Multivariate Analysis for Probabilistic WLAN Location Determination Systems

Moustafa Youssef
Department of CS
University of Maryland
College Park, Maryland 20742
moustafa@cs.umd.edu

Mohamed Abdallah
Department of ECE
University of Maryland
College Park, Maryland 20742
mabdalah@umd.edu

Ashok Agrawala
Department of CS
University of Maryland
College Park, Maryland 20742
agrawala@cs.umd.edu

Abstract

WLAN location determination systems are gaining increasing attention due to the value they add to wireless networks. In this paper, we present a multivariate analysis technique for enhancing the performance of WLAN location determination systems by taking the correlation between samples from the same access point into account. We show that the autocorrelation between consecutive samples from the same access point can be as high as 0.9. Giving a sequence of correlated signal strength samples from an access point, the technique estimates the user location based on the calculated probability of this sequence from the multivariate distribution. We use a linear autoregressive model to derive the multivariate distribution function for the correlated samples. Using analytical analysis, we show that the proposed technique provides better location accuracy over previous techniques especially for the highly correlated samples in a typical WLAN environment. Implementation of the technique in the Horus WLAN location determination system shows that the average system accuracy is increased by more than 64%. This significant enhancement in the accuracy of WLAN location determination systems helps increase the set of context-aware applications implemented on top of these systems.

1. Introduction

As 802.11-based wireless LANs become more ubiquitous, the importance of WLAN location determination systems [4–7,11,14,16,20–22,26–33] increases. Such systems are purely software based and therefore add to the value of the wireless network. A large class of applications, including [8] location-sensitive content delivery, direction finding, asset tracking, and emergency notification, can be built on top of such systems. This set of applications can be broadened as the accuracy of WLAN location determination system increases.

WLAN location determination systems usually work in two phases: *offline* training phase and *online* location determination phase. During the offline phase, the signal strength received from the access points (APs) at selected locations in the area of interest is tabulated, resulting in a so-called *radio map*. During the location determination phase, the signal strength samples received from the access points are used to "search" the radio map to estimate the user location.

Radio-map based techniques can be categorized into two broad categories: deterministic techniques and probabilistic techniques. *Deterministic techniques* [4,5,14,22] represent the signal strength of an access point at a location by a scalar value, for example, the mean value, and use non-probabilistic approaches to estimate the user location. For example, in the *Radar* system [4,5] the authors use nearest neighborhood techniques to infer the user location. On the other hand, *probabilistic* techniques [6,7,16,20,21,26–33] store information about the signal strength distributions from the access points in the radio map and use probabilistic techniques to estimate the user location. For example, the *Horus* system [26–33] uses a Bayesian-based approach to estimate the user location.

WLAN location determination systems need to deal with the noisy characteristics of the wireless channel to achieve higher accuracy. In this paper, we analyze one aspect of the

Proceedings of the Second Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services (MobiQuitous'05)
0-7695-2375-7/05 $20.00 © 2005 IEEE

temporal characteristics of the wireless channel: samples correlation from the same access point. We show that consecutive samples can have correlation as high as 0.9. The main challenge is how to use multiple samples to obtain a better location estimate technique despite this high correlation value. Our approach is to start from an autoregressive model that captures the correlation of samples from the same access point. Based on the autoregressive model, we present a technique that derive the multivariate distribution for multiple consecutive samples from each access point. This multivariate distribution is used to estimate the probability of any sequence of signal strength samples.

We quantify the advantage of the multivariate analysis technique over the simple samples averaging technique analytically and by testing them in an actual testbed in Sections 4 and 5 respectively. The advantage of the multivariate analysis technique increases with the increase in the correlation value. This shows the importance of the proposed technique especially for the high correlation values in typical WLAN environments. We also show that the proposed technique, when implemented in the context of the *Horus* probabilistic WLAN location determination system, enhances the accuracy by more than 64%. This significant enhancement in accuracy helps in increasing the set of applications that can be built on top of the WLAN location determination systems and hence increases their value.

The rest of the paper is structured as follows: in the next Section, we present a brief introduction to probabilistic WLAN location determination systems and analyze the autocorrelation of samples from the same access point. We describe our multivariate analysis technique that handles the correlation between signal strength samples in Section 3. Section 4 analytically quantifies the advantage of the multivariate analysis technique over the simple samples averaging technique. In Section 5, we present the results of implementing the new technique and compare its accuracy to the accuracy of the original *Horus* technique and samples averaging technique. Section 6 discusses related work. Finally, Section 7 highlights the main findings of the paper and provides concluding remarks.

2. Introduction

2.1. Probabilistic WLAN Location Determination Systems

The basic approach used in probabilistic WLAN location determination systems, e.g. [6, 7, 16, 20, 21, 26–33], is Bayesian-inversion. Let $S = [\bar{s}_1, ..., \bar{s}_k]$ be the signal strength matrix from k access points, where each entry is a column vector representing N consecutive signal strength samples from an access point ($\bar{s}_i = [s_i[1]...s_i[N]]^T$), where $(\cdot)^T$ denotes the transpose of a vector. The system returns the location x among the set of radio map locations $\mathbb{X}$ that maximizes $P(x|S)$, *i.e.*

$$\underset{x \in \mathbb{X}}{\text{argmax}}[P(x|S)] \tag{1}$$

Using Baye's theorem, this can be rewritten as:

$$\underset{x \in \mathbb{X}}{\text{argmax}}[P(x|S)] = \underset{x \in \mathbb{X}}{\text{argmax}}[\frac{P(S/x).P(x)}{P(S)}] \tag{2}$$

Since $P(S)$ is constant for all x, we can rewrite Equation 2 as:

$$\underset{x \in \mathbb{X}}{\text{argmax}}[P(x|S)] = \underset{x \in \mathbb{X}}{\text{argmax}}[P(S|x).P(x)] \tag{3}$$

where $P(x)$ represents the probability of finding the user at location x (the user profile), $P(S|x)$ is calculated by:

$$P(S|x) = \prod_{i=1}^{k} P(\bar{s}_i|x) \tag{4}$$

and

$$P(\bar{s}_i|x) = \int_{s_i[1]-\Delta/2}^{s_i[1]+\Delta/2} \cdots \int_{s_i[N]-\Delta/2}^{s_i[N]+\Delta/2} f(\bar{q}|x)\, d\bar{q} \tag{5}$$

where $f(\cdot)$ is the probability density function (PDF) at location x for AP i. The value Δ used in the integration limits denotes the quantization error of the wireless interface card in use.

In this paper, we focus on developing a mathematical representation of $P(\bar{s}_i|x)$ for N consecutive correlated signal strength samples. Also, we assume that the user profile, *i.e.* the distribution of $P(x)$, is computed offline [1]. For the rest of the paper, we will drop the subscript i for sake of clarity.

2.2. Samples Correlation

Figure 1 shows the autocorrelation function of the samples collected from one access point (one sample per second) at a fixed position. The figure shows that the autocorrelation of consecutive samples ($lag = 1$) is as high as 0.9. This high autocorrelation is expected as over a short period of time the signal strength received from an access point at a particular point is relatively stable (modulo the changes in the environment).

This high autocorrelation value should be considered when using the methods that use multiple samples, especially for *probabilistic* location determination techniques.

[1] Interested readers can find more details about estimating the user profile and its effect on accuracy in [25].

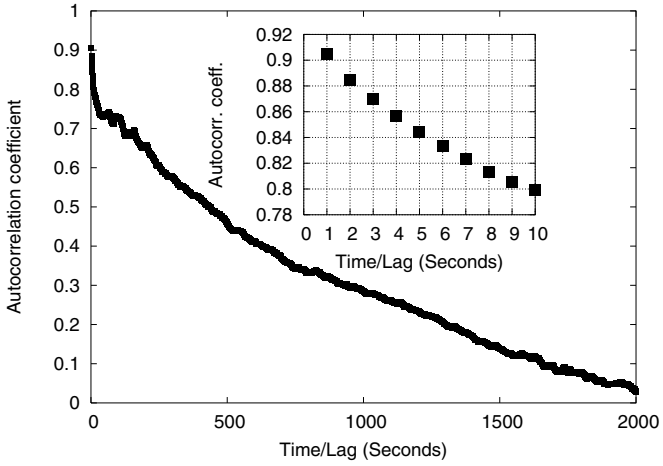

Figure 1. An example of the autocorrelation between samples from an access point. The sub-figure shows the autocorrelation for the first 10 lags.

Figure 5 shows the effect of averaging samples on the accuracy of a probabilistic WLAN location determination system that assumes the *independence* of samples[2]. The figure shows that although averaging increases accuracy, the wrong independence assumption leads to increasing average distance error *increases* as the number of averaged samples increases. The goal of this paper is to take the high samples correlation into account to further enhance the performance of probabilistic WLAN location determination systems.

3. Multivariate Analysis of Signal Strength Samples

Due to the random nature of the wireless medium, the samples collected from each access points can be viewed as a sequence of random variables. Such sequence collected in time forms a stochastic process denoted in the rest of the paper as $s[n]$. In this Section, we present the statistical characterization of N consecutive samples of the stochastic process $s[n]$ collected in the signal strength column vector $\bar{s}$ defined as,

$$\bar{s} = \left[\begin{array}{cccc} s[1] & s[2] & \ldots & s[N] \end{array} \right]^T \quad (6)$$

To this end, we first present the temporal characteristics of $s[n]$. Then, we exploit such characterization to determine the probability density function (PDF) and cumulative distribution function (CDF) of the signal strength vector $\bar{s}$ at a fixed location x which are required to compute $P(\bar{s}|x)$ defined in Equation 5.

[2]The figure is discussed in more details in Section 5.

3.1. Temporal Characteristics of Signal Strength Samples

As shown in [27], analyzing the signal strength samples collected from an access point leads to the following properties about the process $s[n]$,

- The process $s[n]$ is wide sense stationary process where each sample follows a Gaussian distribution with mean μ_s and variance σ_s^2.
- The samples of $s[n]$ are highly correlated and its temporal variations can be modeled as a first order autoregressive process as follows,

$$s[n] = \alpha s[n-1] + \sqrt{1-\alpha^2}\sigma_s v[n] + \mu_s(1-\alpha) \quad (7)$$

  where α denotes the degree of correlation between two consecutive samples taking the values $0 \leq \alpha < 1$ and $v[n]$'s are independent identically-distributed (i.i.d) samples with zero-mean unit-variance Gaussian distribution and are independent of $s[n]$.

In the next Section, we exploit such properties to derive the probability density function and the cumulative distribution function for the signal strength vector $\bar{s}$ of N consecutive samples from the same AP.

3.2. PDF of the Signal Strength Vector

We start by proving that the PDF of the signal strength vector $\bar{s}$ is a multivariate Gaussian. Then we develop expressions of the mean and the covariance of $\bar{s}$ required to compute the Gaussian PDF. Without loss of generality, we assume that the process $s[n]$ has zero mean, *i.e.*, $\mu_s = 0$.

To prove that the PDF of the signal strength vector $\bar{s}$ is a multivariate Gaussian, we use the following theorem [18]:

Theorem 1 *The signal vector $\bar{s}$ is a multivariate Gaussian iff for any constant vector $\bar{a} = \left[\begin{array}{cccc} a[1] & a[2] & \ldots & a[N] \end{array} \right]^T$, the dot product $\bar{a}^t\bar{s}$ has a Gaussian distribution.*

Using the autoregressive model in Equation 7, the N^{th} signal sample $s[N]$ of the vector $\bar{s}$ can be represented in terms of the first sample $s[1]$ and the N samples $\{v[1], v[2], \ldots, v[N]\}$ as follows,

$$s[N] = \alpha^n s[1] + \sqrt{(1-\alpha^2)}\sigma_s \sum_{j=1}^{N} \alpha^{N-j} v[j]. \quad (8)$$

The term $\bar{a}^T\bar{s}$ can be represented as follows,

$$\begin{aligned}\bar{a}^T\bar{s} &= \sum_{i=1}^{N} a[i](\alpha^i s[1] + \sqrt{(1-\alpha^2)}\sigma_s \sum_{j=1}^{i} \alpha^{i-j} v[j]) \\ &= \sum_{i=1}^{N} a[i]\alpha^i s[1] + \\ &\sqrt{(1-\alpha^2)}\sigma_s \sum_{i=1}^{N}\sum_{j=1}^{i} a[i]\alpha^{i-j} v[j].\end{aligned} \tag{9}$$

Equation 9 shows that the term $\bar{a}^T\bar{s}$ is represented by a linear sum of the independent Gaussian variables: $s[1]$ and $\{v[1], v[2], \ldots, v[N]\}$. Since the sum of independent Gaussian variables is a Gaussian variable, the term $\bar{a}^T\bar{s}$ is Gaussian for any selection of the vector $\bar{a}$. Therefore, from Theorem 1, the signal vector $\bar{s}$ has a multivariate Gaussian distribution.

The multivariate Gaussian PDF $f(\bar{s})$ of the signal strength vector $\bar{s}$ is then given by [18]:

$$f(\bar{s}) = \frac{1}{\sqrt{(2\pi)^N |\Sigma_{\bar{s}}|}} \exp(\frac{-\bar{s}^T \Sigma_{\bar{s}}^{-1} \bar{s}}{2}) \tag{10}$$

where $\Sigma_{\bar{s}}$ is the covariance matrix given by

$$\Sigma_{\bar{s}} = E\left[\bar{s}\bar{s}^T\right] \tag{11}$$

with determinant $|\Sigma_{\bar{s}}|$ and inverse $\Sigma_{\bar{s}}^{-1}$.

The elements of the covariance matrix can be represented in terms of the variance σ_s and the correlation factor α. The (i,k) element of the covariance matrix $\Sigma_{\bar{s}}(i,k) = E\left[s[i]s[k]\right]$ is expressed using the autoregressive model (Equation 7) as:

$$\begin{aligned}E\left[s[i]s[k]\right] =& E[(s[i])(\alpha^{(i-k)} s[i] + \\ & \sqrt{(1-\alpha^2)}\sigma_s \sum_{j=1}^{k} \alpha^{k-i} v[j])]\end{aligned} \tag{12}$$

Since $s[i]$ and $\{v[1], v[2], \ldots, v[k]\}$ are independent, then

$$E\left[s[i]s[k]\right] = \alpha^{(i-k)}\sigma_s^2 \tag{13}$$

Hence, the covariance matrix $\Sigma_{\bar{s}}$ can be formulated as follows,

$$\Sigma_{\bar{s}} = \sigma_s^2 \begin{bmatrix} 1 & \alpha & \ldots & \alpha^{N-1} \\ \alpha & 1 & \ldots & \alpha^{N-2} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha^{N-1} & \alpha^{N-2} & \ldots & 1 \end{bmatrix} \tag{14}$$

We remark that the covariance matrix $\Sigma_{\bar{s}}$ is a symmetric semi-definite Toeplitz matrix [12]. Moreover, the covariance matrix is highly structured as it can be calculated by knowing only the value of σ_s and α. This means that the storage requirement for this matrix and the computational requirements for the typical values of N are minimal.

3.3. CDF of the Signal Strength Vector

In this Section, we show how to compute the distribution function $F(\bar{s})$ of the multivariate Gaussian random vector $\bar{s}$, defined as follows,

$$F(\bar{s}) = \int_{-\infty}^{s[1]} \int_{-\infty}^{s[2]} \ldots \int_{-\infty}^{s[N]} f(\bar{s}) \mathrm{d}\bar{s}. \tag{15}$$

where as shown in Section 3.2, $f(\bar{s})$ is given by:

$$f(\bar{s}) = \frac{1}{\sqrt{(2\pi)^N |\Sigma_{\bar{s}}|}} \exp(\frac{-\bar{s}^T \Sigma_{\bar{s}}^{-1} \bar{s}}{2}) \tag{16}$$

Since $\Sigma_{\bar{s}}$ is a symmetric semi-definite Toeplitz matrix, then using Chelosky decomposition [12], $\Sigma_{\bar{s}}^{-1}$ can be decomposed as follows:

$$\Sigma_{\bar{s}}^{-1} = H_{\bar{s}}^T H_{\bar{s}}, \tag{17}$$

where $H_{\bar{s}}$ is an upper diagonal matrix computed as follows

$$H_{\bar{s}} = \Lambda^{-1/2} V, \tag{18}$$

where Λ is a diagonal matrix whose elements are the eigenvalues of $\Sigma_{\bar{s}}^{-1}$ and V is the eigenvector matrix of $\Sigma_{\bar{s}}^{-1}$ respectively.

Substituting Equations 17 and 18 in Equation 15, and applying the following transformation of variables

$$\bar{y} = H_{\bar{s}}\bar{s} \tag{19}$$

the exponent term in the PDF equation $f(\bar{s})$ can be expressed in terms of the vector $\bar{y}$ as:

$$\begin{aligned}\bar{s}^T \Sigma_{\bar{s}}^{-1} \bar{s} &= \bar{s}^T H_{\bar{s}}^T H_{\bar{s}} \bar{s} \\ &= (H_{\bar{s}}\bar{s})^T (H_{\bar{s}}\bar{s}) \\ &= \bar{y}^T \bar{y}\end{aligned} \tag{20}$$

Hence, the PDF $f(\bar{s})$ can be rewritten as follows,

$$f(\bar{s}) = \frac{1}{\sqrt{(2\pi)^N |H_{\bar{s}}^T||H_{\bar{s}}|}} \exp(\frac{-1}{2}\bar{y}^T\bar{y}) \tag{21}$$

Applying transformation of variables requires expressing the differential $\mathrm{d}\bar{s}$ in terms of $\mathrm{d}\bar{y}$ which can be achieved by using the following equation:

$$\mathrm{d}\bar{s} = \mathrm{J}.\mathrm{d}\bar{y} \tag{22}$$

where J is the Jacobian matrix defined as follows [13],

$$\mathrm{J} = \begin{bmatrix} \frac{\partial s[1]}{\partial y[1]} & \frac{\partial s[1]}{\partial y[2]} & \ldots & \frac{\partial s[1]}{\partial y[N]} \\ \frac{\partial s[2]}{\partial y[1]} & \frac{\partial s[2]}{\partial y[2]} & \ldots & \\ \frac{\partial s[2]}{\partial y[1]} & & & \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial s[N]}{\partial y[1]} & \frac{\partial s[N]}{\partial y[2]} & \ldots & \frac{\partial s[N]}{\partial y[N]} \end{bmatrix}. \tag{23}$$

Computing the differential elements of the Jacobian matrix yields J $= |H_{\bar{s}}|^{-1}$. Hence, the distribution function can be rewritten as follows,

$$F(\bar{s}) = \int_{-\infty}^{y[1]} \int_{-\infty}^{y[2]} \cdots \int_{-\infty}^{y[N]} \frac{|H_{\bar{s}}|^{-1}}{\sqrt{(2\pi)^N |H_{\bar{s}}^T||H_{\bar{s}}|}} \exp(\frac{-1}{2}\bar{y}^T\bar{y})\mathrm{d}\bar{y}. \quad (24)$$

By using the relation $|H_{\bar{s}}|^T = |H_{\bar{s}}|$ [12], the distribution function is given as follows:

$$F(\bar{s}) = \int_{-\infty}^{y[1]} \int_{-\infty}^{y[2]} \cdots \int_{-\infty}^{y[N]} \frac{1}{\sqrt{(2\pi)^N}} \exp(\frac{-1}{2}\bar{y}^T\bar{y})\mathrm{d}\bar{y}. \quad (25)$$

Finally, $F(\bar{s})$ can be directly computed in terms of the complementary error function $\mathrm{erfc(x)}$ as follows,

$$F(\bar{s}) = (0.5)^N \prod_{i=1}^{N} \mathrm{erfc(y[i]}/\sqrt{2}) \quad (26)$$

To summarize, given a signal vector $\bar{s}$ and its covariance matrix $\Sigma_{\bar{s}}$, we obtain $\bar{y}$ from Equation 19 and computes $F(\bar{s})$ using Equation 25. We emphasize that the above expression for computing $F(\bar{s})$ also holds if the signal vector $\bar{s}$ has a non-zero mean μ_s.

We note that the computational requirements for calculating $F(\bar{s})$ are minimal due to the fact that the covariance matrix is a Toeplitz matrix and the value of N is typically less than 5 samples.

3.4. Modified *Horus* Algorithm

In this Section, we use the results of the previous Section to obtain the probability of a given vector of N consecutive correlated signal strength samples. We use this probability to determine the most probable user location. The technique works as follows:

- *Offline phase*: the system calculates the parameters of CDF for the multivariate distribution of N samples for each access point in the radio map.
- *Online phase*: Given a vector of N consecutive samples from an access point, the algorithm obtains the probability of this vector using the radio map constructed in the offline phase.

Algorithm 1 shows the details of the modified *Horus* algorithm. Note that the value of α is implicitly used in the online phase as the multivariate distribution of the average of N samples depends on the value of α as discussed in Section 3.3.

Alg. 1 $l_{\max}$= Multi_Horus_GetLocation (N, S, $\mathbb{X}$,$P(\cdot)$)

Input:

N : Number of samples from each access point.

S : Measured signal strength vectors from k access points ($S = [\bar{s_1}, ..., \bar{s_k}]$). Each $\bar{s_i}, 1 \leq i \leq k$ is a vector containing N samples from access point i.

$\mathbb{X}$: Radio map locations.

$P(\cdot)$: A radio map based function, where $P(\bar{s}_i|x)$ returns the probability of receiving the signal strength vector $\bar{s}_i$ from the ith access point at location $x \in \mathbb{X}$.

Output:

The location $l_{\max} \in \mathbb{X}$ that maximizes $P(x/S)$.

1: Max $\leftarrow 0$
2: **for** $x \in \mathbb{X}$ **do**
3: $p \leftarrow \prod_{i=1}^{k} P(\bar{s}_i|x)$
4: **if** $p >$ Max **then**
5: $l_{\max} \leftarrow x$
6: $Max \leftarrow p$
7: **end if**
8: **end for**

4. Multivariate versus Averaging

In this Section, We use the information theory framework [9] to quantify the advantage of the amount of information revealed by the full strength vector compared to averaging the samples, taking correlation into account, as discussed in [27]. In addition, we show that such information increases as the degree of correlation between samples increases which motivate the advantage of using the multivariate analysis in the typical wireless environment where samples are highly correlated. Specifically, we define I_d, the information gain achieved using the multivariate analysis technique, as follows,

$$I_d = \log_2 \frac{1}{\mathrm{P(\bar{s})}} - \log_2 \frac{1}{\mathrm{P(s_{avg})}} \quad (27)$$

where $s_{\mathrm{avg}} = \frac{1}{N}\sum_{i=1}^{N} s[i]$ is the average of vector $\bar{s}$ defined in Equation (6).

From equations 10 and 5, $P(\bar{s})$ is given by,

$$P(\bar{s}) = \int_{s[1]-\Delta/2}^{s[1]+\Delta/2} \int_{s[2]-\Delta/2}^{s[2]+\Delta/2} \cdots \int_{s[N]-\Delta/2}^{s[N]+\Delta/2} \frac{1}{\sqrt{(2\pi)^N |\Sigma_{\bar{s}}|}} \exp(\frac{-1}{2}\bar{y}\Sigma_{\bar{s}}^T\bar{y})\mathrm{d}\bar{y} \quad (28)$$

and s_{avg} is a Gaussian variable with zero mean and variance

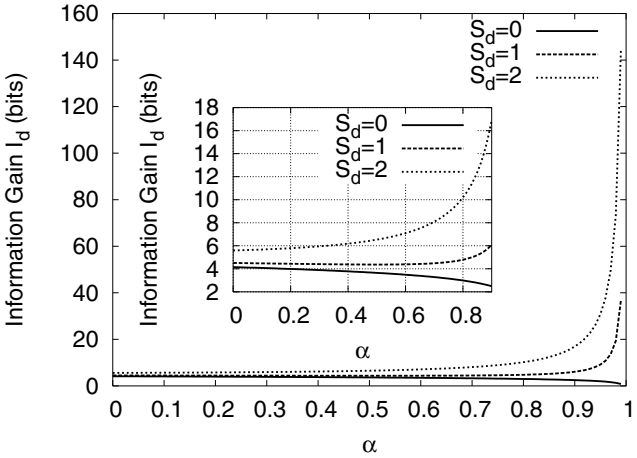

Figure 2. Information gain I_d as a function of the correlation factor α for various values of the sample difference, $s_d = 0, 1, 2$. The subfigure shows the values for $\alpha = 0 - 0.9$.

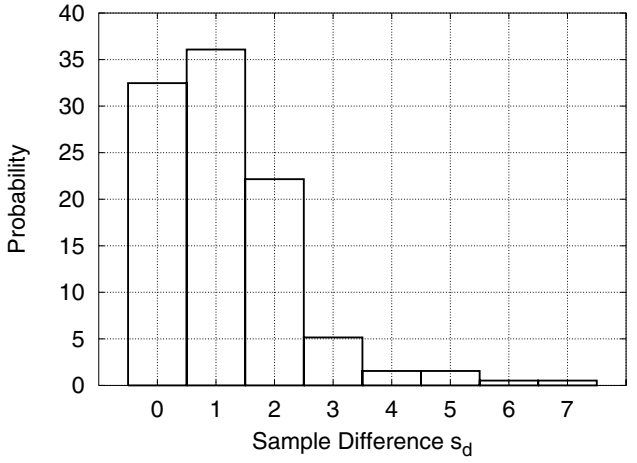

Figure 3. A typical sample difference histogram collected from one access point at one location. For 68% of the time, the sample difference is greater than zero. The higher the the sample difference, the higher the information gain.

$\sigma_{s_{\text{avg}}}$, *i.e.*,

$$P(s_{\text{avg}}) = \int_{s_{\text{avg}}-\Delta/(2N)}^{s_{\text{avg}}+\Delta/(2N)} \frac{1}{\sqrt{(2\pi)\sigma^2_{s_{\text{avg}}}}} \exp(\frac{-1}{2\sigma_{s_{\text{avg}}}} y^2) \mathrm{d}y \tag{29}$$

A detailed analysis of how to compute $P(\bar{s})$ and $P(s_{\text{avg}})$ is give in Section 3.3 and Reference [27] respectively. For the sake of simplifying the analysis, we only show the analysis for the case of 2 samples only. In addition, we approximate the probabilities defined in Equations (28) and (29) by assuming that the PDFs are constant within the range of integration. Such assumption is valid for the typical values of variance in an actual wireless environment where the ratio of variance of the signal σ_s to Δ is large enough to guarantee that the PDF is constant over the integration range. Applying these approximations, the probabilities can be rewritten as follows:

$$P(\bar{s}) \approx \Delta^N \frac{1}{\sqrt{(2\pi)^N |\Sigma_{\bar{s}}|}} \exp(\frac{-1}{2} \bar{s} \Sigma_{\bar{s}}^T \bar{s}) \tag{30}$$

and

$$P(s_{\text{avg}}) \approx \frac{\Delta}{N} \frac{1}{\sqrt{(2\pi)\sigma^2_{s_{\text{avg}}}}} \exp(\frac{-1}{2\sigma_{s_{\text{avg}}}} s^2_{\text{avg}}) \tag{31}$$

Substituting the from equations 30 and 31 in equation 27 for $N = 2$, we obtain the following,

$$I_d \approx \log_2 \frac{1}{2.\mathrm{P}(\mathrm{s_d})} \tag{32}$$

where $s_d = s[1] - s[2]$ is the sample difference and $P(s_d)$ is given by:

$$P(s_d) = \frac{1}{\sqrt{(2\pi)2\sigma_s^2(1-\alpha)}} \exp(\frac{-(s_d)^2}{4\sigma_s^2(1-\alpha)}) \tag{33}$$

Figure 2 shows the information gain, I_d, as a function of the correlation factor α for various values of the sample difference ($s_d = s[1] - s[2]$) (for $\sigma_s^2/\Delta = 10$). The Figure reveals the following remarks about the performance of the multivariate analysis technique compared to the samples averaging technique:

- For the valid range of our approximation (i.e. high values of σ/Δ), I_d is *positive for all* values of α which shows that using the multivariate analysis technique yields more information compared to the samples averaging technique.
- The behavior of the information gain I_d varies based on the value of the sample difference. Figure 3 shows an example of the histogram of the sample difference from an AP at a particular location. For the frequent case of sample difference of one or higher (68% of the time), I_d increases as the correlation increases till it reaches *infinity* at $\alpha = 1$.

  For sample difference value of zero, we observe that I_d is still positive. However, it decreases as the correlation factor α increases till it reaches zero at $\alpha = 1$.

This can be explained by noting that for a correlation of one, $\alpha = 1$, the probability of receiving two identical samples is one for both the multivariate analysis technique and the samples averaging technique. This means that both techniques give zero information and therefore, the information gain is zero.

In summary, the amount of information used by the multivariate analysis technique is more than that available for the samples averaging technique especially for highly correlated samples. In the next Section, we confirm our findings through testing in an actual environment.

5. Experimental Evaluation

In this Section we present the result of implementing the multivariate analysis technique in the context of the *Horus* system.

5.1. Experimental Testbed

We performed our experiment in the south wing of the fourth floor of the Computer Science Department building. The layout of the floor is shown in Figure 4. The wing has a dimension of 224 feet by 85.1 feet. The technique was tested in the Computer Science Department wireless network. The entire wing is covered by 12 access points installed in the third and fourth floors of the building.

For building the radio map, we took the radio map locations on the corridors on a grid with cells placed 5 feet apart (the corridor's width is 5 feet). We have a total of 110 locations along the corridors. On the average, each location is covered by 4 access points. The value of α, autocorrelation degree, for these access points was approximately 0.9 for all access points.

Using the device driver and the API we developed [1], we collected 300 samples at each location, one sample per second. The cards used were Lucent Orinoco silver NICs supporting up to 11 Mbit/s data rate [2]. *To test the performance of the system, we used an independent test set that was collected on different days, time of day, and by different persons than the training set.*

5.2. Multivariate Analysis Results

We start by showing the effect of the wrong independence assumption on the performance of the original *Horus* system. Figure 5 shows the average distance error for different values of N [3] for the multivariate analysis technique. We can see that using more samples can significantly improves performance (average error decreases by about 2 feet from $N = 1$ to $N = 2$). However, as the number of used samples increases, the performance degrades. The minimum value at $N = 3$ can be explained by noting that there are two opposing factors affecting the system accuracy:

1. as the number of the samples used (N) increases, the accuracy of the system should increase.
2. as N increases, the estimation of the multivariate distribution becomes worse due to the wrong independence assumption.

At low values of N the first factor is the dominating factor and hence the accuracy increases. Starting from $N = 4$, the effect of the bad estimation of the distribution becomes the dominating factor and accuracy degrades.

Figure 6 shows the average distance error for different values of α and N. The figure shows that as the value of α, used in calculating the parameters of the distribution of the average of N samples, approaches the true α value (0.9), the system accuracy increases.

Note that at low values of α using samples lead to worse accuracy, as shown in Figure 5, till we reach a switch-over point between $\alpha = 0.4$ and $\alpha = 0.5$ where using more samples starts to give better accuracy. Using the modified technique, the system can achieve an average accuracy of about 1.6 feet, better than the original system by more than 2.95 feet. This represents an accuracy enhancement of more than 64% from the original *Horus* algorithm.

5.3. Comparison with Samples Averaging Technique

Figure 7 compares the performance of the simple samples averaging technique and the multivariate analysis technique. The figure shows that the multivariate technique

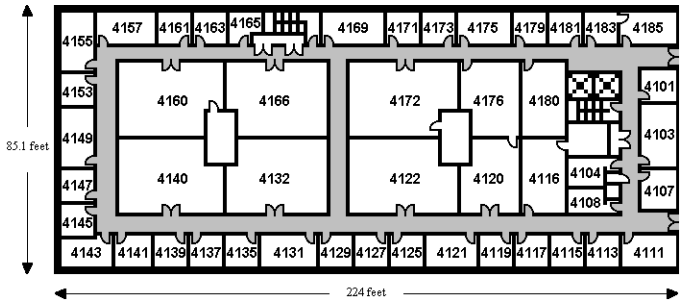

Figure 4. Plan of the south wing of the 4th floor of the Computer Science Department building where the experiment was conducted. Readings were collected in the corridors (shown in gray).

[3] The case of $N = 1$ is equivalent to the original *Horus* system.

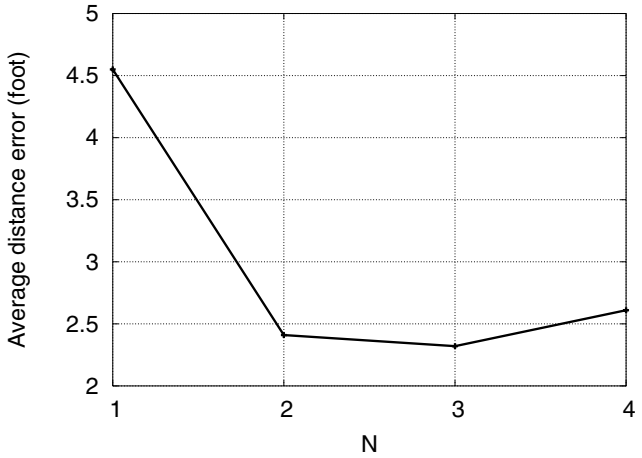

Figure 5. Effect of the wrong independence assumption on the average distance error. As the dimension of the multivariate distribution (number of samples) is increased beyond 3 samples, the average system error increases.

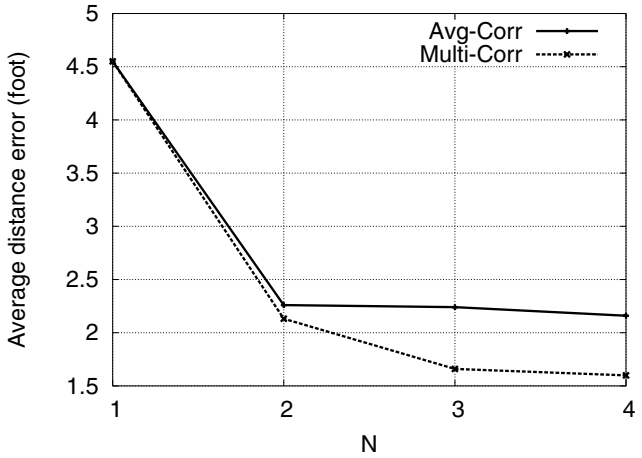

Figure 7. Comparison between the accuracy of the multivariate analysis technique and samples averaging technique. The multivariate analysis technique has a performance advantage of more than 25%

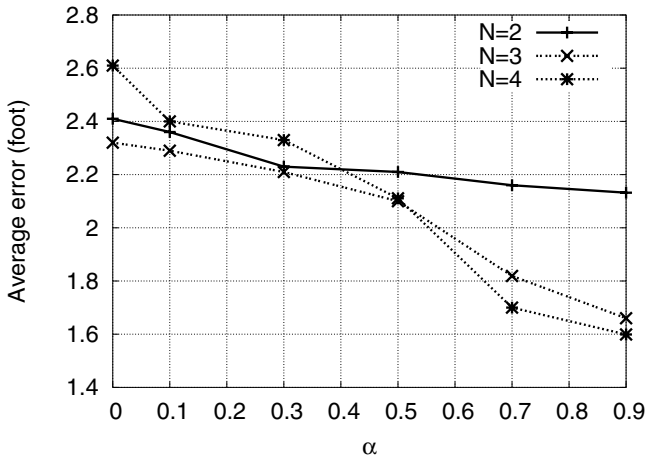

Figure 6. Average distance error for different values of α and N. As the value of α approaches the true value of 0.9, the system performance increases. The case for $N = 1$ (original *Horus* system performance) is shown in Figure 5 for clarity.

gives a performance enhancement of more than 25% over the samples averaging technique. This confirms our previous conclusion in Section 5.2.

6. Related Work

Many systems over the years have tackled the problem of determining and tracking the user position. Examples include GPS [10], wide-area cellular-based systems [23], infrared-based systems [3, 24], ultrasonic-based systems [19], various computer vision systems [15], physical contact systems [17]. WLAN location determination systems provide more ubiquitous coverage and do not require additional hardware for user location determination, thereby enhancing the value of the wireless data network.

In the rest of this Section, we describe techniques that use multiple samples to enhance the performance of WLAN location determination systems. We show how the proposed technique relates to them.

6.1. Physical Location Space Averaging

Different systems, e.g. [4, 5, 16, 20, 21], proposed to use averaging in the *physical-location space*. The system uses a moving time-average of multiple consecutive location estimates to obtain a better location estimate.

Our technique uses multiple samples in the *signal-strength space* to obtain a better location estimate. Moreover our technique can be used in conjunction with the physical-location space averaging to further enhance their accuracy.

6.2. Signal Strength Space Averaging

The authors of the *Radar* [4, 5] system, a *deterministic* location determination technique, were the first to propose using multiple signal strength samples to obtain better estimation accuracy. Their technique is to average the received samples and use the average value in the k-nearest neighborhood algorithm to determine the best location estimate. Their results indicate that using more samples in the averaging process leads to better accuracy.

The work in this paper is concerned with *probabilistic* location determination techniques in which the process of using multiple samples to obtain a location estimate is more involved.

In [27], we discuss how to use the distribution of the average of N correlated signal strength distribution to obtain a better estimate of the user location. In this paper, we extend our previous work to take more information into the estimation process and hence achieve better accuracy. We quantified the advantage of the multivariate analysis technique over the samples averaging technique through analytical analysis and experimental evaluation. The results presented in this paper show that we gain more than 25% performance enhancement when using the multivariate analysis technique over the samples averaging technique.

The proposed technique is unique in using the *multivariate* analysis technique to enhance the accuracy of *probabilistic* location determination systems *taking into account the high correlation degree between samples from the same access point*.

7. Discussion and Conclusions

The main contribution of this paper is three fold: (a) We applied the multivariate analysis technique to the problem of handling high correlation of samples form the same access point (b) We quantified the advantage of using the multivariate analysis technique over samples averaging technique and (c) we analyzed the performance of the proposed technique by implementing it in the context of the *Horus* system and comparing it with previous techniques that use multiple samples to enhance accuracy.

Since samples autocorrelation can be as large as 0.9, it becomes crucial to take this high autocorrelation into account when designing location determination algorithms that uses more than one samples. We presented a technique that uses a linear autoregressive model to estimate the multivariate distribution of N samples from the same access point, taking samples autocorrelation into account. We used the multivariate distribution to enhance the accuracy of the probabilistic WLAN location determination systems by calculating the probability of any sequence of N samples from an access point. The results of testing the proposed technique in the context of the *Horus* WLAN location determination system show that the average distance accuracy is enhanced by more than 2.95 feet (64%).

We also quantified the advantage of using the multivariate analysis technique over the samples averaging technique. Analysis using the information theory framework shows that the amount of information the multivariate technique uses in the estimation process is more than that of the samples averaging technique. This difference in the amount of information increases as the samples autocorrelation increases. This is particularly useful in a typical WLAN environment where the samples are highly correlated. Experimental evaluation shows that the multivariate analysis technique is better than the samples averaging technique by more than 25%.

We believe that the multivariate analysis technique presented in the paper is general and can be applied to other probabilistic WLAN location determination techniques to further enhance their accuracy.

References

[1] http://www.cs.umd.edu/users/moustafa/Downloads.html.

[2] http://www.orinocowireless.com.

[3] R. Azuma. Tracking requirements for augmented reality. *Communications of the ACM*, 36(7), July 1997.

[4] P. Bahl and V. N. Padmanabhan. RADAR: An In-Building RF-based User Location and Tracking System. In *IEEE Infocom 2000*, volume 2, pages 775–784, March 2000.

[5] P. Bahl, V. N. Padmanabhan, and A. Balachandran. Enhancements to the RADAR User Location and Tracking System. Technical Report MSR-TR-00-12, Microsoft Research, February 2000.

[6] P. Castro, P. Chiu, T. Kremenek, and R. Muntz. A Probabilistic Location Service for Wireless Network Environments. *Ubiquitous Computing 2001*, September 2001.

[7] P. Castro and R. Muntz. Managing Context for Smart Spaces. *IEEE Personal Communications*, OCTOBER 2000.

[8] G. Chen and D. Kotz. A Survey of Context-Aware Mobile Computing Research. Technical Report Dartmouth Computer Science Technical Report TR2000-381, 2000.

[9] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Addison Wisley, 1991.

[10] P. Enge and P. Misra. Special issue on GPS: The Global Positioning System. *Proceedings of the IEEE*, pages 3–172, January 1999.

[11] S. Ganu, A.S.Krishnakumar, and P.Krishnan. Infrastructure-based Location Estimation in WLAN Networks. In *IEEE Wireless Communications and Networking Conference*, March 2004.

[12] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1999.

[13] W. Kaplan. *Advanced Calclus*. Addison Wesley, 1984.

[14] P. Krishnan, A. Krishnakumar, W. H. Ju, C. Mallows, and S. Ganu. A System for LEASE: Location Estimation Assisted by Stationary Emitters for Indoor RF Wireless Networks. In *IEEE Infocom*, March 2004.

[15] J. Krumm et al. Multi-camera multi-person tracking for Easy Living. In *3rd IEEE Int'l Workshop on Visual Surveillance*, pages 3–10, Piscataway, NJ, 2000.

[16] A. M. Ladd, K. Bekris, A. Rudys, G. Marceau, L. E. Kavraki, and D. S. Wallach. Robotics-Based Location Sensing using Wireless Ethernet. In *8th ACM MOBICOM*, Atlanta, GA, September 2002.

[17] R. J. Orr and G. D. Abowd. The Smart Floor: A Mechanism for Natural User Identification and Tracking. In *Conference on Human Factors in Computing Systems (CHI 2000)*, pages 1–6, The Hague, Netherlands, April 2000.

[18] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, third edition, 1991.

[19] N. B. Priyantha, A. Chakraborty, and H. Balakrishnan. The Cricket Location-Support system. In *6th ACM MOBICOM*, Boston, MA, August 2000.

[20] T. Roos, P. Myllymaki, and H. Tirri. A Statistical Modeling Approach to Location Estimation. *IEEE Transactions on Mobile Computing*, 1(1):59–69, January-March 2002.

[21] T. Roos, P. Myllymaki, H. Tirri, P. Misikangas, and J. Sievanen. A Probabilistic Approach to WLAN User Location Estimation. *International Journal of Wireless Information Networks*, 9(3), July 2002.

[22] A. Smailagic, D. P. Siewiorek, J. Anhalt, D. Kogan, and Y. Wang. Location Sensing and Privacy in a Context Aware Computing Environment. *Pervasive Computing*, 2001.

[23] S. Tekinay. Special issue on Wireless Geolocation Systems and Services. *IEEE Communications Magazine*, April 1998.

[24] R. Want, A. Hopper, V. Falco, and J. Gibbons. The Active Badge Location System. *ACM Transactions on Information Systems*, 10(1):91–102, January 1992.

[25] M. Youssef and A. Agrawala. On the Optimality of WLAN Location Determination Systems. Technical Report CS-TR 4459, University of Maryland, College Park, March 2003. http://www.cs.umd.edu/Library/TRs/.

[26] M. Youssef and A. Agrawala. Small-Scale Compensation for WLAN Location Determination Systems. In *IEEE WCNC 2003*, March 2003.

[27] M. Youssef and A. Agrawala. Handling Samples Correlation in the Horus System. In *IEEE Infocom*, March 2004.

[28] M. Youssef and A. Agrawala. On the Optimality of WLAN Location Determination Systems. In *Communication Networks and Distributed Systems Modeling and Simulation Conference*, January 2004.

[29] M. Youssef and A. Agrawala. Analysis of the Optimal Strategy for WLAN Location Determination Systems. *International Journal of Modeling and Simulation*, 2005.

[30] M. Youssef and A. Agrawala. Location-Clustering Techniques for Energy-Efficient WLAN Location Determination Systems. *International Journal of Computers and Applications*, 2005.

[31] M. Youssef and A. Agrawala. The Horus WLAN Location Determination System. In *Third International Conference on Mobile Systems, Applications, and Services (MobiSys 2005)*, June 2005.

[32] M. Youssef, A. Agrawala, and A. U. Shankar. WLAN Location Determination via Clustering and Probability Distributions. In *IEEE PerCom 2003*, March 2003.

[33] M. Youssef, A. Agrawala, A. U. Shankar, and S. H. Noh. A Probabilistic Clustering-Based Indoor Location Determination System. Technical Report UMIACS-TR 2002-30 and CS-TR 4350, University of Maryland, College Park, March 2002. http://www.cs.umd.edu/Library/TRs/.