Universidad de Alcalá
BIBLIOTECA

Document downloaded from the institutional repository of the University of Alcala: http://ebuah.uah.es/dspace/

This is a posprint version of the following published document:

Hernández, N., Alonso, J.M. & Ocaña, M. 2016, “Hierarchical approach to enhancing topology-based WiFi indoor localization in large environments”, Journal of Multiple-Valued Logic & Soft Computing, vol. 26, no. 3-5, pp. 221-241.

© 2016 Old City Publishing

(Article begins on next page)

Universidad de Alcalá

This work is licensed under a

Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

J. of Mult.-Valued Logic & Soft Computing, Vol. 26, pp. 221–241
Reprints available directly from the publisher
Photocopying permitted by license only

©2016 Old City Publishing, Inc.
Published by license under the OCP Science imprint,
a member of the Old City Publishing Group.

Hierarchical Approach to Enhancing Topology-based WiFi Indoor Localization in Large Environments

NOELIA HERNÁNDEZ[1], JOSE M. ALONSO[2] AND MANUEL OCAÑA[1]

[1]*Department of Electronics, Polytechnic School, University of Alcalá, Alcalá de Henares (Madrid) 28871, Spain*
E-mail: noelia.hernandez@depeca.uah.es, mocana@depeca.uah.es
[2]*European Centre for Soft Computing, Mieres (Asturias) 33600, Spain*
E-mail: jose.alonso@softcomputing.es

Received: December 23, 2013. Accepted: July 28, 2014.

Traditionally, WiFi has been used for indoors localization purposes due to its important advantages. There are WiFi access points in most buildings and measuring WiFi signal is free of charge even for private WiFi networks. Unfortunately, it also has some disadvantages: when extending WiFi-based localization systems to large environments their accuracy decreases. This has been previously solved by manually dividing the environment into zones. In this paper, an automatic partition of the environment is proposed to increase the localization accuracy in large environments. To do so, a hierarchical partition of the environment is performed using K-Means and the Caliński-Harabasz Index. Then, different classification techniques have been compared to achieve high localization rates. The new approach is tested in a real environment with more than 200 access points and 133 topological positions, obtaining an overall increase in the accuracy of approximately 10% and reducing the mean error to 2.45 metres.

Keywords: WiFi indoor localization, large environments, learning algorithms, clustering, classification

1 INTRODUCTION

Recent years have seen a rapid growth of smartphone and tablet applications [1]. Many of these applications make use of the localization capabilities of these devices and are emerging in very different areas: medical staff

and equipment localization [2, 3], medical assistance [4], inventory control at warehouses [5], robotics [6], assistance and guidance for disabled people [7], guidance at museums or public buildings [8], security [3], etc. All these applications require accurate indoor localization for the strategic planning of the navigation or to provide guidance to the final target. Traditionally, this localization has been carried out through GPS [9] which provides accurate localization when working outdoors. Unfortunately, satellite signals are attenuated and scattered by roofs, walls and other elements making indoor GPS localization not suitable. Thus, providing indoor localization requires the use of other technologies.

Different technologies are being used to provide indoor localization: infrared [10], ultrasound [11], laser [12], computer vision [13], radio frequency (RF) [14, 15] or cellular communication [16]. Among them, WiFi technology is arising as one of the most popular. The use of WiFi for indoor localization is motivated by its two main advantages: WiFi is widely deployed in almost every building and measuring the WiFi signal is free of charge even for private networks. Unfortunately, it has also some disadvantages: although the Received Signal Strength (RSS) decays logarithmically on free space, the multipath effect, obstacles and the small scale effect [17] make the RSS a complex function of the distance. In addition, the presence of people heavily affects the RSS absorbing part of the electromagnetic signal [18]. As a result, it is very difficult to model the RSS in indoor environments and propagation models, which have been proved very effective tools outdoors, are not generalizable and hard to adjust indoors. For this reason, most of indoor WiFi localization systems rely on a pre-learned set of fingerprints to infer the position of the device.

Generally, this kind of systems provide localization using a map as reference. Two map representations have been traditionally used: discrete and continuous. On a discrete map representation, the environment is divided into discrete positions and the localization is obtained in an estimation stage comparing the measures with a previously stored pattern (known as radio map or fingerprint database) [14, 19]. When the discrete positions are selected based on their topological significance it is called a topological representation. On a continuous map representation the environment is considered continuous and the position is obtained updating a probabilistic distribution of the position through action and propagation models as in particle filters [20, 21]. Continuous maps are more often used in robotics where the actuation and motion models are known, although some attempts have been made to model the human movement using Inertial Measurement Units as described in [22].

Topological representations [23–25] discretise the environment using nodes that correspond to a differentiating feature of the environment. These approaches have been especially useful in WiFi-based localization systems where no movement models are available and topological information is more

relevant than metric one (e.g. been at the doorway of office 15 versus being at coordinates x,y,z).

In a previous work, a WiFi-based localization system that uses a fuzzy rule-based classifier [26,27] was designed to estimate the topological position of WiFi devices in indoors [28]. Nevertheless, even though the localization system reported good performance dealing with small sized environments, a decrease in the performance has been observed when the number of positions and Access Points (APs) increased.

In this paper, the challenge of designing a WiFi localization system for large environments, crowded with APs not deployed for localization purposes, will be faced. This kind of environments have been previously neglected in the literature, where most of the proposed systems have been tested in small environments with a low number and a very uniform distribution of APs. This paper extends our previous work by automatically creating a hierarchical partition of the environment to simplify the training and classifying stages at the cost of increasing the number of required classifiers, but reducing the complexity of each one of them. In this new approach the system will have to determine in which of the partitions the device is localized to perform the final localization in a later stage. The effectiveness of this hierarchical approach was previously tested with a manual partition of a simple environment in [29] with performance improvements of 5% in the final localization accuracy. In this new approach the well-known K-Means clustering algorithm [30] is used along with the Caliński-Harabasz Index [31] to automatically create a hierarchical partition of the environment. This new hierarchical approach has been tested in a real multi-floor environment. It has reported a higher improvement than the system using manual partitions, without the need of human intervention.

The remaining of the paper is as follows. Section 2 describes the architecture of the system. Results and discussion are addressed in Section 3. Finally, Section 4 highlights the main conclusions and future work.

2 HIERARCHICAL WIFI-BASED LOCALIZATION SYSTEM

This section presents a description of the proposed localization system. The main objective is to achieve high accuracy even when working in large environments. To do so, the system will create a hierarchical partition of the environment with the objective of improving the localization task by reducing the number of positions in each one of the partitions. First, the hierarchical partition of the environment will be created using a clustering algorithm. Then, different classifiers will be trained to localize the device through the different levels of the hierarchy. This way, the device will be first located inside the higher subzones, to finally decide the position of the device inside the lower ones. A block diagram of the entire system is shown in Figure 1.

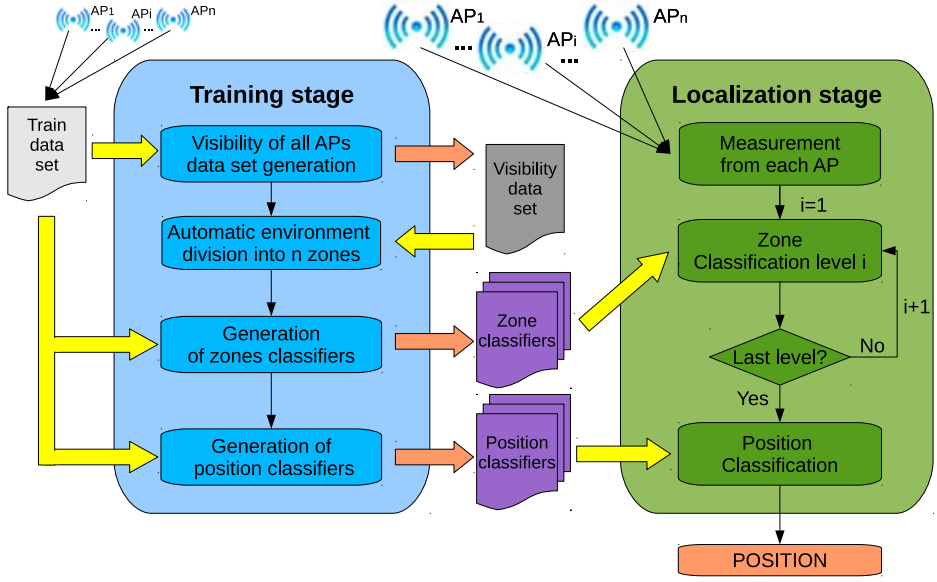


FIGURE 1
General architecture of the system.

This block diagram will be thoroughly explained in the next subsections. First, we will describe both training and localization stages. Then, the used learning algorithms will be briefly presented.

2.1 Training stage

The goal is to obtain a hierarchical tree by dividing the environment into zones. For each zone, a specific classifier will be trained to distinguish between the different zones (zone classifiers) and, in the lowest level of each one of the tree branches, one classifier will be trained to distinguish between the different positions (position classifiers). The training stage consists of the following steps:

- **Visibility dataset generation**: First, RSS is measured for every position of the environment and stored in $RSS_{TRAINDATA}$ (Equation 1).

$$RSS_{TRAINDATA} = \begin{pmatrix} RSS_{AP_1}(P_1,1) & RSS_{AP_2}(P_1,1) & \dots & RSS_{AP_Z}(P_1,1) \\ \vdots & \vdots & & \vdots \\ RSS_{AP_1}(P_1,T) & RSS_{AP_2}(P_1,T) & \dots & RSS_{AP_Z}(P_1,T) \\ RSS_{AP_1}(P_2,1) & RSS_{AP_2}(P_2,1) & \dots & RSS_{AP_Z}(P_2,1) \\ \vdots & \vdots & & \vdots \\ RSS_{AP_1}(P_2,T) & RSS_{AP_2}(P_2,T) & \dots & RSS_{AP_Z}(P_2,T) \\ RSS_{AP_1}(P_Y,1) & RSS_{AP_2}(P_Y,1) & \dots & RSS_{AP_Z}(P_Y,1) \\ \vdots & \vdots & & \vdots \\ RSS_{AP_1}(P_Y,T) & RSS_{AP_2}(P_Y,T) & \dots & RSS_{AP_Z}(P_Y,T) \end{pmatrix} \quad (1)$$

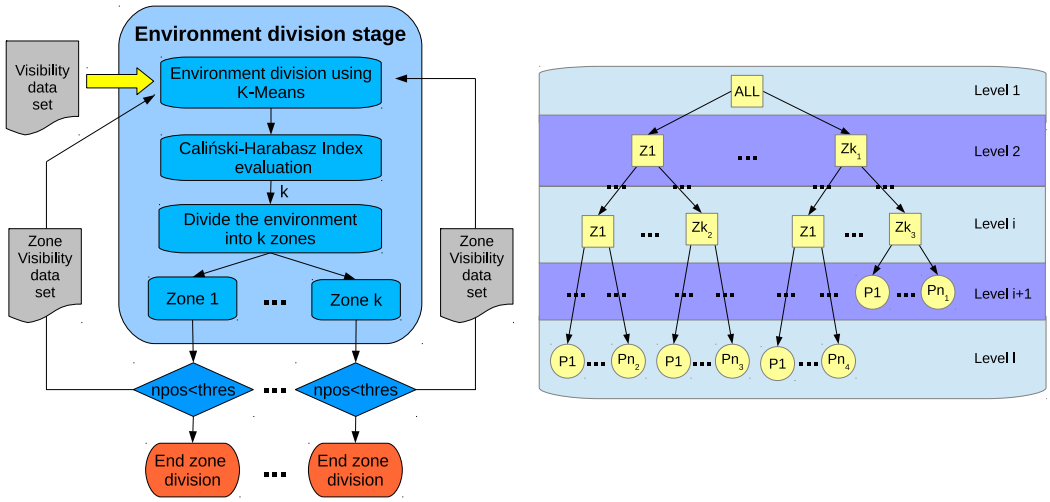
where Z is the number of APs, Y is the number of positions and T is the number of samples collected per position.
The division of the environment is performed using the so-called visibility. The visibility of an AP (AP_i) at a certain position (P_j) is defined by Equation 2:

$$VIS_{AP_i}(P_j) = \frac{1}{T}\sum_{t=1}^{T} d_{ij}(t), \quad d_{ij}(t) = \begin{cases} 1 & , \quad RSS_{AP_i}(P_j,t) > RSS_{thres} \\ 0 & , \quad otherwise \end{cases} \tag{2}$$

$VIS_{AP_i}(P_j)$ is computed as the percentage of samples that were collected with $RSS_{AP_i}(P_j, t)$ greater than a predefined threshold RSS_{thres}. Currently, this threshold is set to the minimum value, this way the sample t is taken into account for visibility purposes for any $RSS_{AP_i}(P_j, t)$. In the future, this threshold could be used to decrease the visibility of those APs with low RSS.
Once the visibility of all APs for each position is evaluated, the visibility dataset ($VIS_{TRAINDATA}$) is generated as described by Equation 3:

$$VIS_{TRAINDATA} = \begin{pmatrix} VIS_{AP_1}(P_1) & VIS_{AP_2}(P_1) & \dots & VIS_{AP_Z}(P_1) \\ VIS_{AP_1}(P_2) & VIS_{AP_2}(P_2) & \dots & VIS_{AP_Z}(P_2) \\ \vdots & \vdots & & \vdots \\ VIS_{AP_1}(P_Y) & VIS_{AP_2}(P_Y) & \dots & VIS_{AP_Z}(P_Y) \end{pmatrix} \tag{3}$$

- **Automatic environment partition**: The environment is automatically divided into zones using K-Means clustering algorithm [30] and the Caliński-Harabasz Index [31] over $VIS_{TRAINDATA}$. Figure 2(a) depicts the flow diagram of the applied procedure. The environment is iteratively divided into zones, in a hierarchical partition that can be represented by a tree (Figure 2(b)). Each zone Z_k is divided into k new sub-zones through K-Means, being the value of k determined by the Caliński-Harabasz Index. A zone is no further divided if it has less than 8 positions. This threshold has been experimentally selected.
- **Zone classifiers training**: Once the environment is divided into hierarchical zones, a classifier is built over the train data with the aim of distinguishing between the different zones belonging to the same level (squares in Figure 2(b)). Three classification algorithms (K-NN [32], FURIA [33] and SVM [34]) have been tested as classifiers.
- **Position classifiers training**: Another classifier is trained for each zone in the lowest level of the tree branches. These classifiers find the closest



(a) Partition diagram. (b) Environment division tree.

FIGURE 2
Partition procedure.

topological position to the current location among all positions belonging to the lower zones (circles in Figure 2(b)). Again, K-NN, FURIA and SVM are used as classifiers.

2.2 Localization Stage

In this stage the WiFi device will obtain its current position using the RSS from all visible APs. The set of classifiers trained in the previous section are now used to hierarchically localize the device, first in the higher zones and, at the end, determining the position of the device in the lower ones. The localization stage comprises two steps as shown in Figure 1:

- **Measurement**: Using a WiFi device, 4 RSS samples are collected from every AP and an averaged RSS sample is generated. This value has been experimentally selected from the analysis carried out in [28].
- **Classification**: The averaged sample is classified through the different levels of the hierarchy previously built in the training stage. Starting from the first level of zone classifiers, the system finds out the zone the sample belongs to. Then, the sample is classified again using the second level of classifiers corresponding to the zone previously identified. This procedure continues until the lowest level in the tree branch is reached. At the end, the estimated position of the WiFi device is determined by the position classifier associated to the lowest zone identified in the previous step.

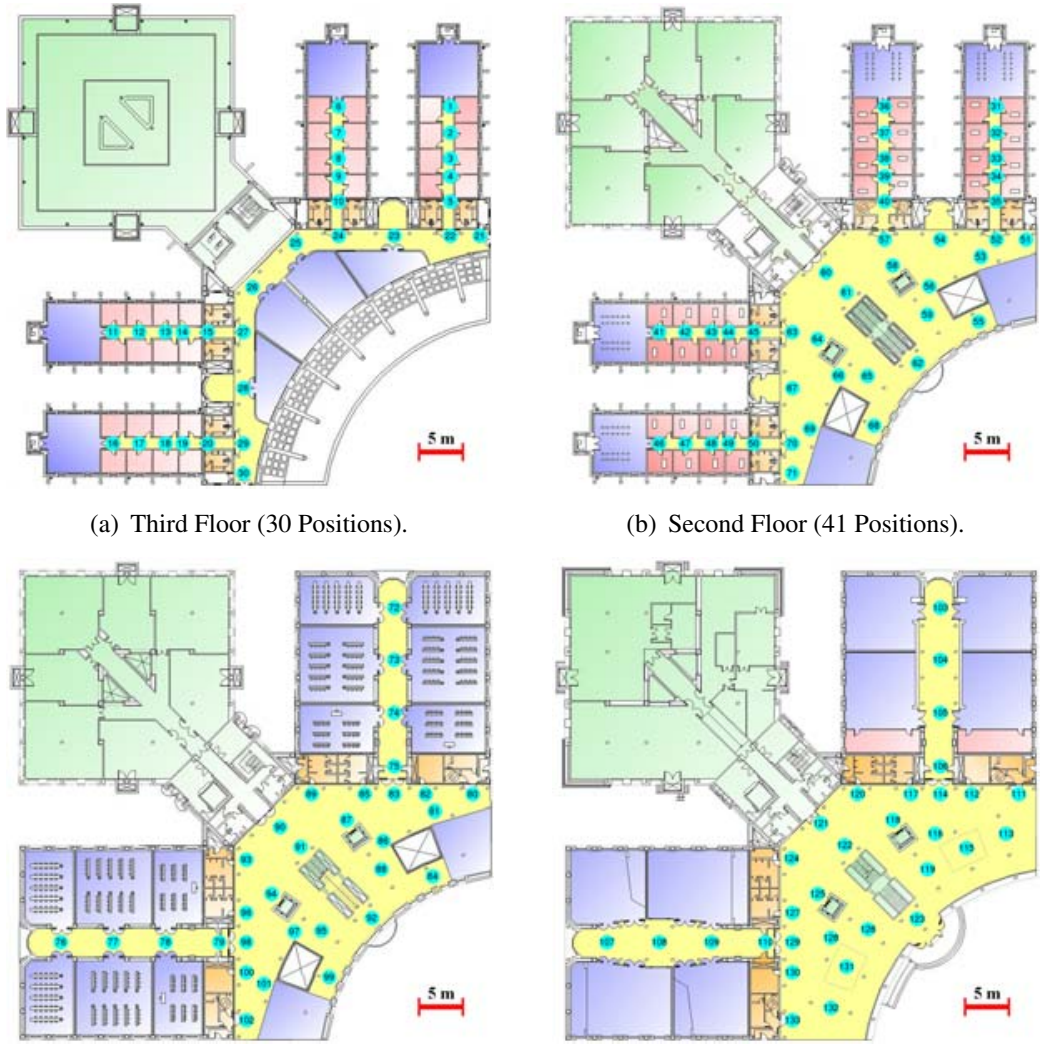
2.3 Learning Algorithms

This section provides a brief revision of the algorithms tested in our system.

- ***Clustering:*** K-Means clustering algorithm [30] along with the Caliński-Harabasz Index [31], also known as Variance Ratio Criterion (VRC), is used to obtain the hierarchical partition of the environment. The objective is to create a partition of the environment maximizing intra-cluster similarity. Setting the right number of clusters is a key task. To do so, the VRC performs a quantitative evaluation of clusters looking for compact and well-separated clusters within the feature space.
  Caliński-Harabasz Index has been chosen since it is one of the criteria providing the highest hit rates while having the lowest computational complexity as stated in the study carried out in [35].
- ***Classification:*** As explained in Sections 2.1 and 2.2, three different kinds of classifiers have been tested to classify the RSS measures into zones at each level and positions at the lowest level: Instance-based, rule induction and kernel-based classifiers.
  Among all existent **instance-based** classifiers, K Nearest Neighbours (K-NN) [32] was selected because it is usually used as baseline to compare with indoor WiFi localization systems [36,37]. K-NN was used in RADAR [14] which is world-wide recognized as one of the pioneers in the research field of WiFi indoor localization.
  Rule induction classifiers have been proved as a powerful tool to deal with noisy data [38]. We have selected the Fuzzy Unordered Rule Induction Algorithm (FURIA) [33]. It is a fuzzy modelling method which extends the well-known RIPPER algorithm [39], a state-of-the-art rule learner, while preserving its advantages, such as simple rule sets.
  Finally, we have chosen the Support Vector Machines (SVM) [34] as the most outstanding **kernel-based** classifier.
  The classifiers were implemented using K-NN, FURIA and SVM algorithm versions provided by the data mining tool Weka [40,41].

3 EXPERIMENTAL ANALYSIS

The hierarchical approach has been tested in a complex real-world environment. The experiments have been performed on the west wing of the Polytechnic School at the University of Alcalá (UAH) (Figure 3). The environment is made up of four floors with a surface of $2400m^2$ each. In our experiments we have detected 216 APs that were deployed over the environment with the aim of providing Internet access to the students but disregarding localization purposes. Notice that, in some related work the APs are deliberately placed for localization purposes what makes easier the localization task. Our system performs the localization using the RSS from the APs with no prior knowledge about their physical location. We have considered 133



(a) Third Floor (30 Positions).

(b) Second Floor (41 Positions).

(c) First Floor (31 Positions).

(d) Ground Floor (31 Positions).

FIGURE 3
UAH test-bed environment.

significant topological positions (distributed over the four floors) represented by circled numbers in Figure 3.

With the aim of evaluating the scalability of our proposal two different scenarios have been tested. Table 1 summarizes the main characteristics of both of them. The tests have been carried out with a laptop computer using its internal Wireless device acquiring 1 sample per second. Two datasets (train and test) have been collected on different days, one week apart, under real conditions. Each dataset has 60 samples per position and per AP.

3.1 Simple scenario. Small test-bed environment

To illustrate the simple scenario division obtained by the proposed system an environment division tree has been used (Figure 4). The horizontal dotted lines show the division between the different levels, the squares represent the

Scenario	Floors	N Positions	Visible APs
Simple	3^{rd} Floor	30	105
Complete	Four floors	133	216

TABLE 1
Main characteristics of the scenarios used to test the hierarchical approach.

zone classifiers and the circles denote the position classifiers, as explained in Section 2.1, Figure 2(b). The number under the circles correspond to the number of positions of the corresponding zone. Finally, the lines joining the nodes represent the hierarchy between the different zones, showing the number of subzones in which a zone is divided. As can be seen, the scenario has been divided in 4 different levels, obtaining 6 final zones with 3 to 7 positions each.

After dividing the environment, the zone and position classifiers are trained for each zone. Then, the test data are classified through the different levels until an inferred position is obtained for each sample.

Figure 5 summarizes the results achieved by the proposed system. The Y axis represents the accuracy of the system, while the X axis represents the number of levels in which the environment is divided. "1 Level" means all positions have been classified using only one classifier, without dividing the environment, while the maximum number of levels means that the whole hierarchical partition, shown in Figure 4, has been used for the classification steps. The results corresponding to the intermediate number of levels are shown just for comparison purposes and are obtained stopping the division process once the corresponding level is reached.

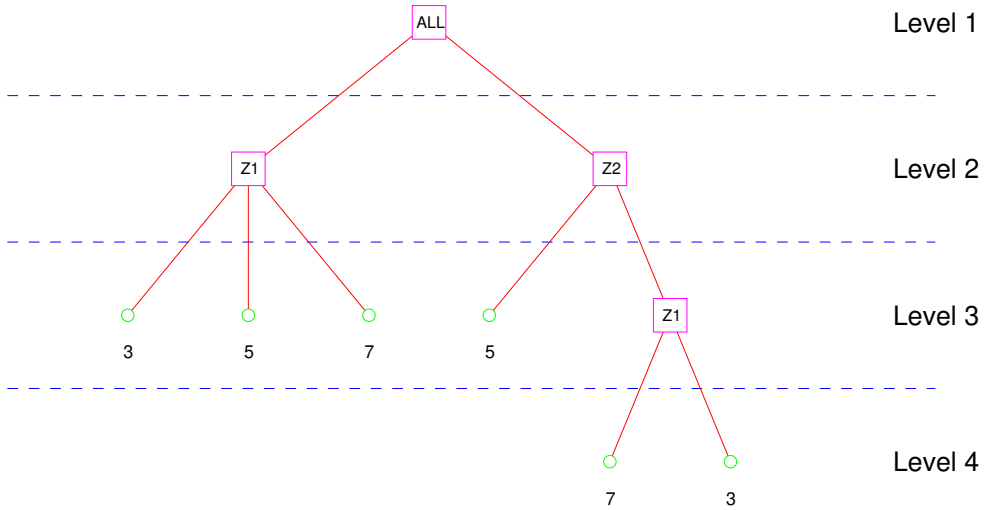
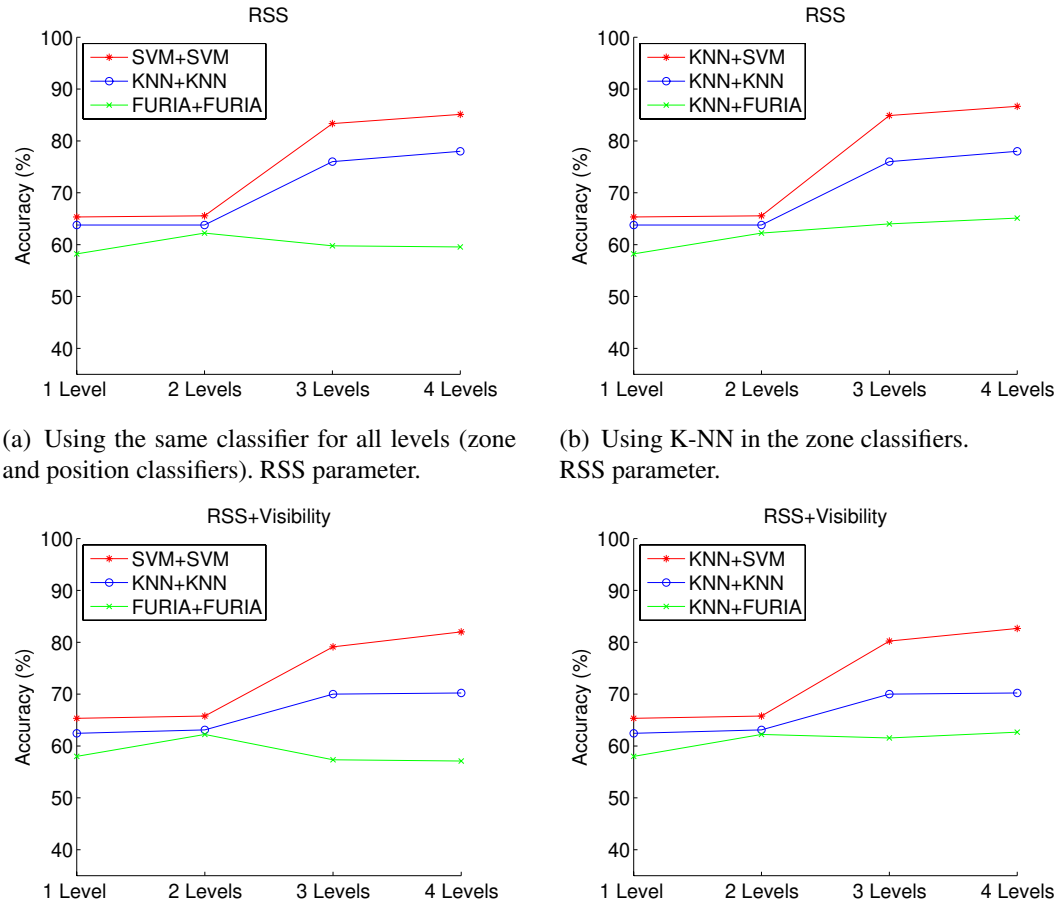


FIGURE 4
Simple scenario division (30 positions in the 3^{rd} floor).



(a) Using the same classifier for all levels (zone and position classifiers). RSS parameter.

(b) Using K-NN in the zone classifiers. RSS parameter.

(c) Using the same classifier for all levels (zone and position classifiers). RSS and Visibility parameters.

(d) Using K-NN in the zone classifiers. RSS and Visibility parameters.

FIGURE 5
Results in the simple scenario (30 positions in the 3^{rd} floor).

Four pictures are plotted, each one illustrating different approaches. On the one hand, the two pictures on top (5(a) and 5(b)) depict the results when using the RSS alone. On the other hand, the two pictures at the bottom (5(c) and 5(d)) correspond to the case when the RSS and the Visibility are used. The pictures on the left side of the figure (5(a) and 5(c)) show the results using the same classifier (FURIA, K-NN or SVM) at every level of the classification hierarchy, while the pictures on the right side of the figure (5(b) and 5(d)) show the results using K-NN in all the zone classifiers and FURIA, K-NN and SVM only at the last classification level (position classifiers).

As can be seen in Figure 5, accuracy increases with the number of levels, except using the FURIA classifier. Accuracy remains almost the same with the first division of the environment into two levels, but it significantly increases with the next hierarchical partition (three levels). Finally, using four levels slightly increases accuracy. No matter the selected classification

	Single Classifier	Hierarchical Classification	Improvement
FURIA	RSS: 58.22%	SCAL: 59.56%	1.34%
		ZCK-NN: 64.89%	6.67%
	RSS + Visibility: 58.00%	SCAL: 57.11%	-0.89%
		ZCK-NN: 62.44%	4.44%
K-NN	RSS: 63.78%	SCAL: 77.56%	13.78%
		ZCK-NN: 77.56%	13.78%
	RSS + Visibility: 62.44%	SCAL: 71.56%	9.12%
		ZCK-NN: 71.56%	9.12%
SVM	RSS: 65.33%	SCAL: 85.11%	19.78%
		ZCK-NN: 86.44%	**21.11%**
	RSS + Visibility: 65.33%	SCAL: 82.00%	16.67%
		ZCK-NN: 82.67%	17.34%

TABLE 2
Summary of results in the simple scenario (30 positions in the 3^{rd} Floor).

technique, adopting the hierarchical approach leads to an improvement of the accuracy versus the "1 Level" except for one case (FURIA+FURIA).

Using the RSS (Figures 5(a) and 5(b)) achieve better results than using the RSS and the Visibility (Figures 5(c) and 5(d)) in all the cases. K-NN and specially SVM significantly increase accuracy using the hierarchical approach and both clearly overcome FURIA. Using K-NN in the zone classifiers (Figure 5(b) and 5(d)), FURIA is able to get better results but they are still worse than those obtained by K-NN and SVM.

Table 2 summarizes the results achieved in the simple scenario when considering the different algorithms and configurations.

The results labelled as "Single Classifier" are the results in the case of classifying all the positions without dividing the environment. The column entitled as "Hierarchical Classification" shows the results achieved after applying the proposed hierarchical approach (all the four levels). SCAL stands for "Same Classifier for All Levels" and it reports the accuracy using the same classifier (FURIA, K-NN or SVM) at every level. ZCK-NN means "Zone Classifiers using K-NN" and it reports the accuracy using K-NN in all the zone classifiers and FURIA, K-NN or SVM (the one appearing in the first column) in the position classifiers at the lowest level of the hierarchy. The last column, "Improvement", gives the difference between the two previous columns, which is the increase (or decrease) in accuracy as result of applying the proposed hierarchical localization in contrast to the non-hierarchical approach.

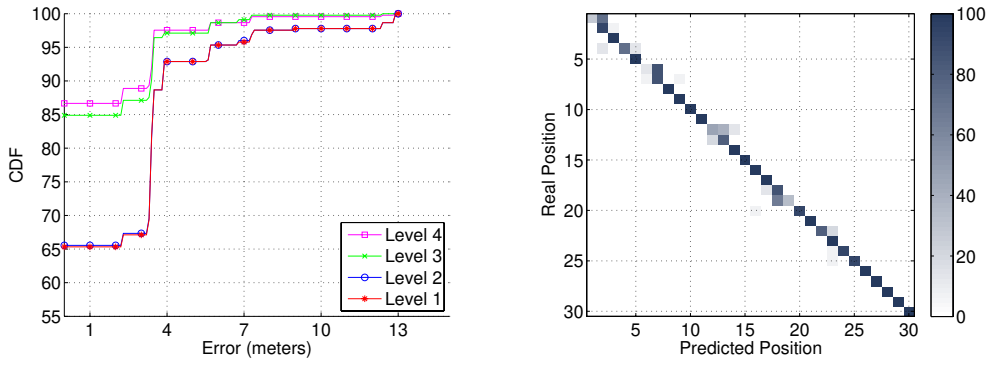
The following conclusions can be drawn after looking at Table 2:

- The hierarchical classification approach clearly overcomes the single classifier approach. In all cases, except considering FURIA using the RSS and

the Visibility, there is an improvement. Moreover, no matter the selected classification algorithm there is always at least one configuration yielding a minimum improvement of 6%.

- The highest accuracy is obtained when using the RSS alone.
- With respect to FURIA and SVM, accuracy increases using K-NN in all zone classifiers and FURIA or SVM only at the lowest level of the hierarchy (position classifiers) versus using FURIA or SVM in all the classifiers. Such behaviour was expected since the environment was divided into zones using K-Means, the "equivalent" clustering algorithm to K-NN.
- The highest accuracy (86.44%) and improvement (21.11%) are obtained by SVM using the RSS, K-NN in all zone classifiers and SVM only at the lowest level of the hierarchy (position classifiers).

Since reporting the accuracy expressed in terms of the misclassification rate only may be misleading for the evaluation of localization systems, Figure 6 shows the Cumulative Distribution Function (CDF) along with the confusion matrix for the configuration providing the highest accuracy. The CDF (Figure 6(a)) shows an analysis of the distance to the real positions in the different levels of the hierarchical system. As can be seen, the error decreases as the number of levels increases, obtaining 95% of the classified samples with an error under 4 metres. The confusion matrix (Figure 6(b)) details the predicted positions by the system related to the positions where the device really was. Looking at the figure, it can be seen that most of the classification errors occur within the nearest positions. Notice that, since we perform a topology-based indoor localization, the minimum error in distance depends on the minimum distance between the topological positions (2.25 metres in this scenario).



(a) CDF. (b) Confusion Matrix.

FIGURE 6
CDF and confusion matrix using ZCK-NN with SVM and RSS parameter. Simple scenario (30 positions in the 3^{rd} floor).

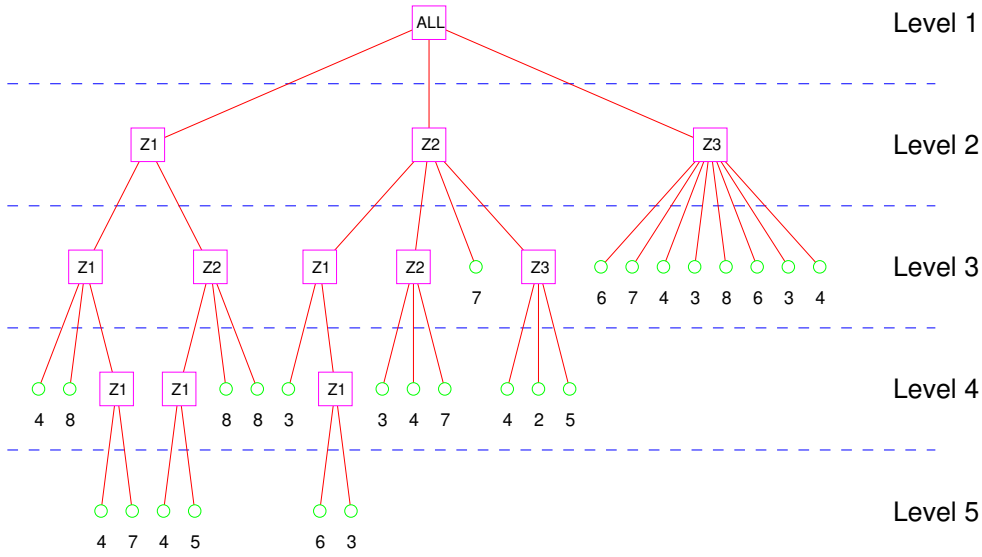


FIGURE 7
Complete scenario division (133 positions in the four floors).

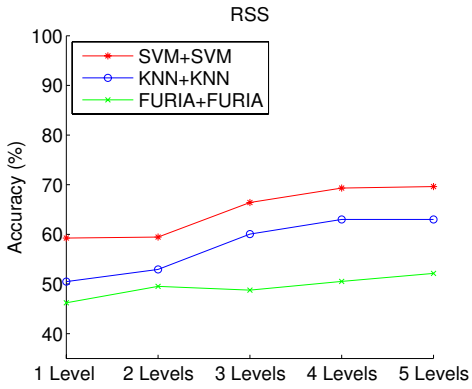
Finally, the mean distance to the real position is 4.03 metres for the misclassified samples and 0.55 metres taking into account all (both correctly and incorrectly classified) samples.

3.2 Complete scenario. Large test-bed environment

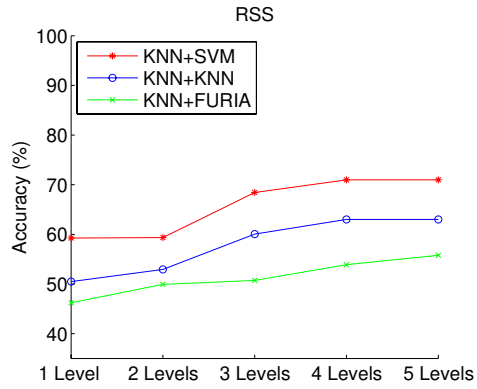
Figure 7 illustrates the environment division tree obtained by the proposed hierarchical localization system in the complete scenario (all the four floors depicted in Figure 3). As can be seen, the scenario has been divided in 5 different levels, obtaining 26 position zones with 2 to 8 positions each.

Figure 8 summarizes the results obtained by the proposed system. The format of this figure is the same as the one described for the simple scenario. The Y axis represents the accuracy of the system, while the X axis represents the number of levels in which the environment is divided ("1 Level" means the environment has not been divided while "5 levels" means that the environment has been fully divided following the proposed hierarchical approach.

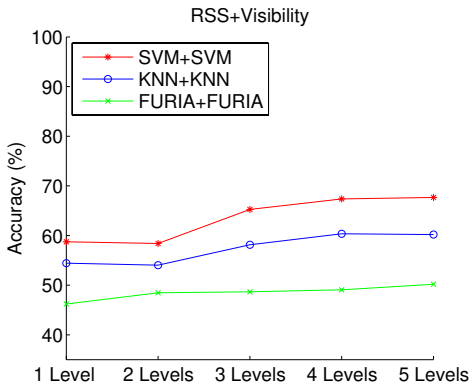
Four pictures are plotted, each one focusing on different learning algorithms and datasets. On the one hand, the two pictures on top (8(a) and 8(b)) depict the results using the RSS alone. On the other hand, the two pictures at the bottom (8(c) and 8(d)) correspond to the case when the RSS and the Visibility are used. The pictures on the left side of the figure (8(a) and 8(c)) show the results using the same classifier (FURIA, K-NN or SVM) at every level of the classification hierarchy, while the pictures on the right side of the figure (8(b) and 8(d)) show the results using K-NN in all the zone classifiers and FURIA, K-NN and SVM at the last classification level (position classifiers).



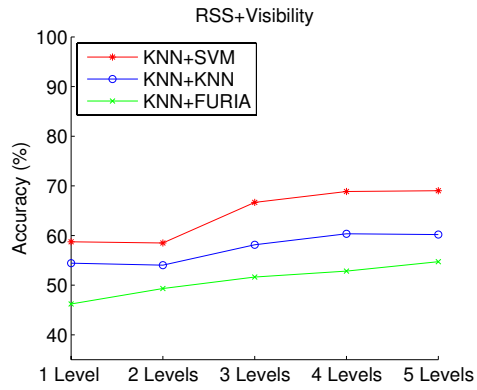
(a) Using the same classifier for all levels (zone and position classifiers). RSS parameter.



(b) Using K-NN in the zone classifiers. RSS parameter.



(c) Using the same classifier for all levels (zone and position classifiers). RSS and Visibility parameters.



(d) Using K-NN in the zone classifiers. RSS and Visibility parameters.

FIGURE 8
Results in the complete scenario (133 positions in the four floors).

As it can be appreciated in Figures 5 and 8, the general trend is the same, accuracy increases with the number of levels. Although in some intermediate levels accuracy decreases, at the end the final accuracy always goes up. Even though, K-NN and SVM seem to work better than FURIA, it is worthy to note that no matter the selected classification technique, adopting the hierarchical approach yields to an accuracy improvement of at least 5% versus the case labelled as "1 Level" (non-hierarchical approach).

Table 3 summarizes the accuracy results achieved in the complete scenario using the different algorithms and configurations. The format of this table is the same as the one described for Table 2 in the case of the simple scenario.

The following conclusions can be drawn after analysing Table 3:

- The hierarchical approach always overcomes the non-hierarchical one. In all cases the improvement is positive and in most of them greater than 5%.

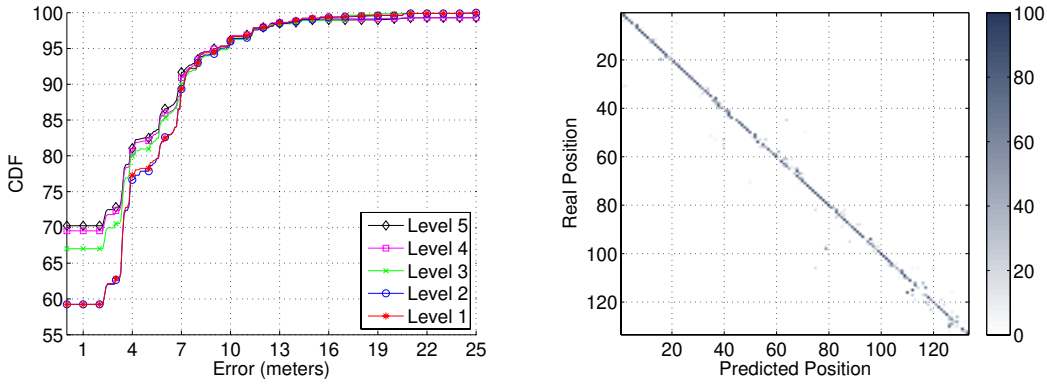
	Single Classifier	Hierarchical Classification	Improvement
FURIA	RSS: 46.22%	SCAL: 51.43%	5.21%
		ZCK-NN: 51.68%	5.46%
	RSS + Visibility: 46.22%	SCAL: 49.72%	3.50%
		ZCK-NN: 53.03%	6.81%
K-NN	RSS: 50.48%	SCAL: 59.20%	8.72%
		ZCK-NN: 59.20%	8.72%
	RSS + Visibility: 54.44%	SCAL: 61.15%	6.71%
		ZCK-NN: 61.15%	6.71%
SVM	RSS: 59.25%	**SCAL: 69.92%**	**10.67%**
		ZCK-NN: 66.42%	7.17%
	RSS + Visibility: 58.75%	SCAL: 67.92%	9.17%
		ZCK-NN: 67.87%	9.12%

TABLE 3
Summary of results in the complete scenario (133 positions in the four floors).

- Using the RSS and the Visibility the accuracy is higher than using the RSS alone when using K-NN in all zone classifiers and FURIA, K-NN or SVM only at the lowest level of the hierarchy (position classifiers) no matter the selected classifier. However, the highest improvement and accuracy is reached using the RSS alone (10.67% and 69.92%).
- Results are similar when using K-NN in all zone classifiers and FURIA, K-NN or SVM only at the lowest level of the hierarchy versus using FURIA, K-NN or SVM in all the classifiers. The FURIA results are slightly better using K-NN in all zone classifiers while the SVM accuracy is higher using SVM in all the classifiers.
- The highest accuracy (69.92%) and improvement (10.67%) are obtained by using SVM in all the classifiers and the RSS alone.

Figure 9 shows the CDF along with the confusion matrix for the configuration providing the highest accuracy. As can be seen looking at the CDF (Figure 9(a)), the distance to the real position decreases as the number of levels increases, obtaining 95% of the classified samples with an error under 9 metres. On the other hand, the confusion matrix (Figure 9(b)) shows that, although the distance error seems to be high, most of the classification errors occur within the nearest positions. It is important to highlight that, since we perform a topology-based indoor localization, the distance error depends on the minimum distance between the topological positions (2.25 metres in this scenario).

The mean distance to the real position is 7.30 metres for the misclassified samples and 2.45 metres taking into account all (both correctly and incorrectly classified) samples.

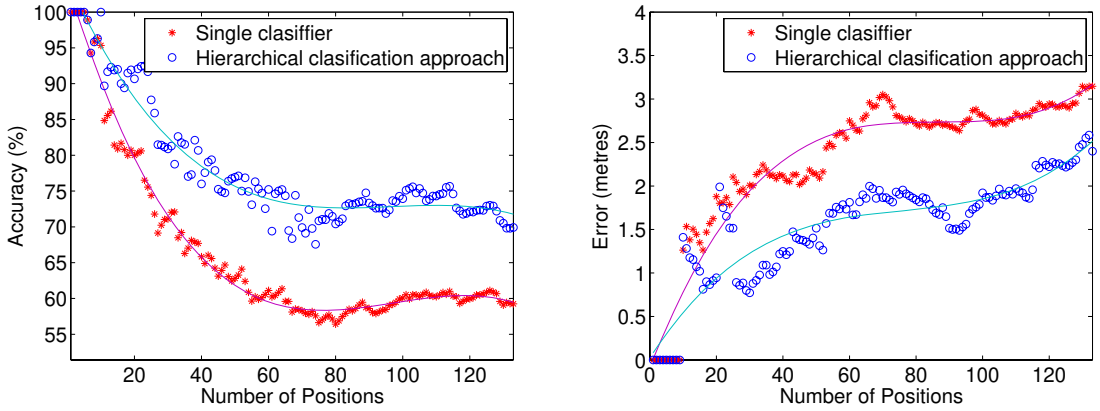


(a) CDF. (b) Confusion Matrix.

FIGURE 9
CDF and confusion matrix using SCAL with SVM and RSS parameter. Complete scenario (133 positions in the four floors).

It is important to remark that this scenario is larger than the previous one. In the simple scenario there are only 30 positions placed in the same floor while in the complete scenario there are 133 positions distributed over four different floors. However, even though the complexity of the problem has been increased (the number of positions is more than four times bigger), the hierarchical approach is still able to yield good results, achieving an accuracy close to 70%. This fact proves that the proposed hierarchical WiFi-based localization system works properly in large environments.

Finally, Figure 10 shows a comparison of the accuracy and mean error variation with the number of positions using the hierarchical approach versus using a single classifier (no environment division).



(a) Accuracy variation. (b) Mean error variation.

FIGURE 10
Accuracy and mean error variation with the number of positions. Single classifier vs. hierarchical approach using SCAL with SVM and RSS parameter.

System	Accuracy (%)	Mean error (m)
Hierarchical system using SVM	69.92%	2.45 m
RADAR [14]	50.48%	3.55 m
SAM-based system [42]	-	2.12 m

TABLE 4
Comparison of results achieved by different systems.

3.3 Comparing the proposed method with others in the literature

In this section, we will compare the results obtained by the proposed hierarchical system, the RADAR system [14] and the SAM-based (Smoothing and Mapping) system proposed in [42]. The RADAR system is an in-building user location system based on a KNN approach that has been traditionally used as baseline for WiFi localization algorithms. While the first two are fingerprint-based systems, the third one is a propagation model-based system that makes use of a robot odometry including a motion model. Therefore, this last system is expected to achieve better results thanks to the use of the motion information from the robot.

Table 4 shows the results obtained by the three systems on the UAH environment. RADAR and the hierarchical approach were tested using the same datasets, while the SAM-based system was tested on the same environment but with a different dataset. As can be seen, the proposed system clearly overcomes the RADAR system improving its accuracy about a 19% and reducing the mean error around a 30%. The SAM-based system achieved a mean error of 2.12 metres using a continuous map representation and a movement model to locate a robot. As explained before, both systems are not directly comparable since they address completely different problems, but the mean error achieved by the SAM-based system can be used as a reference of the minimum expectable error without using a movement model and a tracking and filtering algorithm.

4 CONCLUSIONS AND FUTURE WORK

This work has presented a generic method for the automatic environment division into hierarchical zones with the aim of improving the accuracy of topology-based WiFi localization systems in large environments. This hierarchical approach simplifies the classification task, reducing the number of outputs in the first level and the number of inputs and outputs in the following ones. With this approach the loss of accuracy when the number of positions in the environment increases is reduced. As a result, the mean error of the system is also reduced.

The proposal was tested in a real-world environment considering two different scenarios. The first one was a quite simple but highly illustrative scenario (of relatively small size), while the complete scenario was a much larger environment. The aim of using two different scenarios was to show how thanks to the proposed hierarchical approach our localization system was able to yield very good results no matter the size of the test-bed environment under consideration. Thus, we have proved our proposal successfully deals with indoor localization in large environments.

On the light of the results we can conclude that our proposal emerges as a powerful tool. The highest accuracy was around 86% in the simple scenario and it was slightly reduced in the complete scenario where it was around 70%. In both cases, the best results were reported when considering the RSS and SVM as position classifier at the lowest classification level of the hierarchy. The accuracy improvement due to the hierarchical approach was around 21% in the simple scenario and 11% in the complete scenario. Finally, 95% of the samples were classified with a distance to the real position under 4 metres getting a mean distance of 0.55 metres in the simple scenario, while in the complete scenario 95% of the samples were classified with a distance to the real position under 9 metres getting a mean distance of 2.45 metres. These results outperform the standard RADAR (reporting mean error of 3.55m) but they are also very competitive in comparison with a system performing probabilistic localization including continuous map representation and a movement model which is expected to yield the highest accuracy (reporting mean error of 2.12m).

It is important to highlight that it is not necessary to know where the APs are located to deploy the localization system. This aspect is especially interesting regarding its deployment in new unknown environments. Moreover, since the localization is performed directly on the device, the system can be safely used without dealing with privacy issues.

Moreover, we have proved the goodness of the proposed hierarchical localization approach even though we considered only basic well-established algorithms for each involved task (clustering, classification, and so on).

In the future, we will explore other more advanced classification techniques, for example the fuzzy rule-based classifier previously designed in [28] or a multiclassifier [43] in combination with our hierarchical approach. In addition, alternative and more advanced methods for finding out an optimal partition of the environment will be further analysed [44]. Finally, a procedure for selecting some of the APs will be tested. This AP selection may be made according to the visibility criteria introduced in this work paying attention to the top APs with the highest RSS.

ACKNOWLEDGEMENTS

This work has been funded by the "Ministerio de Economía y Competitividad" through TIN2011-29824-C02-01 and TIN2011-29824-C02-02 (ABSYNTHE project), as well as by the University of Alcalá through CCG2013/ EXP-066 (ROBOSHOP project) and by the Principality of Asturias Government under the project with reference CT13-53.

REFERENCES

[1] BI Intelligence, (2013). Cumulative app downloads since 2008. Accessed on June 2014, http://www.businessinsider.com/comparing-the-growth-of-top-app-markets-2013-7.

[2] G. Molina and E. Alba. (2011). Location discovery in wireless sensor networks using metaheuristics. *Applied Soft Computing*, 11(1):1223–1240.

[3] Ekahau, (2014). Wi-Fi tracking systems, RTLS and WLAN site survey. Accessed on June 2014, http://www.ekahau.com.

[4] E. López, R. Barea, L. M. Bergasa, and M. S. Escudero. (2004). A human-robot cooperative learning system for easy installation of assistant robots in new working environments. *Journal of Intelligent and Robotic System*, 40(3):233–265.

[5] Z. Zhao and X. Zhang. (2011). An RFID-based localization algorithm for shelves and pallets in warehouse. In *Proceedings of the International Conference on Transportation Engineering*, pages 2157–2162.

[6] M. A. Sotelo, M. Ocaña, L. M. Bergasa, R. Flores, M. Marrón, and M. A. García. (2007). Low level controller for a POMDP based on WiFi observations. *Robotics and Autonomous Systems*, 55(2):132–145.

[7] C. Benavente-Peces, M. Puente, A. Domínguez-García, M. Lugilde-Rodríguez, E. de la Serna, D. Miguel, and A. García. (2009). Global system for localization and guidance of dependant people: Indoor and outdoor technologies integration. In Ambient Assistive Health and Wellness Management in the Heart of the City, volume 5597 of *Lecture Notes in Computer Science*, pages 82–89.

[8] O. A. Hammadi, A. A. Hebsi, M. J. Zemerly, and J. W. P. Ng. (2012). Indoor localization and guidance using portable smartphones. In *Proceedings of the IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 3, pages 337–341.

[9] P. Enge and P. Misra. (1999). Special issue on GPS: The global positioning system. *Proceedings of the IEEE*, 87(1):3–172.

[10] R. Want, A. Hopper, V. Falcão, and J. Gibbons. (1992). The active badge location system. *ACM Transactions on Information Systems*, 10(1):91–102.

[11] N. B. Priyantha, A. Chakraborty, and H. Balakrishnan. (2000). The cricket location-support system. In *Proceedings of the Annual International Conference on Mobile Computing and Networking*, pages 32–43.

[12] R. Barber, M. Mata, M. J. L. Boada, J. M. Armingol, and M. A. Salichs. (2002). A perception system based on laser information for mobile robot topologic navigation. In *Proceedings of the Annual Conference of the IEEE Industrial Electronics Society*, volume 4, pages 2779–2784.

[13] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer. (2000). Multi-camera multi-person tracking for easyliving. In *Proceedings of the IEEE International Workshop on Visual Surveillance*, pages 3–10.

[14] P. Bahl and V. N. Padmanabhan. (2000). RADAR: An in-building RF-based user location and tracking system. In *Proceedings of the Annual Joint Conference of the IEEE Computer and Communications Societies*, pages 775–784.

[15] T. García-Valverde, A. García-Sola, H. Hagras, J. Dooley, V. Callaghan, and J. A. Botía. (2013). A fuzzy logic-based system for indoor localization using WiFi in ambient intelligent environments. *IEEE Transactions on Fuzzy Systems*, 21(4):702–718.

[16] B.B. Parodi, H. Lenz, A. Szabo, H. Wang, J. Horn, J. Bamberger, and D. Obradovic. (2006). Initialization and online-learning of RSS maps for indoor / campus localization. In *Proceedings of the IEEE/ION Position, Location, And Navigation Symposium*, pages 164–172.

[17] M. Youssef and A. Agrawala. (2003). Small-scale compensation for WLAN location determination systems. In *Proceedings of the IEEE Wireless Communications and Networking*, volume 3, pages 1974–1978.

[18] A. Bahillo, R. M. Lorenzo, S. Mazuelas, P. Fernández, and E. J. Abril. (2009). Assessment of the shadow caused by the human body on the personal RF dosimeters reading in multipath environments. *Biomedical Engineering*, pages 133–144.

[19] M. Youssef, A. Agrawala, and A. U. Shankar. (2003). WLAN location determination via clustering and probability distributions. In *Proceedings of the IEEE Pervasive Computing and Communication*, pages 143–150.

[20] D. Fox, J. Hightower, L. Liao, D. Schulz, and G. Borriello. (2003). Bayesian filtering for location estimation. *IEEE Pervasive Computing*, 2(3):24–33.

[21] J. Hightower and G. Borriello. (2004). Particle filters for location estimation in ubiquitous computing: A case study. In *Proceedings of International Conference on Ubiquitous Computing*, pages 88–106.

[22] O. Woodman and R. Harle. (2008). Pedestrian localisation for indoor environments. In *Proceedings of the International Conference on Ubiquitous Computing*, pages 114–123.

[23] B. J. Kuipers and Y. Byun. (1988). A robust, qualitative method for robot spatial learning. In *Proceedings of the National Conference on Artificial Intelligence*, pages 774–779.

[24] D. M. Kortenkamp. (1993). *Cognitive maps for mobile robots: A representation for mapping and navigation*. PhD thesis, University of Michigan.

[25] E. López, L. M. Bergasa, R. Barea, and M. S. Escudero. (2005). A navigation system for assistant robots using visually augmented POMDPs. *Autonomous Robots*, 19(1):67–87.

[26] L. A. Zadeh. (1965). Fuzzy sets. *Information and Control*, 8:338–353.

[27] L. A. Zadeh. (1973). Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-3(1):28–44.

[28] J. M. Alonso, M. Ocaña, N. Hernández, F. Herranz, A. Llamazares, M. A. Sotelo, L. M. Bergasa, and L. Magdalena. (2011). Enhanced WiFi localization system based on soft computing techniques to deal with small-scale variations in wireless sensors. *Applied Soft Computing*, 11(8):4677–4691.

[29] N. Hernández, J. M. Alonso, M. Magro, and M. Ocaña. (2012). Hierarchical WiFi localization system. In *Proceedings of the International Workshop on Perception in Robotics, IEEE Intelligent Vehicles Symposium*, pages 21.1–21.6.

[30] J. B. MacQueen. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297.

[31] T. Caliński and J. Harabasz. (1974). A dendrite method for cluster analysis. *Communications in Statistics-Simulation and Computation*, 3(1):1–27.

[32] D. Kibler and D. Aha. (1987). Learning representative exemplars of concepts: An initial case study. In *Proceedings of the International Workshop on Machine Learning*, pages 24–30.

[33] C. Hühn and E. Hüllermeier. (2009). FURIA: An algorithm for unordered fuzzy rule induction. *Data Mining and Knowledge Discovery*, 19(3):293–319.

[34] C. Cortes and V. Vapnik. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

[35] L. Vendramin, R. J. G. B. Campello, and E. R. Hruschka. (2010). Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining*, 3(4):209–235.

[36] M. Youssef and A. Agrawala. (2008). The Horus location determination system. *Wireless Networks*, 14(3):357–374.

[37] B. Wu, C. Jen, and K. Chang. (2007). Neural fuzzy based indoor localization by Kalman filtering with propagation channel modeling. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, pages 812–817.

[38] J. Hühn and E. Hüllermeier. (2010). An analysis of the FURIA algorithm for fuzzy rule induction. In *Advances in Machine Learning I*, volume 262 of *Studies in Computational Intelligence*, pages 321–344.

[39] W. W. Cohen. (1995). Fast effective rule induction. In *Proceedings of the International Conference on Machine Learning*, pages 115–123.

[40] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. (2009). The weka data mining software: An update. *SIGKDD Explorations*, 11(1):10–18.

[41] I. H. Witten, E. Frank, and M. A. Hall. (2011). *Data Mining: Practical machine learning tools and techniques*. Data Management Systems Series. Morgan Kaufmann, third edition.

[42] F. Herranz. (2013). *Simultaneous Localization and Mapping using Range Only Sensors*. PhD thesis, University of Alcalá.

[43] K. Trawinski, J. M. Alonso, and N. Hernández. (2013). A multiclassifier approach for topology-based wifi indoor localization. *Soft Computing*, 17(10):1817–1831.

[44] J. M. Alonso, N. Hernández, and M. Ocaña. (2013). Wifigrams: Design of hierarchical WiFi indoor localization systems guided by social network analysis. In *Computer Aided Systems Theory - EUROCAST 2013*, volume 8112 of *Lecture Notes in Computer Science*, pages 9–16.