WiFi Localization For Mobile Robots based on Random Forests and GPLVM

Reda Elbasiony* and Walid Gomaa*,†
*Department of Computer Science and Engineering
Egypt-Japan University of Science and Technology (E-JUST), Alexandria, Egypt
†Faculty of Engineering, Alexandria University, Alexandria, Egypt
Email: {reda.elbasiony, walid.gomaa}@ejust.edu.eg

***Abstract*—The proliferation of WiFi networks has attracted many research communities to employ WiFi signals in estimating the location of mobile devices in indoor environments. In this paper, we propose a localization framework that is capable of determining the location of mobile robots in indoor limited areas. The proposed framework exploits the random forests algorithm in both classification and regression techniques, which are used to build cooperative supervised localization models. The localization models are trained offline based on training data that contains measurements of WiFi signal strengths and the location of these measurements. We also propose an extension of our framework using the gaussian process latent variable model (GPLVM), which gives our framework the ability to build subjective localization models which do not require any prior knowledge about ground truth on the localization place. Our experimental evaluation of the proposed framework using the KheperaIII mobile robot in one testbed show that it gives high accuracy, where the calculated mean localization error is ± 36 cm.**

Keywords*—*WiFi SLAM; Mobile Robots; Random Forests; GPLVM; Khepera

I. Introduction

Recently, using autonomous mobile robots has gained a lot of interest in research communities, especially in the methods of integrating the localization and mapping capabilities of unknown environments into mobile robots. Many solutions have been proposed to implement the localization issue.

There are two main categories of localization strategies: objective localization, and subjective localization [1]. In objective localization, the location of the robot is determined against a global predefined frame of reference, for example using vision-based methods depending on digital cameras and special pre-located landmarks [2], or using GPS signals to determine the absolute position (longitude, latitude and altitude) of A robot on the earth [3]. Although objective localization has the advantages of not being in need of initial position determination, as well as having decreasing error with successive measurements, it also has many limitations. These limitations arise from the available techniques, e.g., using landmarks which can lead to great expense in time and effort. use of such landmarks can lead to sensitivity to environmental lighting conditions in the digital image used. GPS methods also suffer from the problems of being less accurate and being lost in most indoor places. In the use of subjective localization, the position of the robot is determined relative to the robot's initial position, for example using Odometry-based methods, which use the robot's wheel encoders information in determining its linear and angular velocities, and calculating the best location relatively to its initial location [4]. on the contrary of objective localization, the accuracy of subjective localization decreases significantly over time because the noise is integrated over time causing significant growth of error [5].

Currently, the rapid expansion of using the 802.11 WiFi networks attracts many research communities to make use of WiFi signals in many research fields, especially in the field of mobile robot localization. Using WiFi signals to localize objects can help in overcoming many drawbacks of the previous used methods, where, being an important component in most nowadays buildings [6], WiFi access points can replace the landmarks localization based methods without any effort in installation, adjustment, or any additional hardware. Also, on the contrary of landmarks, WiFi signals already cover the whole area of the localization place, and the robot can sense them at any position. The wireless network interface card, which already embedded in almost all mobile robots, can replace the digital camera used in the vision-based localization methods without any impact of lightening conditions. Compared to GPS localization methods, which can't be applied indoors, WiFi localization can be applied indoors with high accuracy.

The key challenge in the method of localization based on WiFi signal strength is the unpredictability of signal propagation, especially through indoor environments [7], which have many walls, doors, and furniture, causing a lot of reflection, absorption, diffraction, and scattering of WiFi signals [8]. So, we will concentrate on proposing techniques to generate accurate localization models based on training data that is collected from indoor environments, this data consists of WiFi signal strength measurements and the ground truth location of the mobile robot.

In this paper, we propose a novel framework which exploits the random forests algorithm [9] to solve the problem of mobile robot WiFi objective localization as classification and regression problems, then, we will show how the proposed framework can be used also in subjective localization by exploiting the Gaussian process latent variable models (GPLVM) algorithm [10] to produce a subjective map of the tested environment to be used in the training process of our proposed framework. The main advantages of our proposed framework are that it can be used in both objective and subjective localization, it doesn't require any parameter optimization in

978-1-4799-7415-3/14 $31.00 © 2014 IEEE
DOI 10.1109/ICMLA.2014.42

the operation phase, which make it suitable for real-time operations, moreover, training data is collected automatically by the robot while tracking a fixed path.

II. Related Work

WiFi localization techniques are mainly grouped into two categories:

A. Signal Propagation Modeling

The goal of research in this area is to build localization models based on understanding the physical properties of WiFi signals while propagation [11], however, this method not only requires knowing a lot of information about the location like walls, doors, and furniture materials to determine their interaction with WiFi signals, but also gives models with limited accuracy [12].

B. Modeling Through Machine Learning

The goal of research in this area is to build localization models directly based on collected training data that contains wireless signal strengths combined with ground truth locations. In the operation phase, a new observed data set that contains WiFi signal strengths is applied to the trained models to generate the corresponding ground truth locations. Many researches have been done in this area, in [11] nearest neighbor searches are used. In [13] an indoor location determination system based on signal strength probability distributions using histograms for signal strengths is proposed. In [14], [15], and [16] WiFi signal strength of each access point is represented as a Gaussian distribution with parameter values derived from training data, however, in the operation phase [16] uses MCL to detect the location based on signal strengths and motion models, but [15] user HMM to detect the location based on signal strengths only. While in [17] a model called LSVM, which depends on support vector machine (SVM), is proposed to solve the localization problem using supervised classification technique. In [7], a localization map is built using WiFi signal strengths based on Gaussian processes (GPs) and GPLVM algorithms, then the localization problem is solved using MCL.

However, our proposed approach solves the localization problem using a classification and regression complex model, where, the random forests algorithm is exploited for building this model based on training data collected automatically by KheperaIII mobile robot. This data consists of WiFi signal strengths of distributed access points, combined with ground truth as x,y coordinates. We proposed also an extension to our approach based on the proposed method in [7], which can be used to generate a subjective localization model in the environments where collecting ground truth data is difficult, as it can accept WiFi signal strengths in the training phase without providing ground truth locations.

The main motivations behind using the random forests Algorithm is that it gives the lowest error rate in some similar applications [6], runs efficiently on large databases,and can handle thousands of input variables without variable deletion [9], so, it can build the localization models very fast, especially in the environments where many access points are used.

III. Background

A. Random Forests

Random forest is a combination of tree predictors where each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest , where, in each tree, If the number of cases in the training set is N, a sample of N cases is taken at random from the original data. This sample will be the training set for growing the tree, then, if there are M input variables, a number m $\ll$ M is specified such that at each node, m variables are selected at random out of the M input variables, then, the best split on these m is used to split the node. The value of m is held constant during the forest growing.

The random forests algorithm can be used in supervised and unsupervised learning techniques. In supervised technique, it can be used in classification and regression. With regard to classification, the random forests algorithm is consisting of a collection of multiple tree classifiers $\{t(x, \Theta_k), k = 1, ...\}$ where the $\{\Theta_k\}$ are independent identically distributed random vectors and each tree gives a unit vote for the most popular class at a determined input x. To classify an observation, it is passed through each decision tree to calculate the suggested class for this tree, giving a set of votes to all possible classes, and finally, the estimated class of this observation will be chosen as the class having maximum number of votes.

With regard to regression, the random forests are built by growing trees depending on a random vector Θ sampled from training data such that each tree predictor $t(x, \Theta_k)$ holds a numerical value not a class label. Assuming that the training set is drawn independently from the distribution of X, Y, where Y is the output value of X, then, the mean square error for each predictor $t(x)$ is:

$$E_{X,Y}(Y - t(X))^2 \tag{1}$$

The random forests predictor is formed by taking the average $\{t(X, \Theta_k)\}$ over k trees. Thus, as the number of the trees in the forest goes to infinity,

$$E_{X,Y}(Y - avrg_k t(X, \Theta_k))^2 \rightarrow E_{X,Y}(Y - E_\Theta t(X, \Theta))^2 \tag{2}$$

Moreover, the random forests algorithm gives very useful information about the provided training data. It gives estimates of what variables are important in the classification or regression processes [9].

B. Gaussian Process Latent Variable Models

1) Gaussian Processes: A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution [18]. Any gaussian process can be specified through two functions: mean function, and covariance function. Let we have a set of training data $T = \{X, Y\}$ drawn from a noisy process $Y = f(X) + \varepsilon$, then, a Gaussian process can non-parametrically estimate posterior distributions over f from the training data T depending on the correlation between the function f values at different points, where, the covariance between $f(a \in X)$, $f(b \in X)$ depends on the covariance between a , b. Thus, a kernel $k(a, b)$ can represent this dependency, for example, Gaussian kernel can be used:

$$k(a, b) = \sigma^2 exp(-\frac{1}{2\beta^2} |a - b|^2) \tag{3}$$

after adding Gaussian observation noise with variance σ_n^2, and performing some mathematical derivation, $f(a_*)$ follows a Gaussian distribution with mean μ_{a_*} and variance $\sigma_{a_*}^2$: $p(f(a_*)|a_*, X, Y) = \mathcal{N}(f(a_*); \mu_{a_*}, \sigma_{a_*}^2)$, where

$$\mu_{a_*} = k_*^T(K + \sigma_n^2 I)^{-1} Y \tag{4}$$

$$\sigma_{a_*}^2 = k(a_*, a_*) - K_*^T(K + \sigma_n^2 I)^{-1} K_* \tag{5}$$

where K is the covariance matrix, and $K_* = k(a_*, X)$ [7].

2) Latent Variable Models: Latent variable model is the probabilistic model which relates a set of latent variables to a set of observed variables using a set of parameters. Then, the values of latent variables can be obtained by marginalizing the parameters and optimizing the latent variables [10]. If we have latent variables X and a set of observed data Y, then $Y = f(X, W) + \epsilon$ is a latent variable model, where W are the parameters of the function $f(.)$, and ϵ is a zero-mean noise.

If independence is assumed across the dimensions of the data, then

$$p(Y|X, W) = \prod_{ij} p(y_{ij}|x_i, w_j) \tag{6}$$

if we select a zero-mean Gaussian prior, and noise drawn from zero-mean Gaussian with variance σ_n^2, then, the marginal likelihood for this case takes the form:

$$p(Y|X) = \prod_{j} \mathcal{N}(y_j; 0, XX^T + \sigma_n^2 I) \tag{7}$$

this model can be extended to use a covariance function defined through a kernel function, thus, maximizing (7) with respect to X determines the values of X that maximize the data likelihood assuming an underlying Gaussian process model [7].

IV. The Proposed WiFi Localization Framework

In this research, we propose a novel framework based on random forests and GPLVM algorithms to solve the localization of mobile robots problem. As shown in Fig. 1 The framework operates in two phases:

A. Training Phase

In the training phase, training data is provided to the framework to build the required models. Training phase consists of the following components:

1) Training Data: Training data is collected using KheperaIII mobile robot. We implemented a program and installed it on the robot to read all the available WiFi signal strengths while tracking a predefined path marked on the ground, where the WiFi signal strengths are collected at well known locations on the path.

2) GPLVM: In case of inability to get the corresponding ground truth of WiFi signal strengths, we can use the proposed framework to operate in subjective localization mode. As in [7], the GPLVM algorithm can be used to reconstruct a topological connectivity graph from a signal strength sequence which can be used to perform subjective localization using our proposed framework.

3) Random Forests Classification: The random forests algorithm is used to build a classification model based on the provided training data, as mentioned before, random forests algorithm builds a set of decision trees based on the training data. There are two important parameters must be selected efficiently in this component: the number of trees to be built, and the number of random features used to construct each decision tree. In our case, we use cross validation to determine the values of these parameters.

4) Random Forests Regression: The random forests algorithm is used to build a set of regression models, where, a regression model is built for each WiFi access point based on its own signal strengths, and the corresponding coordinates as the output and the input of this model respectively. Like the classification component, the values of parameters of each model are selected using cross validation.

5) Classification and Regression Models: These models are generated from the random forests algorithm and are trained with the provided training data. Thus, we have a single classification model that takes a set of WiFi signal strengths as input, and estimates the corresponding location as the output. However, we have multiple regression models, as described before, one model for each access point that, On the contrary of the classification model, takes the xy coordinates as input, and generates the WiFi signal strength as output.

B. Operation Phase

In the operation phase, the models that have been trained in the training phase are used to estimate the location of a KheperaIII [19] mobile robot based on the WiFi signal strengths sensed during motion.

1) Test Data: Test data is a set of WiFi signal strengths which are obtained by the mobile robot at unknown location on its path.

2) WiFi Classifier: WiFi classifier is a random forests classifier which is responsible for predicting the location at which the provided set of WiFi signal strengths are obtained. This component uses the classification model which has been built using random forests algorithm in the training phase.

3) Random Location Generator: Random location generator is a very important component in our proposed framework. It is used to overcome the limitations of representing localization as a classification problem, which are:

1) The role of the classifier is to select the most suitable class (location) to the provided set of WiFi signal strengths, thus, the accuracy of this method is limited by the minimum distance between training points.
2) Our experiments on the classifier show that the error rate of the classifier itself is very high when we accept the output class directly, however, the error decreases by more than 50% when we consider the classification result as a center of a localization circle with a specific radius.

So, the role of random generator is to sample many points from a bivariate Gaussian distribution, where, its mean is the predicted location (xy coordinates) estimated by the classifier, and its variance is determined based on the test results of the

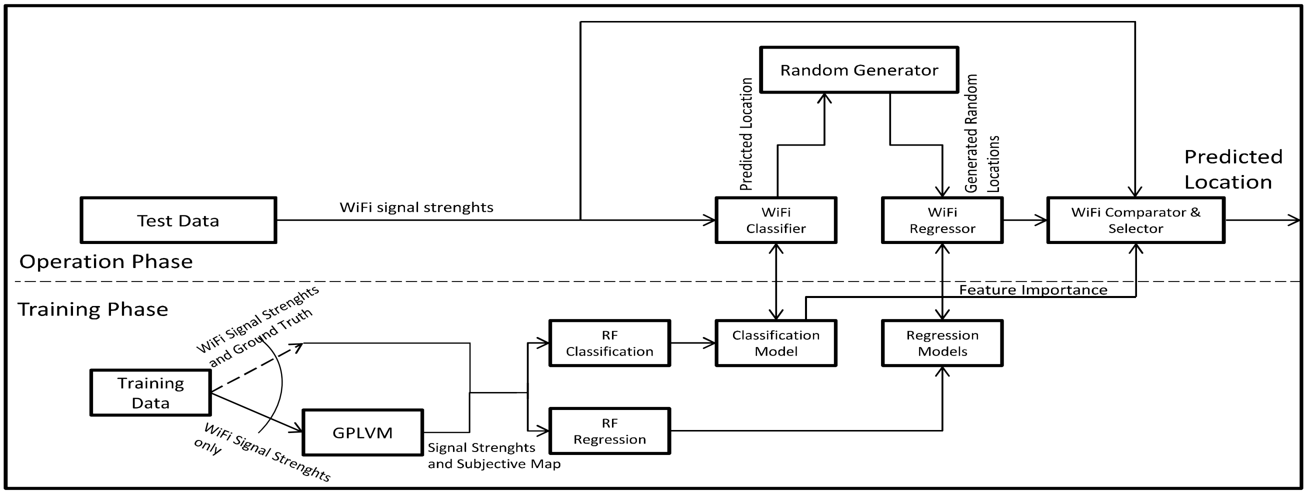

Fig. 1: The proposed WiFi localization framework.

classifier as we will explain in the next section. Thus, instead of accepting the result of the classifier directly, we will expand the search region to cover more area around this point.

4) WiFi Regressor: WiFi regressor is a random forests regressor which is responsible for predicting a WiFi signal strength based on the provided location (xy coordinates), where the regressor uses the previously generated models for each access point to predict a set of WiFi signal strengths for all the provided locations which are generated from the random generator, thus, we have a set of predicted WiFi signal strengths for each generated location.

5) WiFi Comparator and selector: This component receives the set of WiFi signal strengths, which are predicted from the random forests regressor, combined with their randomly generated locations. The role of this component is to calculate the similarity between the record of WiFi signal strengths which are provided to the framework as test data, and all the records of WiFi signal strengths which are predicted from the random forests regressor, and then selects the location having the maximum similarity ratio as the estimated location of the mobile robot. This component also exploits the variable importance estimates given by the random forests algorithm after building the classification model. It uses these importance estimates to enhance the calculated similarity values by applying these estimates, after being normalized, as weights to each feature in the euclidean distances while being calculated making them more expressive.

V. Experiments and Results

In our experiments, we applied the proposed framework to localize KheperaIII mobile robot in a laboratory of 17 meters length and 10 meters width. the lab contains 6 rooms and a wide hall, where all of them contain many pieces of furniture. The path of the robot has been marked on the ground using an adhesive tape, the path is shown in Fig. 2. We also installed seven access points at fixed locations in the lab to depend on their signals in the localization process. We collected our data over six runs, then, we used five runs for training and one for testing.

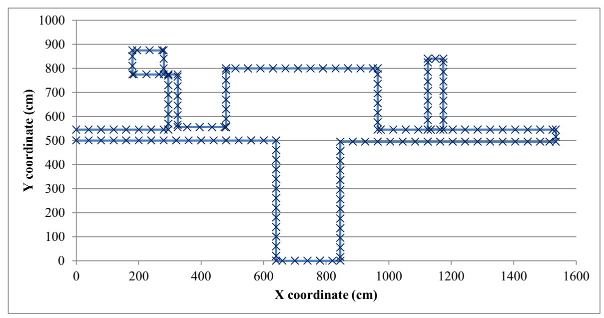

Fig. 2: Motion path and ground truth

In both training and test phases, KheperaIII mobile robot tracked the path automatically and collected a set of WiFi signal strengths readings at each cross point on the path, as shown in Fig. 3

Fig. 3: KheperaIII while collecting data

First, we used the training data to train and select the values of parameters of our proposed framework in objective localization mode, where the WiFi signal strengths and the ground truth of five runs were provided to the Random Forrests classification algorithm to build the classification model. We used cross validation method To validate this model separately, where we measured classification error with variable acceptable euclidean distances. The results are shown in Fig. 4.

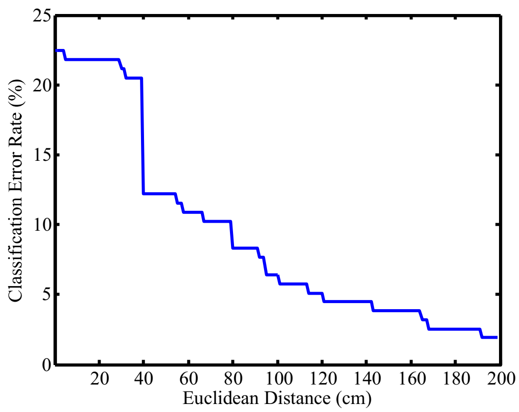

Fig. 4: Classification error rate while changing acceptable euclidean distances

These results show that the classification error reaches about 8% at acceptable distance of 80 cm, so we adjusted the variance of our Gaussian distribution in the random generator to be able to generate points within a circle with radius of 80 cm to cover this area, where the mean of this Gaussian distribution is the point estimated by the classifier.

The random forests algorithm also calculated variable importance estimates of the training data to be used in the similarity measurement as explained before. Fig 5 shows the calculated importance estimates after being normalized.

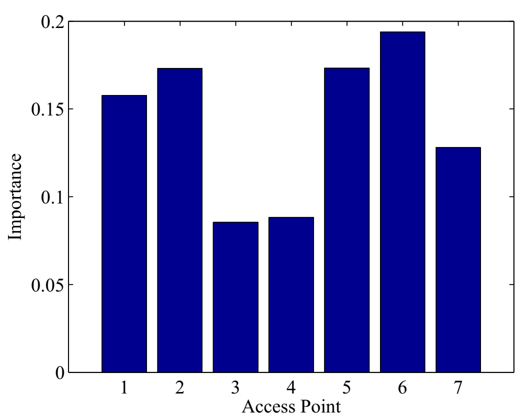

Fig. 5: The normalized variable importance estimates calculated by random forests algorithm

Then, we used the training data also to train seven regression models based on the random forests algorithm, where we built a model for each access point that takes the xy coordinates of a point as input and generates estimated signal strength as output. Fig. 6 and Table. I show the correlation coefficients of each model after being cross validated.

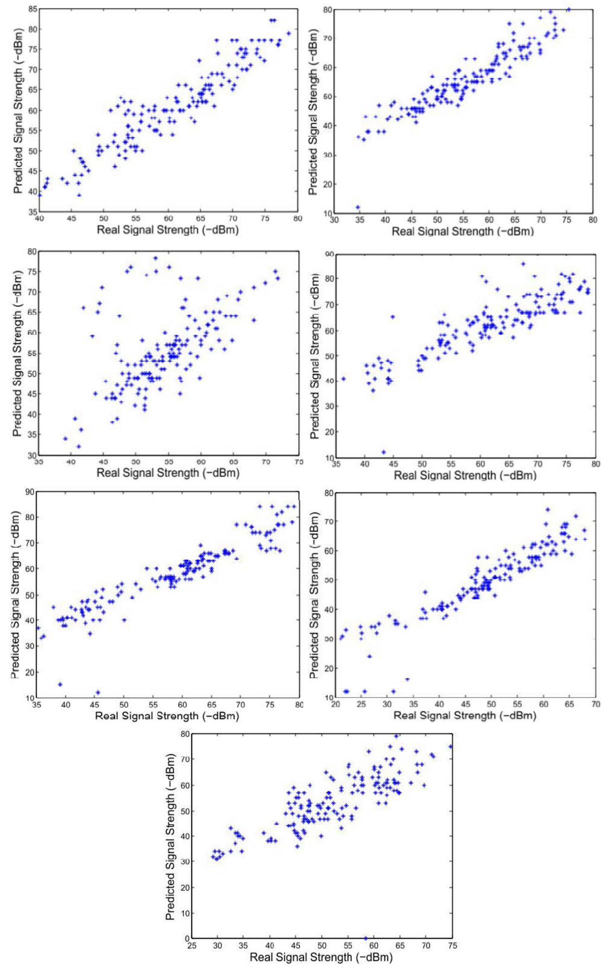

Fig. 6: Correlation coefficients of the generated regression models

TABLE I. CORRELATION COEFFICIENTS OF ALL ACCESS POINTS

Access Point	1	2	3	4	5	6	7	Mean
Correlation Coefficient	0.94	0.93	0.57	0.86	0.93	0.93	0.77	0.85

After being trained, our proposed framework is used to localize KheperaIII while tracking the defined path. The estimated locations compared to ground truth are shown in Fig. 7, where the calculated correlation coefficient for the output was 0.98 with mean localization error about $\pm$ 36 cm over 156 tested locations.

For subjective localization, we also repeated the previous training and test processes on the same training data but without ground truth locations, instead, we used the GPLVM to generate a subjective map that was combined with the provided WiFi signal strengths and provided to the framework as shown in Fig. 1. the output of the test process using this method is shown in Fig. 8, where the calculated correlation coefficient for the output in this case was 0.94. As seen in Fig. 8, the GPLVM algorithm generated a map which was topologically similar to the real path, however it still needs fine tuning which already applied in [7] using a set of provided constraints. In our case,

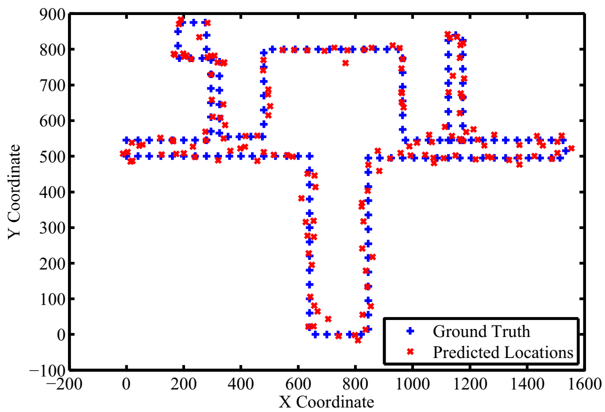

Fig. 7: Ground truth VS. the proposed framework estimated locations

we use the output of GPLVM directly without providing any constraints about the environment.

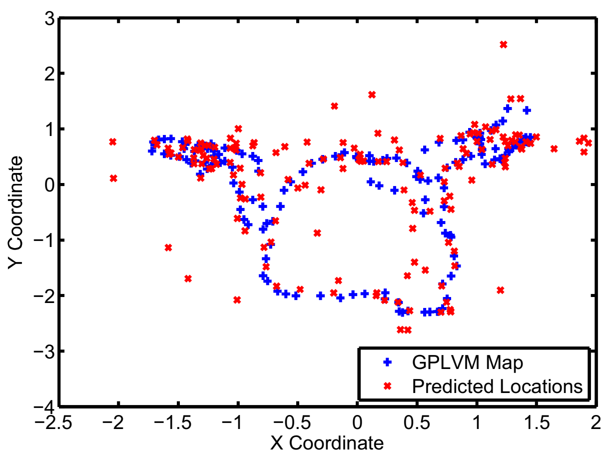

Fig. 8: GPLVM subjective map VS. the proposed framework estimated locations

VI. Conclusion

We have proposed a novel framework to perform the mobile robot localization issue based on classification and regression techniques. The framework depends only on the WiFi signal strengths that the robot can detect while navigation.The main advantages of our framework are the high accuracy (about $\pm$ 36 cm mean localization error), and the ability to be used in real-time localization processes after being trained, moreover, it can be used with or without ground truth for objective or subjective localization respectively.

In the future, we will try to do more experiments on the proposed framework in a wider environment with multiple floors, moreover, we want to employ multiple cooperated robots to collect data in parallel with each other to reduce time for wider places. We want also to study the effect of the number of available access points in more details to determine the minimum number of access points which will be suitable for localization process.

Acknowledgment

This research has been supported by the Ministry of Higher Education (MoHE) of Egypt through a PhD fellowship. Our sincere thanks to Egypt-Japan University of Science and Technology (E-JUST) for guidance and support.

References

[1] M. Bowling, D. Wilkinson, A. Ghodsi, and A. Milstein, "Subjective localization with action respecting embedding," in *Robotics Research*. Springer, 2007, pp. 190–202.

[2] S. Se, D. Lowe, and J. Little, "Vision-based mobile robot localization and mapping using scale-invariant features," in *Robotics and Automation, 2001. Proceedings 2001 ICRA. IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 2051–2058.

[3] G. Reina, A. Vargas, K. Nagatani, and K. Yoshida, "Adaptive kalman filtering for gps-based mobile robot localization," in *Safety, Security and Rescue Robotics, 2007. SSRR 2007. IEEE International Workshop on*. IEEE, 2007, pp. 1–6.

[4] S. Thrun, W. Burgard, and D. Fox, "Probabilistic robotics. 2005," 2005.

[5] J. Borenstein and L. Feng, "Measurement and correction of systematic odometry errors in mobile robots," *Robotics and Automation, IEEE Transactions on*, vol. 12, no. 6, pp. 869–880, 1996.

[6] B. Balaguer, G. Erinc, and S. Carpin, "Combining classification and regression for wifi localization of heterogeneous robot teams in unknown environments," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*. IEEE, 2012, pp. 3496–3503.

[7] B. Ferris, D. Fox, and N. D. Lawrence, "Wifi-slam using gaussian process latent variable models." in *IJCAI*, vol. 7, 2007, pp. 2480–2485.

[8] W. Braun and U. Dersch, "A physical mobile radio channel model," *Vehicular Technology, IEEE Transactions on*, vol. 40, no. 2, pp. 472–482, 1991.

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

[10] N. Lawrence, "Probabilistic non-linear principal component analysis with gaussian process latent variable models," *The Journal of Machine Learning Research*, vol. 6, pp. 1783–1816, 2005.

[11] P. Bahl and V. N. Padmanabhan, "Radar: An in-building rf-based user location and tracking system," in *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 2. Ieee, 2000, pp. 775–784.

[12] S. Zickler and M. Veloso, "Rss-based relative localization and tethering for moving robots in unknown environments," in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*. IEEE, 2010, pp. 5466–5471.

[13] M. A. Youssef, A. Agrawala, A. U. Shankar, and S. H. Noh, "A probabilistic clustering-based indoor location determination system," 2002.

[14] A. Howard, S. Siddiqi, and G. S. Sukhatme, "An experimental study of localization using wireless ethernet," in *Field and Service Robotics*. Springer, 2006, pp. 145–153.

[15] A. M. Ladd, K. E. Bekris, A. P. Rudys, D. S. Wallach, and L. E. Kavraki, "On the feasibility of using wireless ethernet for indoor localization," *IEEE Transactions on Robotics and Automation*, vol. 20, no. 3, pp. 555–559, 2004.

[16] J. Biswas and M. Veloso, "Wifi localization and navigation for autonomous indoor mobile robots," in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4379–4384.

[17] D. A. Tran and T. Nguyen, "Localization in wireless sensor networks based on support vector machines," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 19, no. 7, pp. 981–994, 2008.

[18] C. E. Rasmussen and C. Williams, "Gaussian processes for machine learning. 2006," *The MIT Press, Cambridge, MA, USA*, vol. 38, pp. 715–719, 2006.

[19] Khepera iii. [Online]. Available: http://www.k-team.com/mobile-robotics-products/khepera-iii