best case but it is still far from the 91% of accuracy achieved using three BT devices.

4.3. Pattern recognition algorithms

From the previously presented results, it is clear that RSSdec and RSSstat feature vectors perform better than RSSI and that the 3-feature reduced sets have better performance than the full sets. Then, in this analysis we compare different classification algorithms using 3-feature RSSdec and RSSstat sets. We trained all the algorithms over the training set

using a 3-fold cross validation repeated 10 times. After that, we applied the classifiers to the test set to obtain the final accuracy of each algorithm.

We compared five algorithms: RF, SVM, K-means, hierarchical clustering and model-based clustering. For K-means we selected k=24 classes matching the 24 cells that compose the used room. We used the Ward criteria [21] for calculating distances in hierarchical clustering and for model-based clustering we used a multivariate mixture model using the EDDA (Eigenvalue Decomposition Discriminant Analysis), detailed in [22], for selecting the best models.

Figure 4 shows the results we obtained, in all cases, results are better for the RSSdec feature set (between 81% and 91%) than for the RSSstat set (between 69% and 81%). The algorithm that gives the worst results is K-means, whereas best results are achieved when using RF and model-based clustering.

5. Conclusions

The main objective of this paper was to compare different pattern recognition algorithms to determine their accuracy for indoor location using information from Bluetooth transmitters. Besides, an analysis of the performance of the vectors that can be built from the RSSI and their optimal size was carried out. We tested different combinations of features generated from the RSSI values given by the Bluetooth protocol. Our conclusion is that the best sets of features are the combination of 10, 90 and 100 deciles and the combination of three statistics: trimmed mean, median and coefficient of variation.

Our experiments were carried out in a medium-sized room equipped with three Bluetooth transmitters. We have also tested if all of them were necessary or it is possible to obtain good performance with fewer transmitters. Our conclusion is that one device gives rather poor results but by using two devices good results are achieved (80% of accuracy working with 1 m^2 error location).

Finally, following the main objective of our work, we compared five different pattern recognition algorithms (SVM, K-means, hierarchical clustering, model-based clustering and RF) to determine which one works best. Results indicate that the best performance is achieved by RF and model-based clustering working on deciles 10, 90 and 100, obtaining 91% of accuracy.

Our system gives results that are similar to other approaches that can be found in the literature. For applications such as guidance in museums, surveillance applications or indoor navigation this technology seems to be suitable and it has the advantage of being a low cost approach, since most

mobile phones have Bluetooth capabilities and to deploy Bluetooth beacons in a building has a moderate cost. There are other applications, such as automatic robot driving, which need more spatial resolution or less error rate. In this case, either other technologies must be used, or more complex algorithms (collaborative locations, combining spatial and temporal information) will be needed.

Competing interests

The authors have declared that no competing interests exist.

References

[1] L. Mainetti, L. Patrono and I. Sergi, "A survey on indoor positioning systems," in *22nd International Conference on Software, Telecommunications and Computer Networks*, 2014.

[2] A. K. Nuaimi and H. Kamel, "A survey of indoor positioning systems and algorithms," in *International Conference on Innovations in Information Technology*, 2011.

[3] R. Want and A. Hopper, "Active badges and personal interactive computing objects," *IEEE Transactions on Consumer Electronics,* vol. 38, no. 1, pp. 10-20.

[4] A. Harter, A. Hopper, P. Steggles, A. Ward and P. Webster, "The anatomy of a context-aware application," *Wireless Networks,* vol. 8, no. 2/3, pp. 187-197, 2002.

[5] N. Priyantha, A. Chakraborty and B. H., "The cricket location-support system," in *Proceedings of the 6th Annual International Conference on Mobile Computing and Networking*, Boston, 2000.

[6] L. Ni, Y. Liu, Y. Lau and P. A.P., "LANDMARC: indoor location sensing using active RFID," *Wireless networks,* vol. 10, no. 6, pp. 701-710, 2004.

[7] F. Li, C. Zhao, G. Ding, J. Gong and C. a. Z. F. Liu, "A reliable and accurate indoor localization method using phone inertial sensors," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, Pittsburgh, 2012.

[8] L. Zhang, X. Liu, J. Song, C. Gurrin and Z. Zhu, "A comprehensive study of bluetooth fingerprinting-based algorithms for localization," in *27th International Conference on Advanced Information Networking and Applications Workshops*, March, 2013.

[9] W. Loh, "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* vol. 1, no. 1, pp. 14-23, 2011.

[10] L. Breiman, "Random forests," *Machine learning,* vol. 45, no. 1, pp. 5-32, 2001.

[11] R. Genuer, J. Poggi and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognition Letters,* vol. 31, no. 14, pp. 2225-2236, 2010.

[12] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning,* vol. 20, no. 3, pp. 273-297, 1995.

[13] A. Jain, "Data clustering: 50 years beyond K-means," *Pattern recognition letters,* vol. 31, no. 8, pp. 651-666, 2010.

[14] C. Fraley and A. Raftery, "Model-based clustering, discriminant analysis, and density estimation," 2002.

[15] R Core Team, "R: A language and environment for statistical computing. R Foundation for Statistical Computing," Vienna, 2016.

[16] M. Kuhn, "Caret: Classification and Regression Training. R package v. 6.0-71," 2016.

[17] L. Breiman, "Random Forests for Classification and Regression. R package v.4.6-12," 2015.

[18] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C. Chang and C. Lin, "Misc Functions of the Department of Statistics. Probability Theory Group (Formerly: E1071), TU WienE1071. R package v.1.6-8.," 2015.

[19] C. Fraley and A. E. Raftery, "MCLUST version 3: an R package for normal mixture modeling and model-based clustering.," Washington Univ Seattle Dept of Statistics.

[20] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Transactions on Information Theory,* vol. 14, no. 1, pp. 55-63, 1968.

[21] J. D. Banfield and A. E. Raftery, "Model-based Gaussian and non-Gaussian clustering," *Biometrics,* pp. 803-821, 1993.

[22] H. Bensmail and G. Celeux, "Regularized Gaussian discriminant analysis through eigenvalue decomposition," *Journal of the American statistical Association,* vol. 91, no. 436, pp. 1743-1748, 1996.