Indoor Localization based on Principal Components and Decision Trees in IEEE 802.15.7 Visible Light Communication Networks

David Sánchez-Rodríguez[1,2], Itziar Alonso-González[1,2], Carlos Ley-Bosch[1,2], Javier Sánchez-Medina[3], Miguel Quintana-Suárez[1] and Carlos Ramírez-Casañas[1,2]

Department of Telematics Engineering[1]
Institute for Technological Development and Innovation in Communications[2]
Institute for Cybernetics[3]
University of Las Palmas de Gran Canaria
e-mail: {david.sanchez, itziar.alonso, carlos.ley, javier.sanchez, mangel.quintana, carlos.ramirez}@ulpgc.es

***Abstract* - Indoor positioning estimation has become an attractive research topic due to the growing interest in location-aware services. Research works have been proposed on solving this problem by using wireless networks. Nevertheless, there is still much room for improvement in the quality of the proposed classification or regression models, i.e., in terms of accuracy or root mean squared error (RMSE). In the last years, the emergence of Visible Light Communication brings a brand new approach to high quality indoor positioning. Among its advantages, this new technology is immune to electromagnetic interference, and also, the variance of the received optical power is smaller than other RF based technologies. In this paper, we propose a fingerprinting indoor location estimation methodology based on principal components analysis (PCA) and decision trees as classification learner. The proposed localization methodology is based on the received signal strength from a grid of emitters multiple. PCA is used to transform all of that features into principal components, consequently reducing the data dimensionality, improving the interpretability of the resulting tree models and the overall computational performance of the proposed system. Along with the proposed method, we also share experimental results derived from the received signal strength values obtained from an IEEE 802.15.7 simulator developed by our research group. Results show that the system accuracy is slightly improved by range 1%-10% and the computation time by range 40%-50%, as compared to the system in which PCA is not carried out. The best tested model (classifier) yielded a 95.6% accuracy, with an average error distance of 2.4 centimeters.**

Keywords - Indoor Localization; Visible Light Communication; Decision Trees; Principal Components Analysis; Received Signal Strength.

I. INTRODUCTION

The present paper expands on the indoor localization system described in the original paper [1] proposing the use of principal components analysis (PCA) to improve the system accuracy while reducing the computational cost and carrying out some enhanced experiments.

Indoor localization has gained considerable attention over the past decade due to the emergence of numerous location-aware services. These new services have made it possible to use applications capable of sensing their location and dynamically adjusting their settings and functions [2]. Many indoor localization approaches based on globally deployed radiofrequency systems, such as Wireless Local Area Networks (WLAN), Bluetooth and Ultra-Wide Band (UWB), have been proposed, mainly because of their low cost and mature standardization state. Nevertheless, they usually deliver an accuracy of up to two meters, since hindered by multipath propagation [3]. On the other hand, Visible Light Communication (VLC) is experiencing a growing interest due to improvements in solid state lighting and a high demand for wireless communications. VLC can offer a higher positioning accuracy [4] mainly because of two reasons: this kind of networks is not affected by electromagnetic interferences and the received optical power is more stable than radio signals and can be accurately known. For example, authors in [5] proposed a system with a positioning error about 10 centimeters using a location code and a spatial power distribution map where the received signal strength (RSS) measurements are gathered 5 centimeters separation from each other.

In this paper, we propose an indoor location estimation method based on an ensemble model of decision trees, yielding and optimal tradeoff between accuracy (high) and variance (low), and the added value of being computationally efficient. In order to achieve this tradeoff, PCA is proposed to transform RSS features of a VLC network into principal components, consequently reducing the data dimensionality and improving the computational cost of the system. The main novelty of this work comes from the fact that the positioning systems based on decision trees and principal components have a lower computational complexity and high accuracy. Additionally, the proposed methodology is also novel in the use decision trees and principal components in IEEE 802.15.7 VLC networks for indoor location estimation.

The rest of the paper is organized as follows. Section II summarizes state of the art. In Section III, we describe our simulator that implements the IEEE 802.15.7 VLC standard. Next, in Section IV, we describe the ensemble model of decision trees used for VLC indoor location estimation. Section V describes the two phases of our indoor positioning method based on an ensemble model of decision trees where PCA is considered. In Section VI, we show experimental results that demonstrate the high accuracy of our approach

and its low computational complexity. Finally, we sum up the conclusions and we present the future work.

II. STATE OF THE ART

Indoor positioning techniques for VLC are mainly classified into two groups based on geometric properties: lateration and angulation [6]. Lateration techniques estimate the target location by measuring distances from the receiver to multiple LEDs base stations with known coordinates. The distances can be estimated involving the time of arrival (TOA), time difference of arrival (TDOA) and RSS measurements. On the other hand, with angulation techniques or angle of arrival (AOA) the target location is estimated by measuring angles to multiple base stations. Nevertheless, these techniques often require additional hardware, time synchronization between emitter and receiver, knowing every base station coordinates and extra computation.

A third kind of location techniques are the so called fingerprinting techniques, that combined with VLC can be an alternative to the aforementioned because they estimate positioning by matching online measured data with pre-measured location-related data, such as RSS. Hence, just RSS information is needed and extra sensors are unnecessary. As a matter of fact, fingerprinting is one of the most commonly used techniques for RF indoor location [7].

Localization based on fingerprinting is usually carried out in two phases. The first phase (off-line phase) consists on the sampling RSS measurements for every emitter and each reference location (VLC receiver). With that samples as training set, a positioning model is learned using a particular machine learning technique. During the second phase (on-line phase), the particular receiver position is estimated by using the learned model and the new RSS measurements.

Learned techniques based on decision trees are widely used in classification problems, and are often used in indoor localization. A decision tree is a sequence of branching operations based on comparisons of RSS values for each feature in the dataset. Depending on the training dataset size and the number of features (emitters or principal components), the depth of the tree can be high, and hence, the number of conditions to be evaluated could influence energy savings. Nevertheless, its computational complexity is considerably lower than the number of floating-point multiplication CPU cycles, where experimental results indicate that decimal64 multiplication with binary integer decimal (BID) encoding takes an average of 117 cycles using Intel's BID library [8]. Since no floating-point multiplication takes place to predict the location using decision trees, the computational complexity of our system is O(1). The latter is an extremely important characteristic if the localization system is designed for portable devices, where both processor power and energy availability are constrained. Hence, factors such as the battery power, computation cost, and the memory size need to be jointly considered. Thus, the reduction of the data dimensions leads to a decrease in the computational complexity. In addition, the performance can be further enhanced when the discarded information is redundant noise [9].

PCA is one of the most widely used techniques to carry out the reduction of the data dimensions. The central idea of PCA is to reduce the dimensionality of a dataset in which there are a large number of interrelated variables, while retaining as much as possible of the variation present in the dataset. This reduction is achieved by transforming to a new set of variables, the principal components, which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all the original variables [10].

Regarding to indoor localization, authors in [9] proposed a novel approach based on PCA which transforms RSS into principal components such that the information of all access points (APs) is more efficiently utilized. Instead of selecting APs for the positioning, the proposed technique changes the elements with a subset of principal components improvement of accuracy and reduces the online computation. The proposed approach delivers a significantly improved accuracy. The results show that the mean error is reduced by 33.75% and the complexity is decreased by 40%, as compared to the existing techniques.

On the other hand, in 2011, Institute of Electrical and Electronic Engineers (IEEE) published the IEEE 802.15.7 standard, which defines Physical (PHY) and Medium Access Control (MAC) layers for short-range wireless optical communications using visible light [11]. Within the last few years, many studies on VLC based positioning have been published. Nevertheless, to the best of our knowledge, to this date there is no any published indoor positioning research using this standard.

With the present work, our contribution is the following: we propose an ensemble model of decision trees based indoor positioning methodology, built of principal components from RSS, together with some promising results. We have carried out a wide experimentation and present results showing the achieved high accuracy and low computational complexity. Furthermore, we make use of the IEEE 802.15.7 standard on VLC to obtain RSS values, which may be a useful piece of information for other researchers and practitioners at this stage of (un)deployment of such standard.

III. SIMULATION MODEL BASED ON IEEE 802.15.7

We built our simulator using OMNET++ [12] simulation framework from the model developed by [13] designed for sensor networks based on the IEEE 802.15.4 standard, due to the similarities existing between IEEE 802.15.7 and IEEE 802.15.4 architectures.

OMNeT++ provides built-in support tools not only for simulating, but also for the analysis and visualization of simulation results. Several data can be chosen for simulation results, such as throughput, delay, packet loss and RSS.

The developed simulation model has been designed with the following premises:

- IEEE 802.15.7 star topology has been chosen, due to its importance and wide range of applications.

- For the MAC layer, we opted to use the superframe structure; since it allows the use of both contention (CAP) and no contention (CFP) access methods. In addition, the use of the superframe enables devices to enter the energy save state during the idle period.
- A VLC Personal Area Network (VPAN) identifier is assigned to each emitter in order to identify each coordinator (LED lamp).

Next subsections describe the most important features in our simulator, for a better comprehension of the presented results.

A. Optical channel model

The transmission medium is modeled as free space without obstacles. We chose the directed line of sight (LOS) link configuration to model the optical signal propagation, requiring a LOS between each device and the coordinator. We have considered only the direct component of the received signal to calculate the received power, despising the possible influence of reflections.

Frequency response of optical channel is relatively flat near Direct Current (DC), so the most important quantity for characterizing this channel is the DC gain H(0) [14], which relates the transmitted and received optical average power, see (1):

$$P_r = H(0) \bullet P_t \quad (1)$$

In VLC, received power can be expressed as the sum of LOS and non-LOS components. In directed LOS links, the DC gain can be computed fairly accurately by considering only the direct LOS propagation path. According to the results presented in [15], at least 90% of total received optical power is direct light in VLC when using a receiver field of view (FOV) of 60 degrees. Figure 1 shows an example of a directed LOS link.

An optical source can be modeled by its position vector, a unit-length orientation vector $\vec{o}_t$, transmission power P_t and a radiation intensity pattern $I(\theta,m)$ emitted in direction θ.

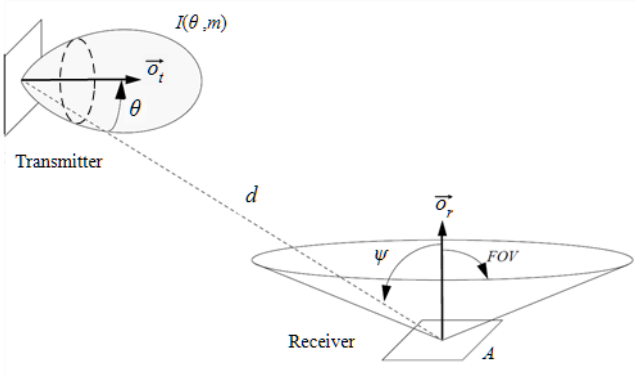

Figure 1. Directed LOS link configuration.

The m parameter is the mode number of the radiation lobe, which specifies the directionality of the source, and is related to the transmitter half power angle $\theta_{1/2}$. Similarly, a receiver is defined by its position, orientation $\vec{o}_r$, photo detector area A, and FOV (ψ_c). The angle formed between the optical incident signal and the orientation vector $\vec{o}_r$ is called the incident angle ψ. The maximum incident angle defines the receiver FOV.

Considering LOS propagation path, the DC gain can be calculated according to [14] as (2):

$$H(0)=(m+1)\bullet cos^m(\theta)\bullet A\bullet T_s(\psi)\bullet G(\psi)\bullet cos(\psi)/(2\pi d^2),\ 0\leq \psi \leq \psi_c$$

$$H(0)=0, \quad \psi > \psi_c \quad (2)$$

$T_s(\psi)$ is the optical filter signal transmission coefficient, $G(\psi)$ is the optical concentrator gain, d is the distance between transmitter and receiver.

The adopted optical channel model facilitates reaching high transmission speeds, since the effects of multipath distortion on the optical signal are not considered. Considering only the direct component of the signal has the additional benefit of improving the efficiency of the implemented simulation model. The computational load required to run simulations of scenarios with multiple nodes including the functionality of different layers of the architecture is reduced significantly.

To ensure the validity of our implemented model, we have configured all optical receivers using a 60 degrees FOV value (ψ_c).

TABLE I. PHY LAYER PARAMETERS

Parameter	Value
Transmission rate	1.25 Mbps
Optical clock rate	3.75 MHz
Coordinator optical transmission power (P_t)	15 W
Half Power Angle $\theta_{1/2}$	60º
Field of Vision (ψ_c)	60º
Photo detector area (A)	100 mm^2
Photo detector responsivity (R)	0.54 A/W
Optical concentrator gain ($G(\psi)$)	15
Optical filter transmission coefficient ($T_s(\psi)$)	1

B. PHY layer simulation parameters

Table I shows the main configuration parameters of PHY layer used in all simulation scenarios. We selected the PHY II operating mode, intended for both indoor and outdoor environments, using MCS-ID number 16, since support for the minimum clock and data rates for a given PHY is mandatory.

Because of the optical channel model used, transmitters' directivity is characterized by its half power angle, $\theta_{1/2}$ while receivers' directivity is defined by its FOV. According

to [16], both parameters are assigned a value of 60 degrees, to ensure validity of the implemented channel model, since the calculation of received optical power takes in account only the direct component of the signal.

In order to simplify the calculation process of the model, the values used for the concentrator gain ($G(\psi)$) and the transmission coefficient of the optical filter ($T_s(\psi)$) are set up as constant values, so they do not depend on the angle of incidence ψ.

The rest of the values selected to characterize VLC transmitters and receivers are commonly used values in literature, similar to those used in [17][18].

IV. Ensemble Model of Decision Trees

Indoor positioning has been a very active research area where several data mining techniques have proved useful to extract knowledge from raw data [19][20]. To solve this problem, in this paper we propose a general approach based on a classifier built as an ensemble model of decision trees.

Decision trees build classification models in the form of a tree structure. In general, they can handle both categorical and numerical data. A decision tree has internal nodes and leaf nodes. An internal node includes a condition or function of any feature of the dataset, which breaks down the dataset into several subsets, corresponding to two or more branches. Each leaf is assigned to a class, representing the classification decision. For instance, in the location problem, the received optical power from luminaries is used in the internal node conditions, and the locations or reference points are used in the leaf nodes. Samples are classified by navigating from the root of the tree down to a leaf, according to the outcome of the condition or function along the path [21].

On the other hand, ensemble models are methods that combine the capabilities of multiple models to achieve better prediction accuracy than any of the individual model could do on its own. Ensemble methods generate multiple base models, and the final prediction is produced as the result of a combination of them, in some appropriate manner, from the prediction of each base model. For instance, the output of each base model is weighted. The success of the ensemble model is based on the ability of generating a set of base models that make errors that are as uncorrelated as possible.

In our indoor localization method, we use a weak classifier based on the C4.5 algorithm [22] to generate a decision tree as a base model. Then, the adaptive boosting (AdaBoost) algorithm [23] is used to build an ensemble model based on previous base models, that is a location estimation model formed by multiple weighted decision trees. AdaBoost aims at improving the accuracy of the weak learner, by concentrating in the samples incorrectly classified by that one. In a previous work, we demonstrated that this combination of machine learning techniques provides excellent results for indoor localization when it is used in WLAN networks [24].

V. Indoor Localization Method

In this section, we describe our positioning method based on an ensemble model of decision trees, and it is divided into two phases. The first phase is the training phase (off-line phase). Coordinators send beacon frames and RSS samples are collected at reference locations (receivers) to build a dataset. From this dataset, the ensemble model is built. The second stage is the test phase (on-line phase) where a receiver infers its position by using the online RSS observations

A. Training phase

In this phase, we aim at building an ensemble model of decision trees using the RSS measurements dataset as training set. Several simulations are carried out at each reference location to calculate different values of RSS. Each simulation is performed with a random orientation vector of each receiver to obtain different values. RSS data are denoted by $\varphi_{i,j}(\tau)$ and indicate the τ-th RSS value measured from i-th coordinator at the j-th receiver. The dataset can be represented by ω as in (3):

$$\omega = \begin{pmatrix} \varphi_{1,1}[\tau] & \cdots & \varphi_{1,R}[\tau] \\ \vdots & \ddots & \vdots \\ \varphi_{A,1}[\tau] & \cdots & \varphi_{A,R}[\tau] \end{pmatrix} \quad (3)$$

A is the number of coordinators, R is the number of receivers or reference locations, $\tau = 1, ..., N$ is the index of RSS samples and N is the number of RSS samples at each reference location.

When principal component analysis is used to reduce the data dimensionality, the dataset can be represented by ω as in (4):

$$\omega = \begin{pmatrix} \varphi_{1,1}[\tau] & \cdots & \varphi_{1,R}[\tau] \\ \vdots & \ddots & \vdots \\ \varphi_{PC,1}[\tau] & \cdots & \varphi_{PC,R}[\tau] \end{pmatrix} \quad (4)$$

$\varphi_{i,j}(\tau)$ is transformed data and indicate the τ-th value transformed from i-th principal component at the j-th receiver, and PC is the number of principal components.

After that, once that dataset of the environment is compiled, an ensemble model of decision trees is built using boosting technique.

B. Test phase

In this phase, a dataset formed by a RSS sample from each coordinator, or its transformation if principal component analysis is used, is taken as input of ensemble model of decision trees to infer the current location. Using similar notations, the online measurements can be represented as in (5):

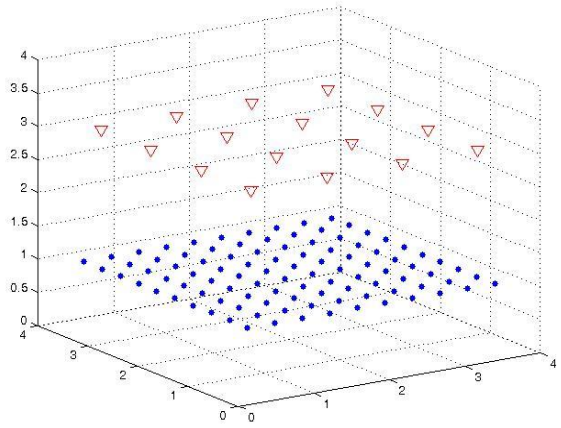

Figure 2. Scenario 1: 16 coordinators and 100 receivers.

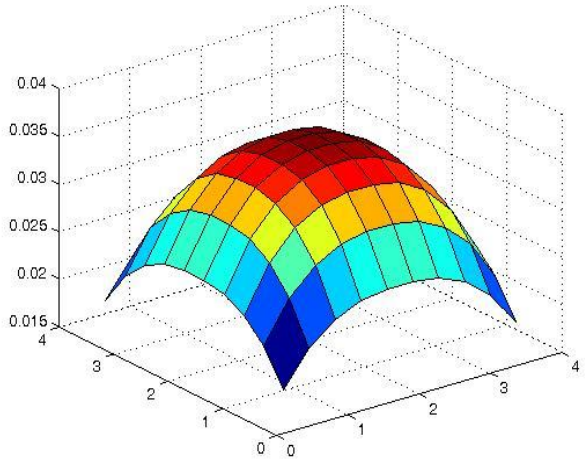

Figure 3. Distribution of the received optical power at 1 meter from the floor.

$$\omega_r = \begin{pmatrix} \varphi_{1,r} \\ \vdots \\ \varphi_{A/PC,r} \end{pmatrix} \quad (5)$$

The location r is unknown.

VI. Experimental Results

In this section, we describe the test environment, and we evaluate the impact of using principal component on the performance of indoor location estimation system. In addition, the accuracy and the computational cost of our system are evaluated.

Experiments were focused to determine the location method accuracy and the computational complexity. The error is the expected distance from the misclassified instance and the real location. The error is calculated by the Euclidean distance between these points, and the arithmetic mean was computed from the results of the experiments. Being a classification problem, an error simply means that a receiver was estimated to be in a wrong positioning cell, in the receiver's grid. All experiments were carried out on an Intel Core i7 3.2 GHz/32 GB RAM non-dedicated Windows machine.

All experiments have been built using the API Weka software [25]. Weka is an open source collection of machine learning algorithms for data mining tasks, more specifically data preprocessing, clustering, classification, regression, visualization and feature selection.

A. Test Environment

Our method was tested in a simulation environment that models a 4 by 4 by 3 meters room. Two scenarios were implemented varying the number of receivers. Scenario 1 is shown in Fig. 2. This environment consists of 16 coordinators or LED lamps (red triangles) configured as 4 x 4 grids placed 1 meter apart from each other on the ceiling. On the lower part, we set up 100 receivers (blue circles) in a 10 x 10 grid configuration, with a 36 centimeters separation from each other. Scenario 2 uses the same number of coordinators, but we set up 361 receivers in a 19 x 19 grid configuration with a 20 centimeters separation from each other. In order to consider different distances between receivers and coordinators, in both scenarios the receivers plane is set up at three different heights: 75, 100 and 125 centimeters from the floor. Receivers orientation was randomly produced for each simulation as follows: they are pointing out to the ceiling with an initial orientation vector [0,0,1] and a random (-0.2,0.2) offset is applied to each axis in each simulation. Thus, each receiver has a different orientation in each simulation.

Eleven simulations were performed on each three vertical layers. One RSS measurement was estimated at each receiver and simulation. This leads to 3.300 RSS and 11.913 RSS measurements for Scenarios 1 and 2, respectively. Fig. 3 shows the received optical power (lux) at 1 meter from the floor with sixteen coordinators. As it can be seen there is enough lighting to receive the beacon frame in every reference location. The simulation parameters are specified in Table 1.

B. Data Transformation using PCA

In order to transform the RSS dataset into principal components *weka.attributeSelection.PrincipalComponents* algorithm was used in conjunction with a Ranker search (implemented in Weka by *weka.attributeSelection.Ranker* class). Dimensionality reduction is accomplished by choosing enough eigenvectors to account for 95% of the variance in the original data. For both scenarios six principal components were obtained. Eigenvectors for Scenario 1 are shown in Table II, where PC1…6 denotes each principal component and L1...16 corresponds to each LED lamp. Similar eigenvectors are obtained for Scenario 2.

C. Analysis of the Training Dataset Size

The size of the training dataset is an important parameter for the performance and the building time of each model

based on decision trees. A large-sized training dataset can provide better accuracy to predict the correct location, but too much data can increase the elapsed time to build the model considerably. In order to test the robustness of the method, different training dataset sizes were used, from 10% to 90% of the whole dataset. For the validity of experimental results, the experiments were performed 100 times, each time selecting the training and testing data after randomizing dataset, picking the same proportion of samples at each class (stratified split). Also, some experiments were carried out using 10-fold cross-validation.

The classification trees were created by the C4.5 algorithm (implemented in Weka by the classifier class: *weka.classifiers.trees.J48*). The boosting method used was the metalearning AdaBoostM1 algorithm implemented by the Weka classifier class *weka.classifiers.meta.AdaBoostM1* with number of iterations equal to 10.

Table III and Table IV for Scenario 1, and Table V and Table VI for Scenario 2 show the ensemble model characteristics when it is built using the original dataset and the transformed dataset (PCA), respectively. As expected, the elapsed time to build each model and the leaves number of the tree increase with the training dataset size. On the other hand, the ensemble model depth is similar for original and transformed datasets. Nevertheless, the time taken to build the ensemble model with transformed dataset is faster than the ensemble model built with original dataset, between about 40% and 50% for Scenario 1 and between about 30% and 40% for Scenario 2, depending on training size.

TABLE II. DATA TRANSFORMATION

PC1	PC2	PC3	PC4	PC5	PC6	
-0.2154	0.288	0.0327	0.39	0.0076	-0.334	L 1
-0.3324	0.1786	-0.1148	0.2057	0.3539	-0.2016	L 2
-0.3691	-0.0749	-0.1094	-0.2066	0.3583	0.1851	L 3
-0.2895	-0.214	0.0393	-0.3825	0.0104	0.3811	L 4
-0.0783	0.3702	-0.1287	0.2083	-0.3464	0.2413	L 5
-0.1772	0.2342	-0.467	0.1221	0.0053	0.1634	L 6
-0.2371	-0.1723	-0.4646	-0.1335	0.0027	-0.1909	L 7
-0.1813	-0.3322	-0.1225	-0.2084	-0.3491	-0.2047	L 8
0.1789	0.333	-0.1251	-0.2069	-0.3504	0.2239	L 9
0.2319	0.1747	-0.4696	-0.1351	0.0089	0.1825	L 10
0.1723	-0.2335	-0.4733	0.1199	0.0102	-0.1725	L 11
0.0768	-0.3706	-0.1281	0.2077	-0.3499	-0.2315	L 12
0.2893	0.216	0.0429	-0.3808	-0.002	-0.3751	L 13
0.3684	0.0779	-0.1132	-0.2044	0.3606	-0.1791	L 14
0.3308	-0.1789	-0.1205	0.2056	0.3591	0.2124	L 15
0.2149	-0.2896	0.0301	0.3865	0.0016	0.3468	L 16

TABLE III. ENSEMBLE MODEL CHARACTERISTICS WITHOUT PCA FOR SCENARIO 1

Training Dataset Size (%)	Time to Build Model (s)	Min Depth	Max Depth	Average Depth	Leaves
10	0.73	6	9	6	1103
20	1.64	5	10	7	1791
30	2.73	5	10	7	2189
40	3.58	6	10	8	2412
50	4.38	5	11	8	2541
60	5.28	5	11	8	2622
70	5.82	5	11	8	2718
80	6.50	5	11	8	2797
90	7.21	5	11	8	2831

TABLE IV. ENSEMBLE MODEL CHARACTERISTICS WITH PCA FOR SCENARIO 1

Training Dataset Size (%)	Time to Build Model (s)	Min Depth	Max Depth	Average Depth	Leaves
10	0.39	6	8	6	1105
20	0.89	6	9	7	1808
30	1.39	6	10	7	2170
40	1.88	5	10	8	2356
50	2.34	5	10	8	2484
60	2.83	5	11	8	2581
70	3.29	5	12	8	2639
80	3.77	5	11	8	2706
90	4.24	5	11	8	2765

TABLE V. ENSEMBLE MODEL CHARACTERISTICS WITHOUT PCA FOR SCENARIO 2

Training Dataset Size (%)	Time to Build Model (s)	Min Depth	Max Depth	Average Depth	Leaves
10	8.89	7	10	8	3909
20	20.07	7	12	9	6555
30	30.56	7	13	9	8439
40	39.81	7	13	10	10016
50	48.21	7	15	10	11288
60	56.17	7	14	10	12305
70	63.64	7	15	10	13330
80	70.66	7	15	10	14310
90	77.58	7	15	10	15027

The leaves number of ensemble model is slightly smaller when principal components are used, and the difference increases when the training dataset size does. Hence, it supposes a considerable reduction of computation cost to build the ensemble model and infer the localization.

On the other hand, Table VII and Table VIII for Scenario 1, and Table IX and Table X for Scenario 2 show the experimental results in terms of correctly classified instance percentage and average error distance using the original dataset and the transformed dataset, respectively. As expected, the accuracy of the system increases when the training dataset size does. Using only 50% dataset size for training the system has an accuracy above 86% and an average error distance less than 9.3 cm for Scenario 1. Nevertheless, an average error distance about 40 cm is reached if misclassified instances are only considered. Obviously, better results are achieved by increasing training dataset size, however, the accuracy is only improved about an 8% using the original dataset and 6% using the transformed dataset from 50% to 90% dataset size, and the average error distance reaches about 2.5 cm.

In all cases, simulations performed with datasets formed by principal components improves the accuracy of system. Although, the average error distance of misclassified instances is higher when these datasets are used. In addition, for the validity of experimental results, experiments were also carried out using 10-fold cross validation yielding a 95.66% accuracy, with an average error distance of 2.4 cm. On the other hand, the system accuracy for Scenario 2 is slightly lower than Scenario 1, yielding an 88.05% accuracy, with an average error distance of 2.5 cm for 10-fold cross validation and using principal components. However, in the second scenario the system achieves an average error distance of misclassified instances about 21 cm. Taking account that the receivers are placed in a grid with a 20 cm separation from each other, most of misclassified instances are the nearest neighbors (receivers) of exact locations.

TABLE VI. ENSEMBLE MODEL CHARACTERISTICS WITH PCA FOR SCENARIO 2

Training Dataset Size (%)	Time to Build Model (s)	Min Depth	Max Depth	Average Depth	Leaves
10	5.31	8	10	8	3963
20	12.36	7	13	9	6593
30	19.62	7	13	9	8300
40	26.57	7	14	10	9407
50	33.52	8	15	10	10250
60	40.51	8	14	10	11117
70	47.12	8	15	10	11847
80	53.90	8	15	10	12461
90	60.89	8	15	10	13195

TABLE VII. EXPERIMENTAL RESULTS WITHOUT PCA FOR SCENARIO 1

Training Dataset Size (%)	Correctly Classified Instances (%)	Average Error Distance ± std (cm)	Average Error Distance ± std (cm) of Misclassified Instances
10	28.30	76.7 ± 0.97	58.2 ± 0.28
20	57.11	42.2 ± 0.98	50.1 ± 0.26
30	72.57	24.0 ± 0.62	46.7 ± 0.24
40	81.34	14.2 ± 0.48	43.4 ± 0.20
50	86.75	9.3 ± 0.38	42.5 ± 0.18
60	89.71	5.8 ± 0.27	40.6 ± 0.16
70	91.94	3.9 ± 0.22	39.4 ± 0.13
80	93.59	3.1 ± 0.19	39.1 ± 0.13
90	94.51	2.7 ± 0.18	37.1 ± 0.03
10-fold Cross Validation	95.18	2.5 ± 0.14	38.2 ± 0.10

TABLE VIII. EXPERIMENTAL RESULTS WITH PCA SCENARIO 1

Training Dataset Size (%)	Correctly Classified Instances (%)	Average Error Distance ± std (cm)	Average Error Distance ± std (cm) of Misclassified Instances
10	37.25	61.3 ± 0.86	56.0 ± 0.27
20	63.28	32.1 ± 0.68	50.0 ± 0.24
30	76.94	18.7 ± 0.53	47.4 ± 0.23
40	84.32	11.4 ± 0.42	47.1 ± 0.23
50	89.00	7.3 ± 0.31	44.4 ± 0.20
60	91.41	5.4 ± 0.26	43.3 ± 0.16
70	93.20	4.4 ± 0.23	46.4 ± 0.23
80	94.14	3.1 ± 0.18	43.0 ± 0.18
90	95.28	2.5 ± 0.16	41.5 ± 0.16
10-fold Cross Validation	95.66	2.4 ± 0.16	43.1 ± 0.19

Fig. 4, Fig. 5, Fig. 6 and Fig. 7 show the cumulative distribution function (CDF) for different training dataset size in Scenario 1 with and without principal components analysis. For Scenario 2, the CDF is show in Fig. 8, Fig. 9, Fig. 10 and Fig. 11. As it can be seen, most of instances are correctly classified and its percentage increases when training dataset size increases. In addition, system accuracy is slightly improved when PCA used. On the other hand, and as it was above commented, most of misclassified locations are the nearest neighbors (in the same receiver's plane) of exact locations, 36 cm and 20 cm for Scenario 1 and Scenario 2, respectively.

TABLE IX. EXPERIMENTAL RESULTS WITHOUT PCA FOR SCENARIO 2

Training Dataset Size (%)	Correctly Classified Instances (%)	Average Error Distance ± std (cm)	Average Error Distance ± std (cm) of Misclassified Instances
10	24.82	25.7 ± 0.23	33.8 ± 0.17
20	46.08	13.8 ± 0.15	25.5 ± 0.11
30	56.84	9.9 ± 0.12	23.1 ± 0.09
40	64.13	7.9 ± 0.11	22.1 ± 0.07
50	69.64	6.5 ± 0.1	21.5 ± 0.06
60	73.82	5.5 ± 0.09	21.3 ± 0.06
70	77.48	4.7 ± 0.09	21.2 ± 0.06
80	80.45	4.1 ± 0.08	21.1 ± 0.06
90	82.87	3.5 ± 0.08	21.0 ± 0.05
10-fold Cross Validation	85.06	3.1 ± 0.07	20.9 ± 0.05

TABLE X. EXPERIMENTAL RESULTS WITH PCA FOR SCENARIO 2

Training Dataset Size (%)	Correctly Classified Instances (%)	Average Error Distance ± std (cm)	Average Error Distance ± std (cm) of Misclassified Instances
10	29.63	23.6 ± 0.22	33.2 ± 0.17
20	52.70	12.0 ± 0.15	25.3 ± 0.11
30	64.47	8.4 ± 0.12	23.6 ± 0.09
40	71.88	6.4 ± 0.11	22.6 ± 0.07
50	76.59	5.1 ± 0.09	22.2 ± 0.06
60	80.15	4.3 ± 0.09	22.1 ± 0.08
70	82.71	3.8 ± 0.08	21.9 ± 0.06
80	84.83	3.3 ± 0.08	21.8 ± 0.06
90	86.56	2.9 ± 0.07	21.7 ± 0.06
10-fold Cross Validation	88.05	2.5 ± 0.07	21.3 ± 0.04

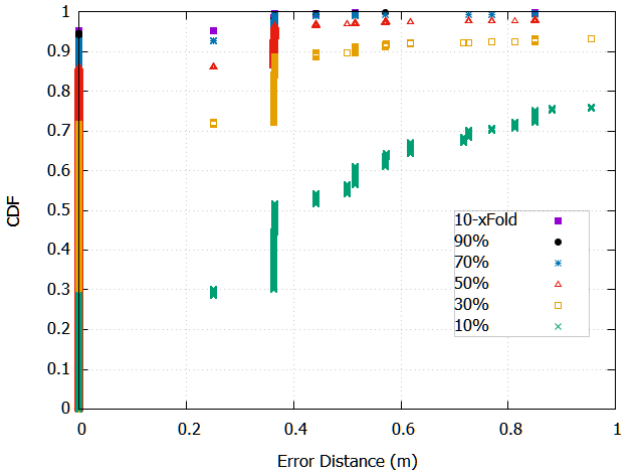

Figure 4. Scenario 1: CDF of performance for different training dataset sizes.

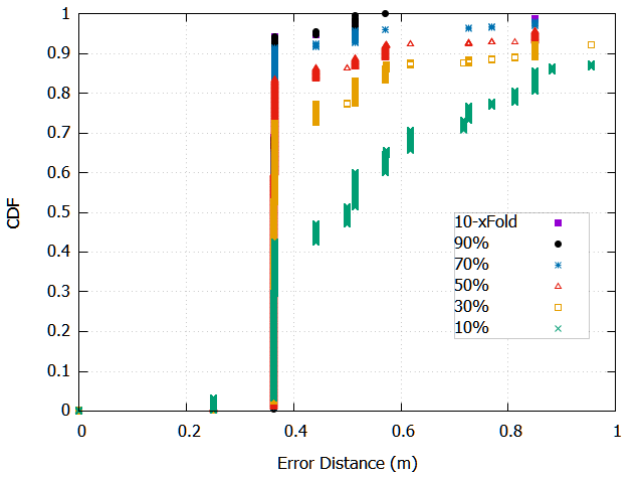

Figure 5. Scenario 1: Misclassified instances CDF of performance for different training dataset sizes.

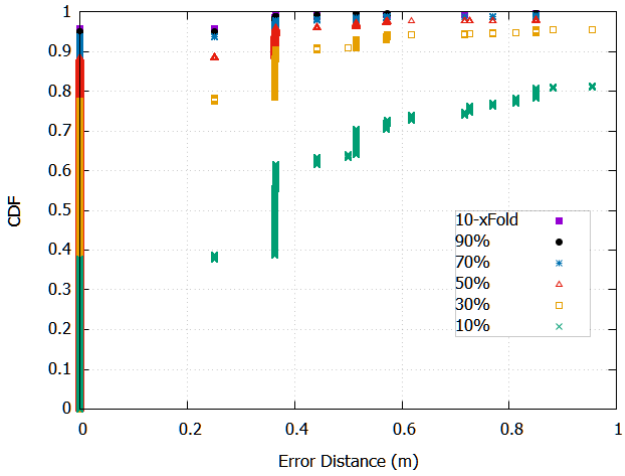

Figure 6. Scenario 1: CDF of performance for different training dataset sizes using PCA.

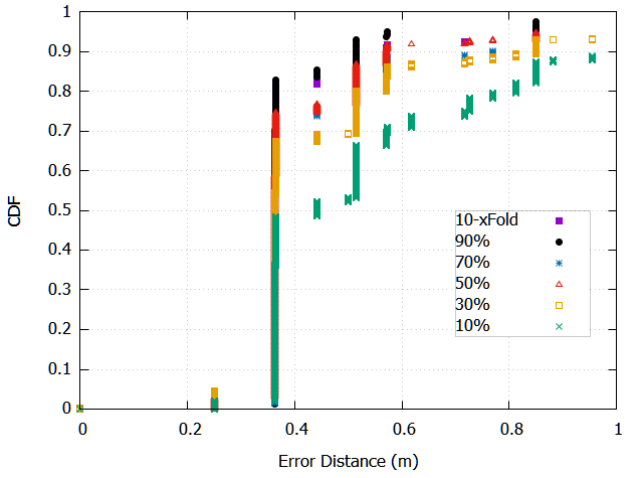

Figure 7. Scenario 1: Misclassified instances CDF of performance for different training dataset sizes using PCA.

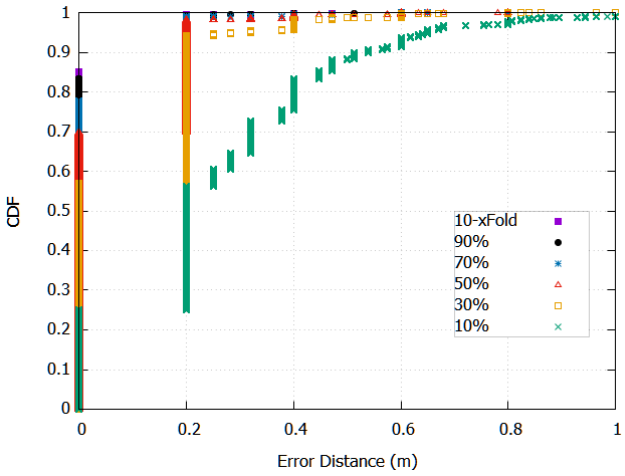

Figure 8. Scenario 2: CDF of performance for different training dataset sizes.

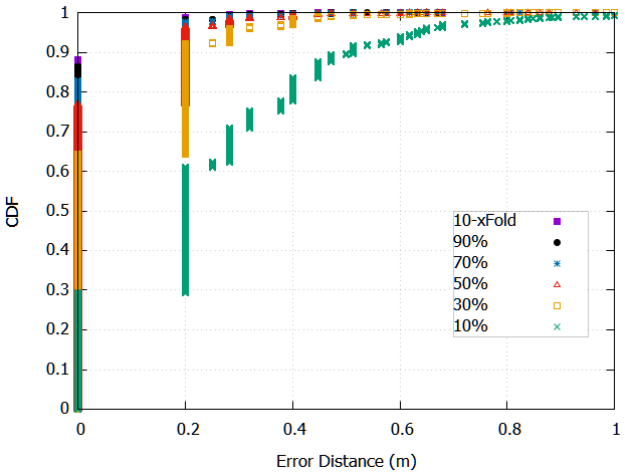

Figure 10. Scenario 2: CDF of performance for different training dataset sizes using PCA.

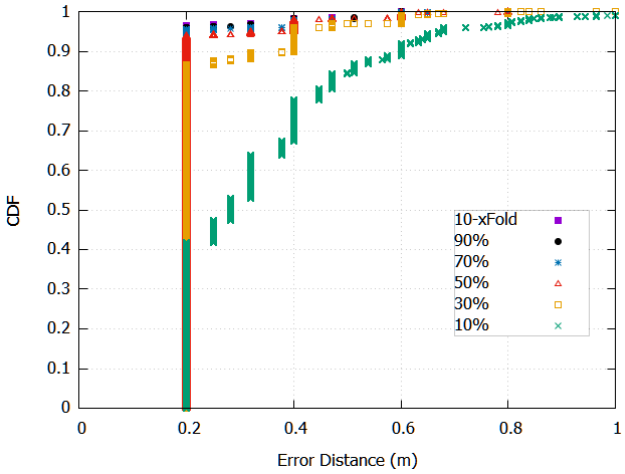

Figure 9. Scenario 2: Misclassified instances CDF of performance for different training dataset sizes.

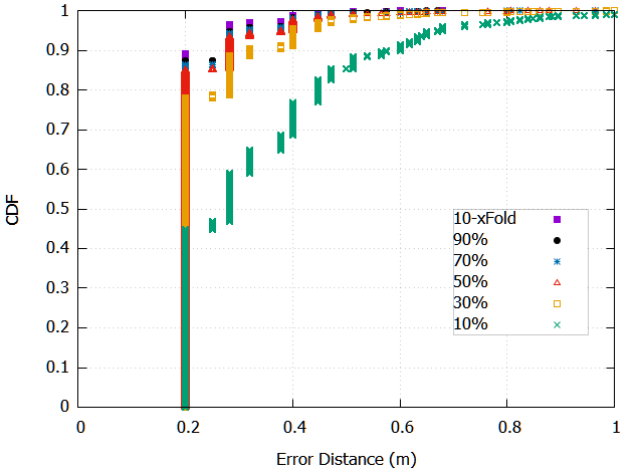

Figure 11. Scenario 2: Misclassified instances CDF of performance for different training dataset sizes using PCA.

VII. CONCLUSION

In this paper, we have demonstrated that decision trees provide a high accuracy for indoor location estimation in VLC networks. This is mainly because the visible light is less susceptible to multipath effects making the propagation and the received optical power more predictable. In addition, principal component analysis provides an efficient mechanism to reduce the data dimensionality, and hence, the system accuracy is improved and the computation time is reduced. Depending on training dataset size the system accuracy can be improved by 10% and the computation time by 50%, as compared to the system when data transformation is not carried out. With regard to accuracy, the best model yielded a 95.6% of instances are correctly classified and average error of 2.4 cm. Furthermore, the ensemble model of decision trees achieves an average error distance of misclassified instances of 43 cm or 21 cm (depending on scenario), taking account that the receivers are placed in a grid with a 36 cm or 20 cm separation from each other, respectively. Thus, most of misclassified instances are the nearest neighbors (receivers) of real locations. On the other hand, the accuracy of the ensemble model improves with the training dataset size, and its effect on the elapsed time to get the model is not meaningful when principal component analysis is used.

Since the average error distance of misclassified instances cannot be less than the distance among receivers when decision trees are used, in our ongoing work, we are planning to use other techniques of data mining, such as regression, to reduce the error distance.

ACKNOWLEDGMENT

This research was partially supported by the Research Program of University of Las Palmas de Gran Canaria (ULPGC2013-15).

REFERENCES

[1] D. Sánchez-Rodríguez, I. Alonso-González, C. Ley-Bosch, J. Sánchez-Medina, M. Quintana-Suárez, and C. Ramírez-Casañas, "Indoor Location Estimation based on IEEE 802.15. 7 Visible Light Communication and Decision Trees," Proceedings of the 12th International Conference on Wireless and Mobile Communications (ICWMC 2016) Barcelona, Spain, pp. 75-79.

[2] R. Want and B. Schilit, "Expanding the Horizons of Location-Aware Computing," IEEE Computer, pp. 31-34, August 2001

[3] P. Bahl and V. N. Padmanabhan, "RADAR: an in-building RF-based user location and tracking system," IEEE Conference on Computer Communications (INFOCOM), pp. 775-784, 2000, doi: 10.1109/INFCOM.2000.832252.

[4] J. Armstrong, Y. A. Sekercioglu, and A. Neild, "Visible light positioning: a roadmap for international standardization," IEEE Communications Magazine, 51(12), pp. 68-73, 2013.

[5] Y. Won, S. H. Yang, D. H. Kim, and S. K. Han, "Three-dimensional optical wireless indoor positioning system using location code map based on power distribution of visible light emitting diode," IET Optoelectronics, 7(3), pp. 77-83, 2013.

[6] W. Xu, J. Wang, H. Shen, H. Zhang, and X. You, "Indoor Positioning for Multiphotodiode Device Using Visible-Light Communications," IEEE Photonics Journal, 8(1), pp. 1-11, 2016.

[7] V. Honkavirta, T. Perala, S. Ali-Loytty, and R. Piché, "A comparative survey of WLAN location fingerprinting methods," Proceedings of the 6th Workshop on Positioning, Navigation and Communication (WPNC 2009) Hannover, Germany, pp. 243–251, 2009.

[8] M.J. Anderson, S. Tsen, L.K. Wang, K. Compton, and M.J. Schulte, "Performance analysis of decimal floating-point libraries and its impact on decimal hardware and software solutions," IEEE International Conference on Computer Design (ICCD 2009), Lake Tahoe, CA, USA, pp. 465–471, 2009.

[9] S.-H. Fang and T. Lin, "Principal component localization in indoor wlan environments," IEEE Transactions on Mobile Computing, vol. 11, no. 1, pp. 100–110, 2012.

[10] I.T. Jollife, Principal Component Analysis. Springer-Verlag, 2002.

[11] S. Rajagopal, R. D. Roberts, and S. K. Lim "IEEE 802.15.7 visible light communication: modulation schemes and dimming support," Communications Magazine, IEEE, 50(3), pp. 72–82, 2012.

[12] OMNeT++ Discrete Event Simulator. Available from: https://omnetpp.org 2017.05.08.

[13] F. Chen, N. Wang, R. German, and F. Dressler, "Performance Evaluation of IEEE 802.15.4 LR-WPAN for Industrial Applications," Fifth Annual Conference on Wireless on Demand Network Systems and Services, pp. 89-96, 2008.

[14] M. Kahn, J. Barry, "Wireless Infrared Communications," Proceedings of the IEEE, Vol. 85, No. 2, pp. 265-298, 1997.

[15] T. Komine and M. Nakagawa, "Fundamental analysis for visible-light communication system using LED lights," IEEE Transactions on Consumer Electronics, Vol.50, Issue 1, pp. 100–107, 2004.

[16] P. Chvojka, S. Zvanovec, P.A. Haigh, and Z. Ghassemlooy, "Channel Characteristics of Visible Light Communications Within Dynamic Indoor Environment," J. Lightwave Technology , 33, pp. 1719–1725, 2015.

[17] D. Deqiang, K. Xizheng, and X. Linpeng, "An Optimal Lights Layout Scheme for Visible-Light Communication System," 8th International Conference on Electronic Measurement and Instruments, pp. 2-189 - 2-194, 2007.

[18] D. Tronghop, J. Hwang, S. Jung, and Y. Shin, "Modeling and analysis of the wireless channel formed by LED angle in visible light communication," International Conference on Information Networking, pp. 354-357, 2012.

[19] M. Youssef and A. Agrawala, "The Horus location determination system," Wireless Networks, 14, pp. 357–374, 2008.

[20] Y. Chen, Q. Yang, J. Yin, and X. Chai, "Power-efficient access-point selection for indoor location estimation," IEEE Trans. Knowl. Data Eng, 18, pp. 877–888, 2006.

[21] O. Z. Maimon and L. Rokach, "Data Mining and Knowledge," Discovery Handbook; Springer: New York, NY, USA, Volume 1 , 2005.

[22] J. R. Quinlan, "C4.5: Programs for Machine Learning," Morgan Kaufmann: San Francisco, CA, USA, Volume 1, 1993.

[23] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," J. Jpn. Soc. Artif. Intell, 14, pp. 771–780, 1999.`

[24] D. Sánchez-Rodríguez, P. Hernández-Morera, J. M. Quinteiro, and I. Alonso-González, "A Low Complexity System Based on Multiple Weighted Decision Trees for Indoor Localization," Sensors, no. 6, pp. 14809-14829, 2015.

[25] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, "The WEKA data mining software: an update," ACM SIGKDD explorations newsletter, 11(1), pp. 10-18.