Identification and Mitigation of Non-line-of-sight conditions Using Received Signal Strength

Zhuoling Xiao*, Hongkai Wen*, Andrew Markham*, Niki Trigoni*, Phil Blunsom*, and and Jeff Frolik†
*Department of Computer Science, University of Oxford. Email: firstname.lastname@cs.ox.ac.uk.
†School of Engineering, University of Vermont. Email: jfrolik@uvm.edu

***Abstract*—Various applications, such as localisation of persons and objects could benefit greatly from non-line-of-sight (NLOS) identification and mitigation techniques. However, such techniques have been primarily investigated for ultra-wide band (UWB) signals, leaving the area of WiFi signals untouched. In this study, we propose two accurate approaches using only received signal strength (RSS) measurements from WiFi signals to identify NLOS conditions and mitigate the effects. We first explore several features from the RSS which are later demonstrated as very effective in identifying and mitigating NLOS conditions. After that, we develop and compare two major optimization problems based on a machine learning technique and hypothesis testing according to different user requirements and information available. Extensive experiments in various indoor environments have shown that our techniques can not only accurately distinguish between LOS/NLOS conditions, but also mitigate the impact of NLOS conditions as well.**

***Keywords*—NLOS identification and mitigation, machine learning, hypothesis testing, localisation.**

I. Introduction

Line-of-sight/non-line-of-sight (LOS/NLOS) information can greatly benefit many location-related applications. Typical examples include the localisation of people and objects inside buildings or in urban landscapes (like victim location detection in emergencies, equipment tracking in hospitals, and other location-based commerce). The accuracy of indoor localisation techniques, especially RSS-based, is hampered by multi-path effects especially in NLOS conditions when the received signal is only composed of these reflected signals. Therefore, it is necessary that NLOS identification and mitigation techniques are introduced to improve the accuracy of RSS-based localisation.

NLOS identification and mitigation techniques so far have been primarily investigated for ultra-wide band (UWB) signals [1]–[5]. The ultra-wide bandwidth of UWB signals makes it possible to identify and extract the LOS component from the received signal, which makes the NLOS identification and mitigation convenient. The identification techniques mainly include hypothesis testing [1], [3] or machine learning algorithms [5] based on features from the received UWB signals, such as root-mean-square (RMS) delay spread, mean/excess delay, and amplitude. The detailed comparison between different variables and approaches can be found in [4]. However, it is impossible to extract such detailed set of features from WiFi signals due to its narrow bandwidth.

Another generic NLOS mitigation technique [6] tries to recover the NLOS errors with convex programming. But it requires more LOS samples than NLOS samples for localisation, which is not the case in scenarios where we often have only one LOS access point in view at majority of the time, e.g., corridors in buildings.

NLOS mitigation is also achieved in [7] and [2], [8] by means of minimum residual and residual weighting algorithms. [7] proposed the minimum residual and residual weighting algorithms to mitigate the NLOS conditions by selecting a subset from the available access point set that minimizes the distance estimation residual. [2], [8] improves the algorithm to reduce the computation complexity. However, the accuracy of the aforementioned NLOS mitigation techniques is not satisfactory in an environment with few LOS measurements and many NLOS measurements.

In this paper, we propose and compare two NLOS identification and mitigation techniques with only RSS measurements from WiFi signals. Based on the observations of multiple RSS measurements from experiments, we explore several variables and adapt a least square support vector machine (LS-SVM) and a Neyman-Pearson testing to identify LOS/NLOS conditions and mitigate their impact on various applications. The main contributions of this paper are as follows:

- We propose the concept and implementation of NLOS identification and mitigation with only RSS measurements, which greatly improves the potential of RSS-based localisation and secure data transmission.
- We explore several novel features from the collected RSS measurements, which are shown to be effective in LOS/NLOS discrimination.
- We formulate two optimization problems to identify the NLOS conditions, which could output accurate NLOS identification results in different environments.
- We design and run extensive experiments to test the accuracy of the proposed techniques.

The remainder of the paper is organized as follows. Section II presents the problem formulation and system model. Section III proposes the feature selection and extraction scheme. Section IV develops the machine learning algorithm used to perform the identification and mitigation. Section V describes the hypothesis testing strategy used in this study. Section VI introduces the experiments and the performance of the NLOS identification and mitigation algorithms. Section VII describes the impacts of our algorithm on positioning system and compares the performance with related approaches. Section VIII concludes the whole paper.

II. Problem Formulation and System Model

This section formulates the problems to be solved in this study and presents the system model for the NLOS identi-

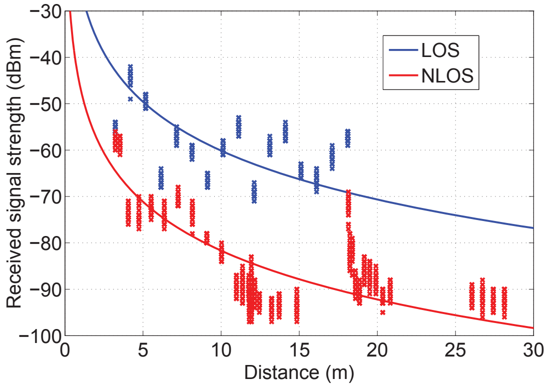


Fig. 1. RSS measurements in LOS and NLOS conditions and the corresponding propagation models estimated using least squares

fication and mitigation algorithms. This study addresses two problems: NLOS identification based on RSS measurements such as Fig. 1, and NLOS mitigation primarily for RSS-based localisation services.

Whether some obstacles block the LOS path of the wireless signals makes a significant difference on the determination of the location, which is the major reason for the inaccuracy of localisation approaches based on the indoor propagation models. In general, for a given distance the RSS in LOS conditions can be over a hundred times stronger than the RSS in NLOS conditions (as shown in Fig. 1). The general form of indoor propagation models can be presented as

$$d = f(\mathbf{r}, \varepsilon, \mathbf{e}), \tag{1}$$

where $\mathbf{r}$ is the signal path loss, ε is the path loss factor, and $\mathbf{e}$ is the environment factor dependent on the walls, floors, windows, etc. For instance, the standard log-normal indoor propagation model is

$$P(d)[dBm] = P(d_0) + 10\gamma\log\frac{d}{d_0} + WAF + X_\sigma, \tag{2}$$

in which $P(d)$ is the received signal strength in a location d meters away from the anchor, d_0 is the reference distance, γ is the distance power loss coefficient, WAF is the wall attenuation factor, and X_σ is a Gaussian distributed random variable with variance σ^2.

The machine learning approach first performs an extensive indoor measurement campaign to collect training data. From the collected RSS measurements we extract some features that could help distinguish between the LOS and NLOS signals or give distance predictions. Afterwards, we select different sets of features to train a least square support vector machine (LS-SVM) to identify LOS/NLOS conditions before we put the algorithm into practice. Note that the parameters we obtained from the measurement campaign could also be used to make accurate predictions in other buildings, which we are going to discuss in Section VI.

The hypothesis testing approach works in a different way. Suppose we can determine *a priori*, denoted with α, from the collected RSS measurements to distinguish between LOS and NLOS conditions, then the two competing hypotheses are

$$\begin{aligned} H_l &: \alpha \le \alpha_t, && \text{LOS conditions,} \\ H_n &: \alpha > \alpha_t, && \text{NLOS conditions.} \end{aligned} \tag{3}$$

A proper function α and threshold α_t which we will develop in Section V could identify the NLOS conditions.

In addition to the NLOS identification problem, our study also addresses the problem of mitigating the influence of NLOS conditions and accurately estimating the distances between transmitters and receivers. With features from the RSS measurements, we build a regression model in machine learning approach to directly predict these distances. However, in hypothesis testing approach we cannot explicitly predict the transmitter-receiver distance but instead we develop different models for LOS and NLOS conditions which could give more accurate distance estimations.

III. NLOS Feature Extraction

In this section, through observation of the RSS samples from LOS and NLOS conditions, we extract typical features from the collected RSS measurements to identify NLOS conditions, including the mean, the standard deviation, the Kurtosis, the Rician K factor, and the χ^2 goodness of fit test parameters. Both the machine learning approach and hypothesis testing approach are developed based on these features. The distributions of these variables are derived for hypothesis testing either from theories or empirical data observation.

A. The Mean and the Standard Deviation (μ, σ_s)

The mean μ and standard deviation σ_s alone cannot distinguish between LOS/NLOS conditions. However, together with features discussed below, the mean and standard deviation can help in NLOS identification. We assume, as indicated by our data, that Gaussian conditions hold for N samples with mean μ_l and variance σ_l^2/N in LOS conditions and with mean μ_n and variance σ_n^2/N in NLOS conditions. Then the probability that a set of RSS samples are from LOS/NLOS conditions can be determined as

$$p(M = m|H) = \frac{\sqrt{N}}{\sigma_h\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{m-\mu_h}{\sigma_h/\sqrt{N}}\right)\right] \tag{4}$$

in which μ_h and σ_h are μ_l and σ_l in LOS conditions, and μ_n and σ_n in NLOS conditions.

B. Kurtosis ($\mathcal{K}$)

The Kurtosis measures the peakedness of the probability distribution, which is defined as $\mathcal{K} = \mu_4/\sigma_s^4 - 3$ in which σ_s is the standard deviation of the sample and μ_k is the kth moment about the mean.

In LOS conditions, the received signal contains a major component which is much stronger than the rest. As a result, the RSS in LOS conditions remains comparatively stable even when other components suffer from changing fading effects. However, the received signal from NLOS conditions is composed of signal components which are all highly variable. Therefore, the RSS measurements in LOS conditions are more centralized than the samples in NLOS conditions. Generally speaking, the RSS distribution in LOS conditions has a higher

Kurtosis than the RSS distribution in NLOS conditions. Existing empirical research [2] has proved that the Kurtosis of UWB signals can be well modeled by a log-normal distribution. Our data also indicates that the log-normal model is a suitable choice for the Kurtosis data.

$$p(\mathcal{K}=\kappa|H)=\frac{1}{\kappa\sqrt{2\pi\sigma_\kappa}}\exp\left[-\frac{(\ln(\kappa)-\mu_\kappa)^2}{2\sigma_\kappa^2}\right] \tag{5}$$

where μ_κ and σ_κ are the mean and standard deviation of $\ln(\kappa)$.

C. Skewness ($\mathcal{S}$)

The skewness measures the asymmetry of the probability distribution. The skewness of Rayleigh distribution is a constant (aprox. 0.63) which is generally larger than the skewness of Rician distribution. In other words, the LOS measurements should be more symmetrical than the NLOS samples.

The skewness is defined as the third standardized moment $\mathcal{S}=\mu_3/\sigma_s^3$ where σ_s is the standard deviation of the sample and μ_3 is the third moment about the mean.

D. The Rician K factor (K_r)

LOS environments have a major dominant signal, which makes the RSS measurements follow the Rician distribution whereas NLOS environments tend to follow a Rayleigh distribution. Existing theoretical and empirical studies have shown that there is a link between the Rician K factor and the presence of LOS conditions [9]. The Rician K factor is defined as the ratio between the power in the direct path and the power in other scattered paths $K=\nu^2/(2\sigma^2)$. Therefore, in NLOS conditions where no direct path exists, the Rician K factor should be zero.

To estimate the Rician distribution and the Rician K factors from the RSS measurements, we use the fixed point technique [10] which converges quickly. We define the ratio of the mean and standard deviation of RSS measurements at one location as $\gamma=\frac{\mu}{\sigma_s}$ where σ_s is the standard deviation of samples and the ratio of Rician distribution parameters as $\theta=\frac{\nu}{\sigma}$. The fix point θ is given as follows.

$$\theta=\sqrt{\xi(\theta)\left(1+\gamma^2\right)-2}, \tag{6}$$

where the correction factor $\xi(\theta)$ is defined as

$$\xi(\theta)=2+\theta^2-\frac{\pi}{2}\left[{}_1F_1\left(-\frac{1}{2},1,-\frac{\theta^2}{2}\right)\right]^2, \tag{7}$$

in which ${}_1F_1$ is the confluent hypergeometric function.

Then we can derive the fix point θ by iterations of Equations (6) and (7), where Newton's method could speed the convergence. Since the convergence speed of Newton's method depends largely on the initial value, we can take $\theta_0=(\gamma-\mathcal{L}_\theta)$ as the start point of iterations, where $\mathcal{L}_\theta$ is the lower bound of θ determined by

$$\mathcal{L}_\theta=\sqrt{\frac{2}{\xi(0)}-1}. \tag{8}$$

Since the fixed point formula has a unique solution for every γ, Newton's method could converge at $\theta=\theta^*$. Then the Rician distribution parameters σ and ν can be determined as follows.

$$\begin{aligned}\sigma &= \frac{\sigma_s}{\sqrt{\xi(\theta^*)}}, & (9)\\ \nu &= \sqrt{\mu^2+(\xi(\theta^*)-2)\sigma^2}. & (10)\end{aligned}$$

Then the Rician K factor can also be determined. The data indicates that the distribution of Rician K factors could be approximated as normal distributions. Then the probability that a set of RSS measurements is taken from LOS/NLOS conditions is

$$p(K=k|H)=\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_k}{\sigma_k}\right)\right] \tag{11}$$

E. The χ^2 goodness of fit (χ^2)

Compared with other scattered signals, the LOS signal reacts minimally, which leads to the different shapes in the empirical distributions in LOS and NLOS conditions. As a result, the goodness of fit parameters to their underlying distributions are different in LOS and NLOS conditions. The disadvantage of this variable is that its performance largely depends on the number of samples. A larger number of measurements results in better performance.

The χ^2 goodness of fit test parameter is defined as

$$\chi^2=\sum_{i=1}^{N}\frac{(O_i-E_i)^2}{E_i}, \tag{12}$$

where O_i and E_i are the observed and expected frequency of the ith sample, respectively.

From the definition, the χ^2 parameter indicates the distance between the RSS measurements and the underlying distribution. A large χ^2 value implies a poor fit between the observed and expected distribution. We also assume that the χ^2 follows Gaussian distribution.

IV. Machine Learning Approach

Since our algorithm is designed for the potential use in mobile devices, the quality of generalization and ease of training possess the highest priority in the selection of machine learning algorithms. Therefore, we use the Support Vector Machine (SVM) whose capabilities in these two aspects are far beyond those of other machine learning approaches. SVM is a supervised machine learning algorithm which can be used as a classifier to separate data sets with different features or as a regressor to estimate the unknown dependent variable (like distances from transmitters to receivers) from some independent variables.

A. Classification

Given a set of training data $\{\mathbf{x}_k, y_k\}_{k=1}^N$ where $\mathbf{x}_k\in\mathbb{R}^n$ and $y_k\in\{-1,1\}$ are the input variables and labels, respectively, linear machine learning algorithms are designed to separate the data set in the following form.

$$y(x)=\text{sign}\left[\mathbf{w}^T\psi(\mathbf{x})+w_0\right], \tag{13}$$

in which $\psi(\cdot)$ is the predetermined feature mapping function, **sign** is the signum function which extracts the sign of a real number, and $\mathbf{w}$ and w_0 are parameters learned from the training data. According to our initial experiments, the LOS/NLOS RSS measurements are not linearly separable, therefore we use a Gaussian radial basis function (RBF) to get a better result than a linear feature mapping.

To make this algorithm feasible in practical implementations, it is necessary to reduce the computation complexity. Therefore, to avoid the quadratic programming problem of standard SVM, the LS-SVM [11] is used in this study which simplifies the optimization problem as follows.

$$\begin{aligned} \underset{\mathbf{w},w_0,\mathbf{e}}{\text{argmin}} \quad & \tfrac{||\mathbf{w}||^2}{2} + c\tfrac{1}{2}\textstyle\sum_{k=1}^{N} e_k^2 \\ \text{s.t.} \quad & y_k\left[\mathbf{w}^T\psi(x_k) + w_0\right] = 1 - e_k, \quad \forall k, \end{aligned} \tag{14}$$

where c is the weighted factor and e_k is the penalty of misclassification. This optimization problem can be solved with its Lagrangian dual and Karush-Kuhn-Tucker (KKT) conditions. It can be shown that the optimization problem (14) is a linear programming problem [11].

In the NLOS identification problem, the input of the classifier are different sets of features discussed in Section III and the output is the classification result ($b = 1$ in LOS conditions and $b = -1$ in NLOS conditions). We are interested in which subset of the features gives the best identification result and how many RSS samples should be combined to make a decision with acceptable accuracy.

B. Regression

The NLOS mitigation is achieved in machine learning approach with regression technique. Again, the input of the regressor in the NLOS mitigation problem are the features extracted from the RSS samples and also the distances derived from the propagation model in (2). The output of the regressor is the estimated distance between the anchor and the location where the samples are collected. We are also interested in which subset of all features presents the least errors.

The SVM regressor is very similar to the classifier on the optimization problem. The regressor is just a function from $\mathbb{R}^n$ to $\mathbb{R}$, which is in the same form as the classifier without the sign function.

$$y(x) = \mathbf{w}^T\psi(\mathbf{x}) + w_0, \tag{15}$$

The distances between the support vectors and the separating hyperplane are maximized by objective function (16).

$$\begin{aligned} \underset{\mathbf{w},w_0,\mathbf{e}}{\text{argmin}} \quad & \tfrac{||\mathbf{w}||^2}{2} + c\tfrac{1}{2}\textstyle\sum_{k=1}^{N} e_k^2 \\ \text{s.t.} \quad & y_k = \mathbf{w}^T\psi(x_k) + w_0 + e_k, \quad \forall k. \end{aligned} \tag{16}$$

Similar to the classification optimization problem (14), the regression optimization problem (16) can also be solved by standard optimization tools.

V. Hypothesis Testing Approach

The reason why we develop an identification approach based on the hypothesis testing is that the hypothesis testing does not require any training phase at the cost of degraded performance and flexibility compared with the machine learning approach.

A. NLOS Identification

We use the Neyman-Pearson test to determine the NLOS conditions with aforementioned distributions. Denote the four variables (i.e., M, $\mathcal{K}$, K, and χ^2) in the last section with $V_i, i = 1, 2, 3, 4$, if we choose N_v $(1 \leq N_v \leq 4)$ variables to make the decision, the joint distribution could be denoted with $p(V_1, \cdots, V_{N_v}|H)$. We can define *a priori* α as

$$\alpha = \frac{p(V_1, \cdots, V_{N_v}|H_l)}{p(V_1, \cdots, V_{N_v}|H_n)}, \tag{17}$$

Then the competing hypotheses are the same as Equation (3) with the threshold $\alpha_t = 1$.

Since the joint distribution requires the convolution of the PDFs of the aforementioned four random variables, the computation complexity could be extraordinarily high. Therefore, in order to put the algorithm into practice in future, a suboptimal solution is to assume all four distributions are independent. Then α could be simplified to

$$\alpha = \prod_{i=1}^{N_v} \frac{p(V_i|H_l)}{p(V_i|H_n)}. \tag{18}$$

in which $p(V_i|H_l$ and $p(V_i|H_n)$ are the distributions of feature V_i in LOS and NLOS conditions, respectively.

B. NLOS mitigation

The NLOS mitigation approach is relatively simple with hypothesis testing. For LOS and NLOS conditions the parameters in Equation 2 (especially γ and σ) can vary greatly thus producing significant errors in distance estimation should distinct models not be used. Therefore, rather than a single propagation model for both conditions as 2, we add the LOS/NLOS conditions as a separate input to the indoor propagation model which then becomes

$$d = f(\mathbf{r}, \varepsilon, \mathbf{e}, y), \tag{19}$$

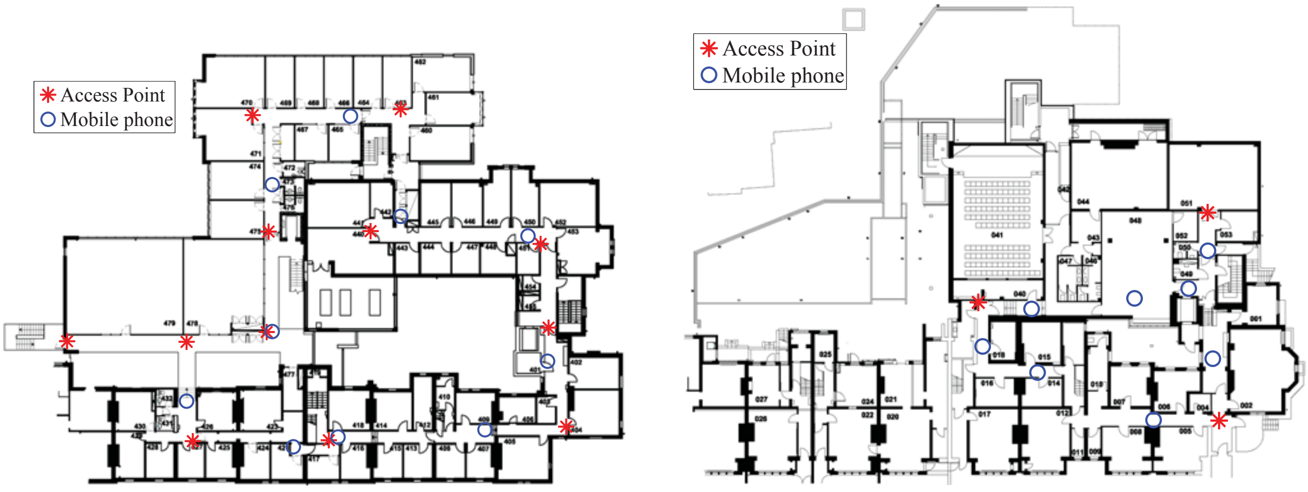
where y is the indication of the LOS/NLOS conditions.

VI. Evaluation

This section evaluates the proposed algorithms through extensive data collection at different times and places. To make the evaluation reliable, RSS samples are collected during different periods of the day in two different experimental sites.

A. Experimental Settings

Collecting RSS measurements to allow for the identification of LOS/NLOS conditions is the primary objective of the experiments. To achieve this goal with machine learning algorithms, we first collect a large number of RSS measurements in an indoor environment to build a database of training samples for the LS-SVM. The antenna of each phone is kept in the same orientation in the data collection during both the training and testing phases to avoid variations caused by antenna orientation. Since the numbers of LOS and NLOS samples differ in various scenarios, we collect half of the samples in LOS conditions and the other half in NLOS conditions. To make this approach more practical for future implementation in localisation, we build the experiments on Huawei U8160 mobile phones running Android 2.3.3. Measurements from



(a) $65m \times 45m$

(b) $55m \times 40m$

Fig. 2. The two experimental sites and locations of anchors.

over 10 agents are fused in the experiments to account for hardware variations between these mobile phones.

As shown in Fig. 2, the anchors are marked in the experimental sites and the agents moved along the corridors in the experiments. To collect the RSS samples in different situations, the distance between the anchor and the agent varies significantly from roughly 0.8 m to 20 m.

B. Database

It is necessary to take into consideration different indoor environments in the implementation of the algorithm. As we know, the accuracy of NLOS identification techniques can be easily decreased by external interference such as people walking around and other signal noise. Although people walking around may not block the LOS signal, they can block and absorb other components of the received WiFi signal which leads to the variation of the measurement distribution. Moreover, from the long-term perspective of practical use, it is impossible to avoid interference from people because it is them who hold their mobile phones and use these services.

To consider the interference from people separately we have two categories of RSS samples in the database. The first group of samples is collected during nights and weekends when there are few people walking around the anchors and the agents to absorb and block the WiFi signal (called static environment hereafter). The other group of samples are collected during busy office hours when there are many people working in their offices and walking around the corridors, which causes severe interference to the RSS measurements and the final measurement distributions (called dynamic environment hereafter). Each of the two groups contains approximately 1500 sets, each of which is composed of 1000 RSS samples (3,360,000 RSS samples in total). We divide each sample set into subsets according to the sample size discussed in the next subsection and extract features from each subset. As stated, half of the sample sets in each group are taken from LOS conditions and the other half from NLOS conditions. The two groups of data are analyzed and discussed in the following subsection.

TABLE I. MISSED DETECTION PROBABILITY (p_m), FALSE ALARM PROBABILITY(p_f), AND OVERALL MISCLASSIFICATION PROBABILITY (p_e) OF MACHINE LEARNING ALGORITHM. FEATURES ARE EXTRACTED FROM EVERY 1000 RSS SAMPLES IN STATIC ENVIRONMENT.

Identification features	p_m	p_f	p_e
$\{\mu\}$	0.683	0.0358	0.1041
$\{\mu, K_r\}$	0.0304	0.0284	0.0588
$\{\mu, K_r, \chi^2\}$	0.0324	0.0324	0.0648
$\{\mu, \sigma_s, K_r, \chi^2\}$	0.0324	0.0331	0.0655
$\{\mu, \sigma_s, D_n, D_p, \mathcal{P}_m\}$	0.0412	0.0331	0.0743

TABLE II. MISSED DETECTION PROBABILITY (p_m), FALSE ALARM PROBABILITY(p_f), AND OVERALL MISCLASSIFICATION PROBABILITY (p_e) OF MACHINE LEARNING ALGORITHM. FEATURES ARE EXTRACTED FROM EVERY 1000 RSS SAMPLES IN DYNAMIC ENVIRONMENT.

Identification features	p_m	p_f	p_e
$\{\mu\}$	0.1620	0.0947	0.2567
$\{\mu, \mathcal{S}\}$	0.1316	0.0647	0.1963
$\{\mu, \sigma_s, \mathcal{S}\}$	0.0853	0.0572	0.1425
$\{\mu, K_r, \mathcal{K}, \mathcal{S}\}$	0.0883	0.0529	0.1412
$\{\mu, K_r, \chi^2, \mathcal{K}, \mathcal{S}\}$	0.0893	0.0508	0.1401

C. NLOS Identification

After the database has been built, we train the LS-SVM with the training data sets and then evaluate the performance of the identification algorithms using over 1500 extra test items. Tables I and II show the performance of the machine learning based NLOS identification algorithm in static and dynamic environments. The performance of the algorithm is measured in terms of missed detection probability p_m (deciding LOS when the RSS samples are from NLOS conditions), false alarm probability p_f (deciding NLOS when the RSS samples are from LOS conditions), and overall misclassification probability $p_e = p_m + p_f$.

We observe that the identification results of RSS samples in the static environment (Table I) are far better than those in the dynamic environment (Table II). The best feature set in Table I can achieve a misclassification probability as low as 0.0648 while the most powerful feature set in Table II still maintains an error probability of 0.1401. For each feature set size, only the feature set with the lowest misclassification probability is presented in the table.

TABLE III. Missed detection probability (p_m), false alarm probability(p_f), and overall misclassification probability (p_e) of hypothesis testing algorithm. Features are extracted from every 1000 RSS samples in static environment.

Identification features	p_m	p_f	p_e
$\{\mu\}$	0.1108	0.0577	0.1685
$\{\mu,\, K_r\}$	0.0684	0.0993	0.1676
$\{\mu,\, K_r,\, \mathcal{K}\}$	0.0518	0.1050	0.1568
$\{\mu,\, K_r,\, \mathcal{K},\, \chi^2\}$	0.0041	0.3464	0.3505

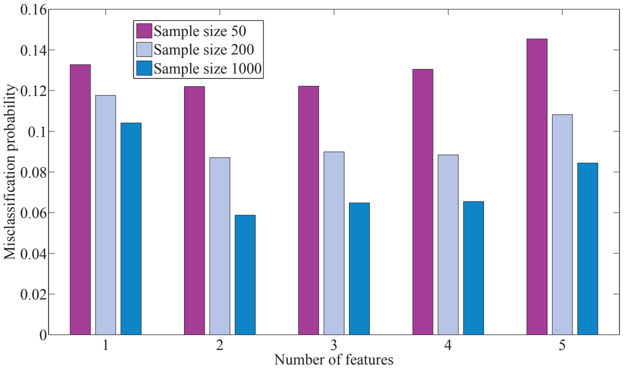


Fig. 3. Overall misclassification probability in static environment for different sample size. Features of different sizes are consistent with Table I.

From Table I, except for the mean of RSS samples, the Rician K factor K_r is contained in most feature sets, which indicates that the Rician K factor is a good indicator of the LOS/NLOS conditions in static environment. Although the Kurtosis $\mathcal{K}$ is a crucial feature in NLOS identification in UWB localisation, it is not included in any feature set here.

Different from Table I, the Rician K factor is not an essential feature in Table II any more. Instead, except for the mean, the skewness appears in each data set and thus becomes the most crucial feature. More importantly, the Kurtosis becomes a strong indicator for the NLOS identification.

The goodness of fit parameters between the RSS samples and their estimated distribution in LOS and NLOS conditions are also important for the NLOS identification in both static and dynamic environments.

Table III shows the performance of hypothesis testing based NLOS identification algorithm in static environments. It is observed that the best feature set (μ, K_r, $\mathcal{K}$) gives a misclassification rate of 0.1568. The reason for the difference in the performance of the two algorithms is that the hypothesis testing approach simplifies the relationship between different features as independent to reduce the computational complexity, which results in the loss of feature correlation information. Due to the limitation of space, we do not present the identification performance of hypothesis testing approach for dynamic environment in detail, where the best misclassification rate is 0.1917.

Figs. 3 and 4 compare the accuracy of identification using different sample sizes in static and dynamic environments, respectively. From the results we can see that the identification accuracy increases with sample size, which indicates that the number of samples collected at each location also has an impact on the identification results.

The reason for the impact of sample size on the identification accuracy is that a larger number of samples can reduce the influence of noisy RSS samples, which leads to a more precise fit of the samples to a distribution. As stated, our features from the measurements largely correlate with the estimated distribution. A better fit to the distribution makes the features more accurate and gives a better result.

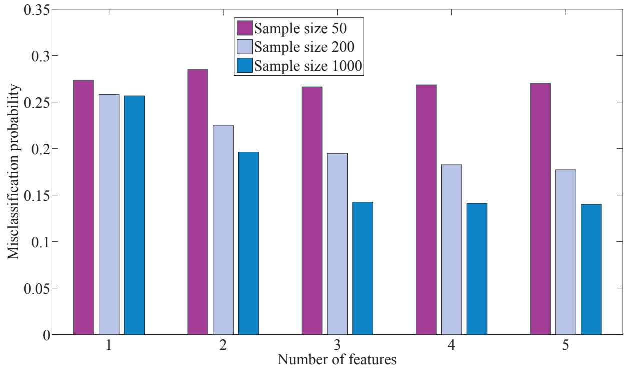


Fig. 4. Overall misclassification probability with external interference for different sample size. Features of different sizes are consistent with Table II.

Based on the number RSS samples required from the experiments, the number of packets exchanged during the receiving of a normal text email including the overhead (e.g. beacons, handshake, handoff) would be sufficient for this technique to provide an acceptable NLOS identification accuracy. An email with picture attachments contains hundreds of MAC layer packets which can make the NLOS identification very accurate without any change to the existing protocol stacks or other infrastructures.

We also observe from Figs. 3 and 4 that human interference plays an important role in the identification accuracy. In static environments, features from 50 samples can give a misclassification rate as low as 0.1248 which is better than the misclassification rate from 1000 samples in a dynamic environment. Therefore, among the factors that impact on the identification accuracy, influence from human interference is more important than the sample size.

D. NLOS Mitigation

In this subsection, we will discuss the accuracy improvement of distance estimation with NLOS mitigation techniques. All RSS measurements in the experiments are divided into two groups, the training group which is used to train the propagation models or the regression model and the test group which tests the accuracy of the models.

1) Standard Propagation Model (SPM): This is the first strawman algorithm that we use as a basis for comparison. We use least squares approximation to derive the parameters γ and WAF of the propagation model in Equation (2) that best fit the training RSS measurements for each anchor. Here the training RSS data include both LOS and NLOS measurements.

2) Breakpoint Propagation Model (BPM): This is the second benchmark algorithm that we use for comparison [12]. It is observed that the path loss in indoor environments as a function of distance has two distinct regions [13] which differ significantly in terms of propagation parameters. The breakpoint propagation model takes the point that separates the two regions, called breakpoint, into account in the log-normal

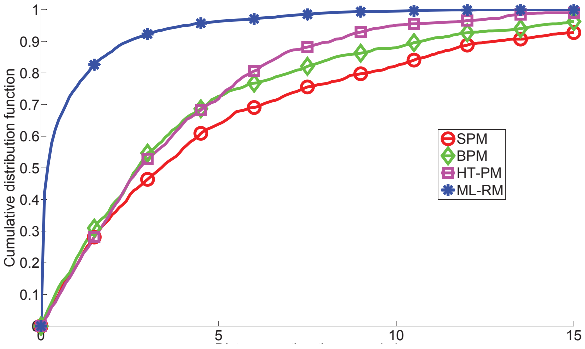


Fig. 5. Mean and RMS of distance estimation errors. Features are extracted from every 100 RSS samples.

propagation model and could estimate the transmitter-receiver distance more accurately than SPM. Similarly, we also use least squares approximation to derive the BPM parameters.

3) Hypothesis Testing Propagation Model (HT-PM): This is the first proposed algorithm which uses the hypothesis testing to identify LOS/NLOS conditions so as to estimate distances more accurately. We divide the training RSS measurements into two subsets: LOS and NLOS. We then use least squares approximation to derive propagation parameters in Equation (2) separately for the LOS subset, denoted with $\gamma^{(l)}$ and $WAF^{(l)}$, and the NLOS subset, denoted with $\gamma^{(n)}$ and $WAF^{(n)}$. After that, in practical implementations we use the hypothesis testing classifier to distinguish between LOS and NLOS conditions before we choose which propagation model (LOS or NLOS) should be used to estimate the distances.

4) Regression Model (ML-RM): This is the second and most important proposed algorithm. Instead of deriving the distance with only the mean of RSS measurements in the propagation models, the regression model takes into account more features of the RSS measurements and uses the LS-SVM regressor in the optimization problem (16) to estimate the distances. In addition to the same features introduced in Section III, we also use the distances estimated by the standard propagation model without NLOS identification as an input to the LS-SVM regressor. The use of the distance estimations in the regression model is to convert the RSS values from linear space to log space.

The distance estimation errors of different models are shown in Fig. 5. It is observed that the HT-PM technique only improves the distance estimation accuracy by around 20 percent compared to BPM. However, the ML-RM technique greatly outperforms other models in distance estimation accuracy, with the mean error from 6.61m for SPM to 0.86m for ML-RM. The main reason is that the propagation model only considers the mean of the RSS measurements while the regression model takes into account more features of the RSS measurements. In addition, the mean RSS and the distance estimated by SPM are key features in the regression model analysis.

E. Robustness

To test the robustness of the machine learning based NLOS identification and mitigation algorithms, we use the same set of parameters trained from the experimental site in Fig. 2(a) (site (a) hereafter, same for site (b)) to identify and mitigate the NLOS conditions in a different experimental site in Fig. 2(b).

The overall misclassification rate is 0.0909 with the best feature set $(\mu,\ K_r,\ \chi^2)$ in Table I. The average distance estimation error is 2.84m, with an improvement of over 50 percent in accuracy compared with the propagation model.

With the best 3-feature set $(\mu,\ K_r,\ \chi^2)$, the overall misclassification rate of the algorithm is 0.0909 when tested in a different building than trained, as opposed to 0.0648 when tested in the same building as trained (see Table I). In terms of average distance estimation error, we observed it to be 2.84m when the regressor was tested in a different building than the one used for training, as opposed to 0.86m when it was trained and tested in the same building. The observed error of 2.84m already offers over 50 percent improvement over the standard propagation model (SPM) approach.

VII. Impact on Positioning System

Our NLOS identification and mitigation techniques are implemented in indoor localisation. With the distance estimation techniques in Section VI, the location of the mobile phones can be determined with simple trilateration. After that, these locations are smoothed with a simple particle filter.

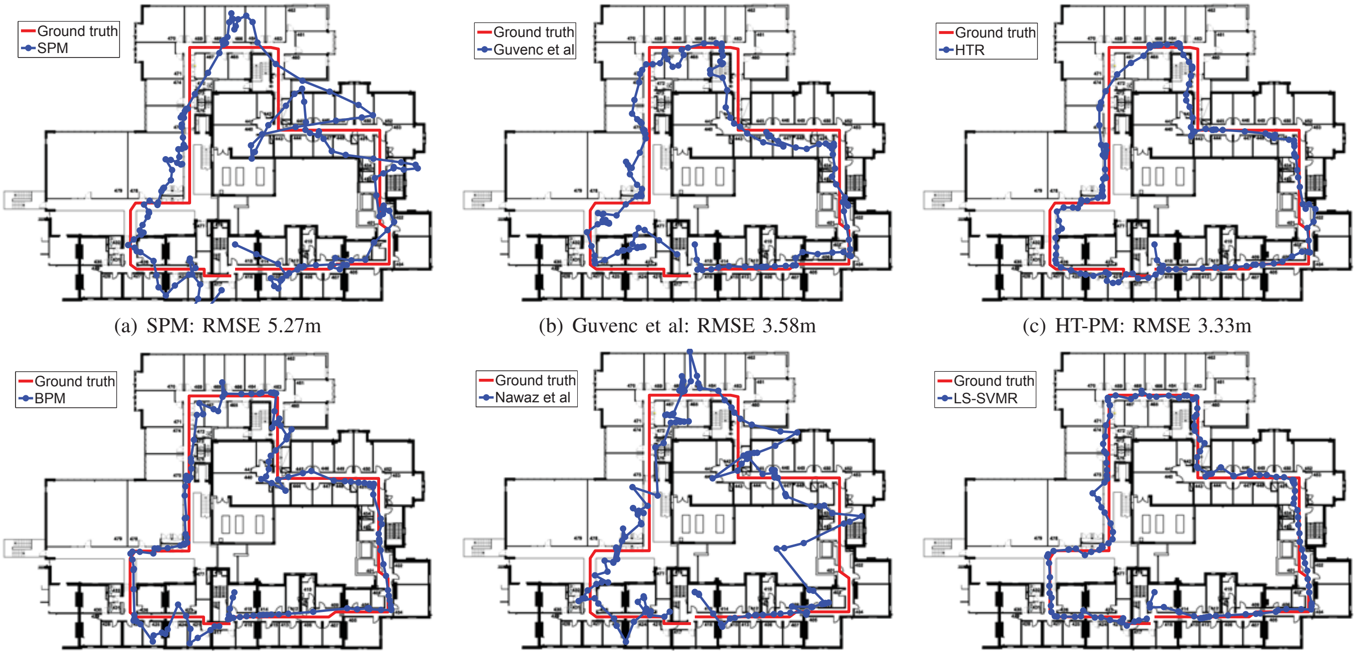
Fig. 6 compares the trajectories estimated from different range-based localisation approaches. All trajectories in this figure are estimated from the same raw RSS measurements. The trajectories in the two figures in the left column (Figs. 6(a) and 6(d)) are calculated with distances estimated from SPM and BPM. The trajectories in the middle column (Figs. 6(b) and 6(e)) are generated from distance estimations using the state-of-the-art NLOS identification and mitigation approaches with RSS measurements. The trajectories in the right column (Figs. 6(c) and 6(f)) are estimated from the two distance estimation algorithms proposed in this paper: HT-PM (hypothesis testing) and ML-RM (machine learning).

It is observed from Fig. 6 that the approach proposed by Guvenc et al [2], which selects the access points subset with minimum weighted residual, works fairly well in our experiments. In addition, recall that the generic NLOS mitigation approach proposed by Nawaz et al in [6] tries to recover the estimation errors by assuming that LOS anchors outnumber NLOS anchors. In our experiments their proposed algorithm results in very poor performance at many locations where there are more NLOS anchors than LOS anchors.

Fig. 6 also shows that there is only slight improvement in accuracy for the propagation model derived from hypothesis testing results, compared with the generic NLOS identification and mitigation approaches proposed in existing works [2], [6]. The incapability for a simple propagation model to capture major features of complicated indoor environments results in this phenomenon.

We can also see from Fig. 6 that the localisation system based on the regression model could improve the localisation accuracy by 60 percent compared with the trajectories estimated with the state-of-the-art NLOS mitigation algorithms, which greatly increases the potential of using WiFi-based localisation in practical settings.

In addition, our model outperforms fingerprinting approach [14] in terms of training time and complexity because



(a) SPM: RMSE 5.27m

(b) Guvenc et al: RMSE 3.58m

(c) HT-PM: RMSE 3.33m

(d) BPM: RMSE 4.14m

(e) Nawaz el al: RMSE 3.85m

(f) LS-SVMR: RMSE 1.30m

Fig. 6. Trajectories generated by different algorithms, showing the efficiency of our NLOS identification and mitigation algorithms.

our model parameters trained from one site can also be used in another site while fingerprinting approach has to train its model for each individual site before it can be used.

VIII. Conclusion and Future Work

In this paper, we have proposed two NLOS identification and mitigation algorithms using only WiFi RSS measurements. Two optimization problems are developed to solve the problems of NLOS identification and distance estimation from RSS measurements. To our knowledge, this is the first identification and mitigation method that is solely based on RSS samples from a WiFi service on mobile phones. The extensive experimental results have shown the accuracy of the proposed NLOS identification and mitigation algorithms.

Our future work will incorporate some other information available in the building, like the map and the location of access points for localization purposes, to reduce or even eliminate the training phase of the machine learning based algorithm and develop online learning algorithms or unsupervised machine learning algorithms to identify the LOS/NLOS conditions. Alternatively, our model can also be initialized with our data sets and then finetuned to new environments.

Acknowledgements The authors would like to acknowledge the support of the EPSRC through grants EP/J012017/1 and the anonymous reviewers for their helpful comments.

References

[1] K. Yu, Y. J. Guo, and S. Member, “Statistical NLOS identification based on AOA , TOA , and signal strength,” *IEEE Trans. Veh. Technol.*, vol. 58, no. 1, pp. 274–286, 2009.

[2] I. Guvenc, C. Chong, and F. Watanabe, “NLOS identification and mitigation for UWB localization systems,” in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC’07)*, pp. 1571–1576, Ieee, 2007.

[3] S. Venkatesh and R. M. Buehrer, “Non-line-of-sight identification in ultra-wideband systems based on received signal statistics,” *IET Microw. Antennas Propag.*, vol. 1, no. 6, pp. 1120–1130, 2007.

[4] J. Khodjaev, Y. Park, and A. Saeed Malik, “Survey of NLOS identification and error mitigation problems in UWB-based positioning algorithms for dense environments,” *Ann. of Telecommun.*, vol. 65, pp. 301–311, Aug. 2010.

[5] S. Maran, W. M. Gifford, and H. Wymeersch, “NLOS identification and mitigation for localization based on UWB experimental data,” *IEEE J. Sel. Areas Commun.*, vol. 28, no. 7, pp. 1026–1035, 2010.

[6] S. Nawaz and N. Trigoni, “Convex Programming Based Robust Localization in NLOS Prone Cluttered Environments,” in *Proc. 10th Int. Conf. Informat. Process. Sensor Netw (IPSN)*, (Chicago, IL, USA), pp. 318–329, 2011.

[7] P.-c. Chen, “A non-line-of-sight error mitigation algorithm in location estimation,” in *Proc. IEEE Wirel. Commun. Netw. Conf. (WCNC’99)*, pp. 316–320, Ieee, 1999.

[8] X. Li, “An iterative NLOS mitigation algorithm for location estimation in sensor networks,” in *Proc. 15th IST Mob. Wireless Commun. Summit*, (Miconos, Greece), 2006.

[9] C. Tepedelenlioglu, A. Abdi, and G. B. Giannakis, “The Ricean K factor: estimation and performance analysis,” *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 799–810, 2003.

[10] C. G. Koay and P. J. Basser, “Analytically exact correction scheme for signal extraction from noisy magnitude MR signals.,” *J. magn. resonance*, vol. 179, pp. 317–22, Apr. 2006.

[11] J. SuYkens and J. Vandewalle, “Least squares support vector machine classifiers,” *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, 1999.

[12] K. Cheung and R. D. Murch, “A new empirical model for indoor propagation prediction,” *IEEE Trans. Veh. Technol.*, vol. 47, no. 3, pp. 996–1001, 1998.

[13] W. Honcharenko, H. L. Bertoni, and J. Dailing, “Mechanisms Governing Propagation Between Different Floors in Buildings,” *IEEE Trans. Antennas Propag.*, vol. 41, no. 6, pp. 787–790, 1993.

[14] P. Bahl and V. N. Padmanabhan, “RADAR : An in-building RF-based user location and tracking system,” in *Proc. 19th Annu. IEEE Int. Conf. Comput. Commun. (INFOCOM 2000)*, vol. 2, (Tel-Aviv, Israel), pp. 775–784, 2000.