Complexity Reduction in Machine Learning-Based Wireless Positioning: Minimum Description Features

Myeung Suk Oh*, Anindya Bijoy Das*, Taejoon Kim†, David J. Love*, and Christopher G. Brinton*
*Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA
†Electrical Engineering and Computer Science, the University of Kansas, Lawrence, KS, USA
*{oh223, das207, djlove, cgb}@purdue.edu, †taejoonkim@ku.edu

***Abstract*—A recent line of research has been investigating deep learning approaches to wireless positioning (WP). Although these WP algorithms have demonstrated high accuracy and robust performance against diverse channel conditions, they also have a major drawback: they require processing high-dimensional features, which can be prohibitive for mobile applications. In this work, we design a positioning neural network (P-NN) that substantially reduces the complexity of deep learning-based WP through carefully crafted minimum description features. Our feature selection is based on maximum power measurements and their temporal locations to convey information needed to conduct WP. We also develop a novel methodology for adaptively selecting the size of feature space, which optimizes over balancing the expected amount of useful information and classification capability, quantified using information-theoretic measures on the signal bin selection. Numerical results show that P-NN achieves a significant advantage in performance-complexity tradeoff over deep learning baselines that leverage the full power delay profile (PDP).**

***Index Terms*—Convolutional neural network, KL divergence, Minimum description length (MDL), Wireless positioning**

I. Introduction and Related Work

Wireless positioning (WP) systems provide location awareness in many mobile applications, from intelligent vehicles to inventory control. Typically, WP is conducted using a set of sensors that exchange signals with a target device to obtain measurements that are informative for location estimation. Ultra-wideband (UWB) sensors are popularly used for this purpose, as they communicate on a large bandwidth that provides high distance resolution[1] for accurate positioning. UWB is also known to have high signal-to-noise ratio (SNR) and penetration ability, from which more reliable WP can be performed [2].

Existing WP algorithms can be mostly categorized into two classes: geometric methods and fingerprinting methods [3]. With geometric methods, distance-dependent measurements, e.g., received signal strength (RSS) and time of arrival (TOA), are first acquired by the sensors. Then, traditional estimation techniques, e.g., weighted least squares or gradient descent, are applied to predict the target location. However, geometric methods are generally prone to high errors when the channel condition is harsh. For example, TOA measurements under a strong non-line-of-sight (NLOS) condition are often unreliable, and compensation techniques are needed to recover performance [4].

Fingerprinting methods, on the other hand, take a data-driven approach, relying on a pre-acquired set of labeled measurements (i.e., with the location information being available for each measurement). The labeled data can be used either for non-parametric estimation as new measurements arrive, e.g., through nearest-neighbor methods, or for training parametric models, e.g., support vector machines (SVM).

For fingerprinting methods, large dimensional data, e.g., power delay profile (PDP), is often used to achieve robustness against diverse channel conditions, and the models employed to handle large dimensional data are becoming increasingly complex as well. Recently, WP based on deep learning has been considered [5], where neural network (NN) approaches have shown improved performance across different channel conditions and positioning environments, e.g., via convolutional neural networks (CNN) [6] and gated recurrent units (GRU) [7].

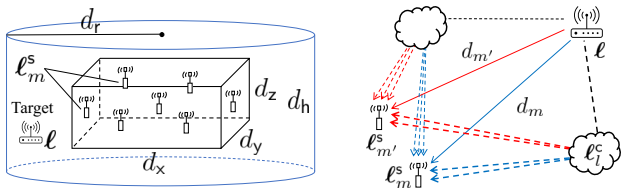
Nevertheless, using PDP as learning features for WP imposes a large bandwidth and/or long latency on a sensor network, as it must be measured and stored for each positioning instance. Also, NNs with high-dimensional input features may require high computational power and associated hardware costs to support real-time positioning rates [8]. These constraints can be undesirable in mobile settings where both latency and cost are critical factors, e.g., consider the requirements of WP in vehicular applications. On the other hand, there exist some works in WP that rely on lower dimensional input features, e.g., the approach in [9] where NNs combined with a linear estimator operate on TOA and RSS measurements. However, their performance is still heavily impacted by channel conditions, and need to be combined with additional learning tasks like ranging error detection [10] anyway. This emphasizes a steep tradeoff between performance/robustness and complexity in WP.

Summary of contributions. In this work, we develop a novel WP technique which reduces complexity compared with PDP-based deep learning without significantly impacting performance. We summarize our contributions as follows:

- We design a positioning neural network (P-NN) that employs the largest power measurements and their temporal locations as its features. Compared to PDP, the feature set has significantly reduced dimensions yet still provides information needed to conduct accurate WP.
- As a component of P-NN, we develop a method for adapting the size of our features to preserve performance based on the channel conditions. Our method adopts principles of model order selection and uses the criterion which we formulate

[1]In the IEEE 802.15.4 standard [1], UWB radio pulses are designed with the maximum duration of 2 ns, which yields a distance resolution of 0.6 m.

978-1-7281-9054-9/24/$31.00 ©2024 IEEE



(a) Layout of sensor and target spaces (b) Channel propagation with clusters

Fig. 1: Illustrations of positioning spaces (left) and channel propagation (right).

based on information-theoretic and classification capability metrics that quantify the impact of varying the number of power measurements used.

- We provide a set of numerical experiments to evaluate P-NN. The results show that our minimum description feature set provides accuracies that approach the PDP-based baseline while using less than 20% of the feature size, thus achieving a desirable performance-complexity tradeoff.

II. System Model

As shown in Fig. 1a, we assume M single-antenna sensors in a rectangular sensor space defined by the length parameters d_{x}, d_{y}, and d_{z}. We denote the location of sensor $m \in \{0, 1, \ldots, M-1\}$ using $\boldsymbol{\ell}^{\mathsf{s}}_m = [x^{\mathsf{s}}_m, y^{\mathsf{s}}_m, z^{\mathsf{s}}_m]^\top$. We are interested in a target that is positioned outside the sensor space but inside a cylindrical target space defined by the radius d_{r} and height d_{h}. We assume both the sensor and target spaces to be centered at $(0, 0, 0)$, and set $d_{\mathsf{h}} > d_{\mathsf{z}}$ and $d_{\mathsf{r}} > \|\frac{d_{\mathsf{x}}}{2} + \frac{d_{\mathsf{y}}}{2}\|_2$ such that the sensor space is always placed inside the target space.

The overall WP procedure is illustrated in Fig. 2. Suppose that a target located at $\boldsymbol{\ell} = [x, y, z]^\top$ transmits a reference signal pulse $s(t)$ that is known to both the target and sensors. The signal received by sensor m is then expressed as [11]

$$r_m(t) = \sum_{l=0}^{L} \sum_{k=0}^{K_l - 1} \alpha_{m,l,k}\, s\Big(t - \frac{d_m}{c} - T_{m,l} - \tau_{m,l,k}\Big) + w_m(t), \quad (1)$$

where L is the number of propagation paths imposed by channel clusters present in the target space (i.e., $l = 0$ refers to the line-of-sight (LOS) path), and K_l is the number of rays existing in each path l. We denote the complex channel gain using $\alpha_{m,l,k} = a_{m,l,k} e^{j\phi_{m,l,k}}$, where $a_{m,l,k}$ is the weight obeying Nakagami-$\mu_{m,l,k}$ distribution with the scale value $\Omega_{m,l,k}$ and $\phi_{m,l,k}$ is the uniformly distributed phase. $w_m(t)$ is the zero-mean complex Gaussian noise with variance σ^2_m.

With d_m denoting the Euclidean distance between the target and sensor m (Fig. 1b) and c being the speed of light, d_m/c represents the TOA of the LOS path. $T_{m,l}$ is the relative delay of path l with respect to the LOS path, which is expressed as

$$T_{m,l} = \begin{cases} 0 & \text{if } l = 0, \text{ (LOS path)} \\ \frac{\|\boldsymbol{\ell}^{\mathsf{c}}_l - \boldsymbol{\ell}\|_2 + \|\boldsymbol{\ell}^{\mathsf{s}}_m - \boldsymbol{\ell}^{\mathsf{c}}_l\|_2 - d_m}{c} & \text{if } l > 0, \end{cases} \quad (2)$$

where $\boldsymbol{\ell}^{\mathsf{c}}_l = [x^{\mathsf{c}}_l, y^{\mathsf{c}}_l, z^{\mathsf{c}}_l]^\top$ is the location of cluster that imposes path $l \in \{1, \ldots, L\}$. $\tau_{m,l,k}$ is the relative delay of ray k with respect to $T_{m,l}$. Hence, $\tau_{m,l,0} = 0$, $\forall m, l$. For $k > 0$, we assume each ray follows Poisson process of the ray arrival rate κ [11].

The pathloss of each path is expressed as [11]

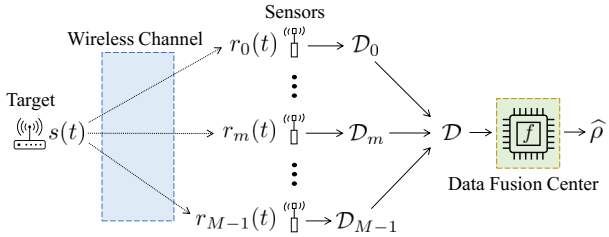


Fig. 2: An overall diagram on wireless positioning.

$$\beta_{m,l,k} = \mathbb{E}[a^2_{m,l,k}] = \overline{P}_m \big(d_m/\overline{d}_m\big)^{-\xi} S^{\mathsf{s}}_m S^{\mathsf{c}}_l\, e^{-\frac{T_{m,l}}{\Gamma} - \frac{\tau_{m,l,k}}{\gamma}}, \quad (3)$$

where $\overline{P}_m$, $\overline{d}_m$, and ξ are respectively the reference power, reference distance, and pathloss exponent. S^{s}_m and S^{c}_l are the random shadowing applied to sensor m and path l, respectively. Γ and γ are respectively the path and ray decaying constants. With (3), each pathloss becomes strongly dependent on the channel propagation distance, which allows the channel paths to convey spatial correlation.

To avoid signal interference, we assume that the signal $s(t)$ is transmitted over a frame of duration T_{f} such that $T_{\mathsf{f}} > \max_{m,l,k}(\frac{d_m}{c} + T_{m,l} + \tau_{m,l,k})$. Upon receiving the frame, each sensor breaks it down to $N_{\mathsf{b}} = \lfloor \frac{T_{\mathsf{f}}}{T_{\mathsf{g}}} \rfloor$ temporal bins, where T_{g} is the integration period. Then, for the signal of bandwidth W, the power contained in each temporal bin $n \in \{0, 1, \ldots, N_{\mathsf{b}} - 1\}$ of sensor m is measured via energy detection [12], [13]

$$\varepsilon_{m,n} = \frac{1}{2W} \sum_{i=0}^{2WT_{\mathsf{g}}-1} \Big| r_m\Big(nT_{\mathsf{g}} + \frac{i}{2W}\Big)\Big|^2, \quad (4)$$

and $\boldsymbol{\varepsilon}_m = [\varepsilon_{m,0}, \varepsilon_{m,1}, \ldots, \varepsilon_{m,N_{\mathsf{b}}-1}]^\top$ becomes the instant PDP vector measured at sensor m.

Each sensor m generates a data set $\mathcal{D}_m$ from $\boldsymbol{\varepsilon}_m$ and transfers it to the data fusion center (DFC). Using the collected set $\mathcal{D} = \{\mathcal{D}_m\}_{m=0}^{M-1}$, the DFC estimates the target location. In this work, we frame our WP as an N_{z}-zone classification task for the following reasons. First, rather than coordinate-level localization, positioning via N_{z} spatial zones is often sufficient in many mobile applications. Note that the value of N_{z} can be adjusted to satisfy the positioning sensitivity. Second, it is more difficult to obtain coordinate-labeled training data than the zone-labeled one. Hence, we define our positioning task using a function $f : \mathcal{D} \to \widehat{\rho}$, where $\widehat{\rho} \in \{0, 1, \ldots, N_{\mathsf{z}} - 1\}$ is the output indicating one of the N_{z} zones. Let $\rho \in \{0, 1, \ldots, N_{\mathsf{z}} - 1\}$ denote the zone in which the target is truly located. Then, the target is correctly positioned if $\widehat{\rho} = \rho$. In Fig. 3, we provide example layouts for $N_{\mathsf{z}} = 8$ and $N_{\mathsf{z}} = 32$, where the zones are created using radius and angle for mobile settings.

III. Positioning Neural Network (P-NN)

A. Features of Minimum Description Length

Many WP algorithms directly use PDP data (i.e., $\mathcal{D} = \{\boldsymbol{\varepsilon}_m\}_{m=0}^{M-1}$) to achieve good performance. Processing such high-dimensional data, however, may increase the operation requirement (e.g., bandwidth, memory, and power). Here, we follow the principle of minimum description length (MDL) [14], which defines that the best model for describing data is one with the smallest size, and propose to use only a small number of the largest power measurements and their temporal locations.

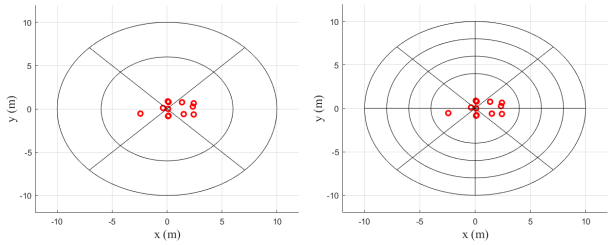


Fig. 3: Zone layouts with $N_{\mathsf{z}} = 8$ (left) and $N_{\mathsf{z}} = 32$ (right). Red circles indicate sensor positions.

Suppose that each sensor m receives the signal $r_m(t)$ and measures the PDP vector $\boldsymbol{\varepsilon}_m$ of size N_{b}. The elements of $\boldsymbol{\varepsilon}_m$ are then sorted to yield $\boldsymbol{\varepsilon}_m^{\text{ord}} = [\varepsilon_{m,0}^{\text{ord}}, \varepsilon_{m,1}^{\text{ord}}, \ldots, \varepsilon_{m,N_{\mathsf{b}}-1}^{\text{ord}}]^\top$, where $\varepsilon_{m,0}^{\text{ord}} \geq \varepsilon_{m,1}^{\text{ord}} \geq \ldots \geq \varepsilon_{m,N_{\mathsf{b}}-1}^{\text{ord}}$. Next, we denote the index vector $\boldsymbol{b}_m^{\text{ord}} = [b_{m,0}^{\text{ord}}, b_{m,1}^{\text{ord}}, \ldots, b_{m,N_{\mathsf{b}}-1}^{\text{ord}}]^\top$, where $b_{m,n}^{\text{ord}}$ is the index of the element $\varepsilon_{m,n}^{\text{ord}}$ in $\boldsymbol{\varepsilon}_m$ (i.e., $b_{m,n}^{\text{ord}}$ indicates the temporal location in $\boldsymbol{\varepsilon}_m$ where the n-th largest power has been measured). The sensor then takes the first F entries of $\boldsymbol{\varepsilon}_m^{\text{ord}}$ and $\boldsymbol{b}_m^{\text{ord}}$ to generate $\mathcal{D}_m = \{\varepsilon_{m,0}^{\text{ord}}, \ldots, \varepsilon_{m,F-1}^{\text{ord}}, b_{m,0}^{\text{ord}}, \ldots, b_{m,F-1}^{\text{ord}}\}$ of size $2F$ and transfers it to the DFC, resulting in a set $\mathcal{D}$ of size $2FM$.

The key motivation for our feature set is an assumption that information needed for accurate WP is more likely present in the temporal bins of the largest powers. Effective TOA estimation algorithms, e.g., [13], [15], are based on this assumption and use the power threshold to detect signals. Since both RSS and TOA of the detected signals become useful information for WP [3], we use both $\boldsymbol{\varepsilon}_m^{\text{ord}}$ and $\boldsymbol{b}_m^{\text{ord}}$ to generate our feature set.

Using PDP is informative as the entire $N_{\mathsf{b}}M$ measurements are perceived as an image for NNs to train and learn. However, if only a small fraction of N_{b} measurements actually convey useful information, it is more desirable to process those measurements only. However, taking the largest powers from N_{b} measurements (i.e., the first F entries of $\boldsymbol{\varepsilon}_m^{\text{ord}}$) can essentially lose information within the time domain. Hence, we directly include the temporal information (i.e., the first F entries of $\boldsymbol{b}_m^{\text{ord}}$) into our feature set.

Compared to having a PDP of size N_{b}, using our feature set reduces the dimension by a factor of $\frac{2F}{N_{\mathsf{b}}}$ (e.g., $F = 5$ and $N_{\mathsf{b}} = 100$ yield the size reduction by $\frac{1}{10}$). Since deep learning algorithms (e.g., CNN of per-layer complexity that quadratically increases with feature dimensions [16]) typically involve large data to be stored, transferred, and/or processed, reduction in feature dimensions can result in benefits such as less storage, smaller bandwidth, and lower computational complexity.

B. Network Architecture and Operation

The overall architecture of our P-NN is illustrated in Fig. 4. From the collected data $\mathcal{D}$, we separate the power and time measurements, normalize them (using mean and standard deviation) [17], and generate two $M \times F$ matrices $\mathbf{E} = [\varepsilon_{0,0}^{\text{ord}}, \ldots, \varepsilon_{0,F-1}^{\text{ord}}; \ldots; \varepsilon_{M-1,0}^{\text{ord}}, \ldots, \varepsilon_{M-1,F-1}^{\text{ord}}]$ and $\mathbf{B} = [b_{0,0}^{\text{ord}}, \ldots, b_{0,F-1}^{\text{ord}}; \ldots; b_{M-1,0}^{\text{ord}}, \ldots, b_{M-1,F-1}^{\text{ord}}]$. We feed each $\mathbf{E}$ and $\mathbf{B}$ into a separate NN first to handle the data obtained from two different domains. Here we use two convolutional layers with rectified linear unit (ReLU) activation to capture spatial correlation across both the measurements and sensors. The outputs of two separate networks are then flattened and concatenated to be fed to a set of two fully connected (FC) layers with ReLU activation. The last layer is designed with N_{z} neurons and softmax activation to output a classification vector that is directly translated to $\widehat{\rho}$. The latter set of FC layers is to combine the information separately extracted from $\mathbf{E}$ and $\mathbf{B}$ and determine the output for our zone-based positioning task.

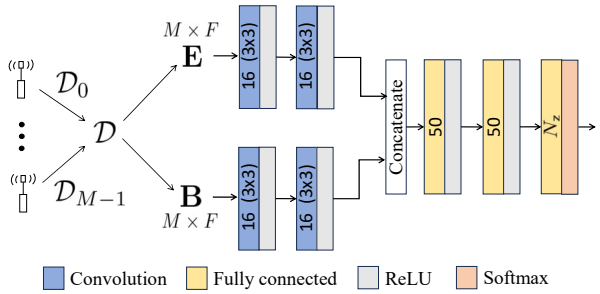


Fig. 4: Architecture of our positioning neural network (P-NN).

To train our P-NN, we pre-acquire a training set of size D, where each data point indexed by $i \in \{0, 1, \ldots, D-1\}$ consists of the set $\mathcal{D}^{(i)} = \{\mathcal{D}_m^{(i)}\}_{m=0}^{M-1}$ and zone index ρ_i for its label. The network is trained offline via Adam optimizer. During the testing phase, the feature set $\mathcal{D}$ is obtained from the sensors in real-time and forward-fed through the NN to determine the positioning outcome $\widehat{\rho}$.

IV. Adaptive Feature Size Selection

As discussed in Sec. III-A, the F largest powers and their temporal locations are collected from each sensor to form our feature set of size $2FM$. Here we develop a strategy to adaptively select the value of F as the number of measurements to be taken by each sensor for accurate WP varies by channel conditions. To determine the value of F, we adopt the principle of model order selection [18] and develop a unique feature size selection method. In the following, we define three metrics that are used to evaluate the effectiveness of our feature set.

1) Information coming from F signal bins: Note that taking the F largest power measurements for our feature set can be seen as assuming F out of N_{b} bins to contain the signal. Since each sensor measures the power according to (4), these F signal-contained bins are assumed to follow non-central chi-square distribution [12], which we approximate using central chi-square distribution of probability density function (PDF) given as [15]

$$f(x; \psi^2, \lambda, \nu) = \left(\frac{1}{2\eta^2}\right)^{\frac{\nu}{2}} \frac{x^{\frac{\nu}{2}-1}}{\Gamma(\frac{\nu}{2})} \exp\left(-\frac{x}{2\eta^2}\right), \qquad (5)$$

where $\eta^2 = \sqrt{\frac{2\nu\psi^4 + 4\psi^2\lambda + (\nu\psi^2+\lambda)^2}{\nu(2+\nu)}}$ with ψ^2, λ, and ν being the non-central chi-square parameters and $\Gamma(\cdot)$ is the Gamma function. The rest $N_{\mathsf{b}} - F$ noise-only bins are assumed to follow central chi-square distribution [12] of the PDF given as

$$f(x; \psi^2, \nu) = \left(\frac{1}{2\psi^2}\right)^{\frac{\nu}{2}} \frac{x^{\frac{\nu}{2}-1}}{\Gamma(\frac{\nu}{2})} \exp\left(-\frac{x}{2\psi^2}\right). \qquad (6)$$

Note that, with $\lambda = 0$, only ψ^2 and ν characterize (6).

Using multiple measurements of $\boldsymbol{\varepsilon}_m^{\text{ord}}$ from each sensor as samples, we can compute $\overline{\boldsymbol{\varepsilon}}^{\text{ord}} = [\overline{\varepsilon}_0^{\text{ord}}, \overline{\varepsilon}_1^{\text{ord}}, \ldots, \overline{\varepsilon}_{N_{\mathsf{b}}-1}^{\text{ord}}]^\top$, where $\overline{\varepsilon}_n^{\text{ord}}$ is the power of the n-th largest temporal bin averaged over

both the sensors and measurements. Using (5) and (6), we define the joint PDF of F non-central and $N_{\mathsf{b}} - F$ central chi-square variables and derive the likelihood of having $\overline{\varepsilon}^{\text{ord}}$ as [15]

$$\begin{aligned}&\ln f(\bar{\varepsilon}^{\text{ord}}; \psi_0^2, \ldots, \psi_{N_{\mathsf{b}}-1}^2, \lambda_0, \ldots, \lambda_{F-1}, \nu)\\ &= \sum_{n=0}^{F-1} -\frac{\nu}{2}\ln(2\eta_n^2) + \frac{\nu-2}{2}\ln(\bar{\varepsilon}_n^{\text{ord}}) - \ln\Gamma\Big(\frac{\nu}{2}\Big) - \frac{\bar{\varepsilon}_n^{\text{ord}}}{2\eta_n^2}\\ &+ \sum_{n=F}^{N_{\mathsf{b}}-1} -\frac{\nu}{2}\ln(2\psi_n^2) + \frac{\nu-2}{2}\ln(\bar{\varepsilon}_n^{\text{ord}}) - \ln\Gamma\Big(\frac{\nu}{2}\Big) - \frac{\bar{\varepsilon}_n^{\text{ord}}}{2\psi_n^2}. \quad (7)\end{aligned}$$

Note that (7) is characterized by N_{b} values of ψ_n^2, F values of λ_n, and a single value of $\nu = 2WT_{\mathsf{g}}$. Since we do not have the knowledge of $\{\psi_n^2\}_{n=0}^{N_{\mathsf{b}}-1}$ and $\{\lambda_n\}_{n=0}^{F-1}$ to evaluate (7), we estimate each term using

$$\psi_F^2 = \frac{1}{N_{\mathsf{b}} - F}\sum_{n=F}^{N_{\mathsf{b}}-1} \bar{\varepsilon}_n^{\text{ord}} \approx \psi_n^2, \;\; \forall n = 0, \ldots, N_{\mathsf{b}} - 1, \quad (8)$$

$$\lambda_n^{(F)} = \bar{\varepsilon}_n^{\text{ord}} - \psi_F^2 \approx \lambda_n, \;\; \forall n = 0, \ldots, F-1. \quad (9)$$

Using (8) and (9), we now define the estimated likelihood of having $\overline{\varepsilon}^{\text{ord}}$ when the F largest powers are taken for our feature set (i.e., F bins are assumed to contain signals) as

$$\mathsf{LL}_F = \ln f(\bar{\varepsilon}^{\text{ord}}; \psi_F^2, \ldots, \psi_F^2, \lambda_0^{(F)}, \ldots, \lambda_{F-1}^{(F)}, \nu). \quad (10)$$

For a given $\overline{\varepsilon}^{\text{ord}}$, the value of (10) varies by F, and we utilize this metric to evaluate the expected amount of information when F measurements are taken for our feature set. Note that the log-likelihood is an effective metric popularly used for the information theoretic model order selection [14], [15], [18].

2) Information acquisition probability: Another metric we define is the probability of acquiring the useful information when we consider the F largest power measurements. Due to the time-varying nature of wireless channels, the power across the N_{b} temporal bins are randomly measured at each positioning instance. In other words, despite the effort to generate our feature set using only the signal-contained bins, it is possible for the set to include measurements from the noise-only bins. Such a case is not desirable since data with no useful information can degrade the performance of our P-NN.

Thus, for a given value of F, we quantify the chance of our feature set to take measurements from the signal-contained bins. Recall that taking the F largest power measurements is to assume F signal-contained bins out of N_{b}. First, we define $P_{\mathsf{th}}^{(F)} = (\bar{\varepsilon}_{F-1}^{\text{ord}} + \bar{\varepsilon}_F^{\text{ord}})/2$ be the power threshold that separates the first F bins from the rest $N_{\mathsf{b}} - F$ bins. Our logic is that the feature set will likely include these signal-contained bins if their power is measured greater than $P_{\mathsf{th}}^{(F)}$. Hence, using (8) and (9), we define the probability of a signal-contained bin $n \in \{0, \ldots, F-1\}$ to have the power greater than $P_{\mathsf{th}}^{(F)}$ as [13]

$$\begin{aligned}p_n^{(F)} &= \mathbb{P}\left\{ \frac{\varepsilon_n^{\text{ord}}}{\psi_F^2} > \frac{P_{\mathsf{th}}^{(F)}}{\psi_F^2} \;\middle|\; \frac{\lambda_n^{(F)}}{\psi_F^2} \right\}\\ &= Q_{\frac{\nu}{2}}\left(\sqrt{2(\lambda_n^{(F)}/\psi_F^2)^2}, \sqrt{2P_{\mathsf{th}}^{(F)}/\psi_F^2} \right), \quad (11)\end{aligned}$$

where $Q_{\frac{\nu}{2}}(\cdot,\cdot)$ is the $\frac{\nu}{2}$-th order Marcum Q-function [19]. Based on (11), we define the acquisition probability of our F largest powers to include the measurements from $f \in \{0, 1, \ldots, F\}$ signal-contained bins as

$$\mathsf{P}_f^{(F)} = \textstyle\sum_{\boldsymbol{q} \in \mathcal{Q}_f^{(F)}} \prod_{i=1}^{F} (p_{i-1}^{(F)})^{\boldsymbol{q}[i]} (1 - p_{i-1}^{(F)})^{(1-\boldsymbol{q}[i])}, \quad (12)$$

where $\mathcal{Q}_f^{(F)}$ is the set of all F-length binary vectors containing f ones (i.e., $\mathcal{Q}_f^{(F)}$ considers all $\frac{F!}{f!(F-f)!}$ cases where f out of F bins have their power greater than $P_{\mathsf{th}}^{(F)}$). The product term in (12) computes the joint probability of each case in $\mathcal{Q}_f^{(F)}$, and the summation provides the overall probability. Note that (12) quantifies the chance of taking f useful measurements when we consider the F largest measurements for our feature set.

3) Inter-zone Kullback-Leibler divergence: Dissimilarity among the class distributions is one of the key factors that impact classification performance, and how we form our feature set directly affects this dissimilarity. Hence, for a given value of F, we propose to quantify the dissimilarity across the data samples from each zone via Kullback-Leibler (KL) divergence and use it for our feature size selection. To evaluate KL divergence, the PDFs must be known. Since we only have empirical measurements (i.e., training data), we take the k-nearest neighbors (KNN) density estimation approach to directly estimate the KL divergence [20]. If we subgroup the training data by each zone in terms of our feature set and denote each group using $\mathcal{D}_z^{\mathsf{z}}$ for $z \in \{0, 1, \ldots, N_{\mathsf{z}} - 1\}$, the estimated KL divergence between the zone z and z' using the KNN density estimation with u nearest neighbors is given by

$$\widehat{D}_u(P_z || P_{z'}) = \frac{F}{|\mathcal{D}_z^{\mathsf{z}}|}\sum_{x \in \mathcal{D}_z^{\mathsf{z}}} \log\frac{r_{u,z'}(x)}{r_{u,z}(x)} + \log\frac{|\mathcal{D}_{z'}^{\mathsf{z}}|}{|\mathcal{D}_z^{\mathsf{z}}| - 1}, \quad (13)$$

where $r_{u,z}(x)$ is the Euclidean distance between x and its u-th nearest neighbor in $\mathcal{D}_z^{\mathsf{z}}$. Now we define the average KL divergence upon taking the F largest power measurements as

$$\mathsf{KL}_F = \tfrac{1}{N_{\mathsf{z}}^2\sqrt{F}} \textstyle\sum_{i=0}^{N_{\mathsf{z}}} \sum_{j=0}^{N_{\mathsf{z}}} \widehat{D}_u(P_i || P_j), \quad (14)$$

which we use to quantify how effectively our feature set of size $2FM$ can separate the classes. Note that, regardless of the distributions being compared, (13) yields a steady increase with F due to the volume expression used in the KNN density estimation. Hence, a factor of $\sqrt{F}$ is applied in (14) to account for the increase in the expected Euclidean distance across F.

Using the metrics (10), (12), and (14), we now formulate our feature size selection criterion, which is expressed as

$$F^\star = \underset{F \in [F_{\min}, F_{\max}]}{\arg\max} \Bigg(\underbrace{\epsilon \sum_{f=0}^{F} \mathsf{P}_f^{(F)} \frac{f}{F} \overline{\mathsf{LL}_F - \mathsf{LL}_0}}_{(a)} + (1-\epsilon) \underbrace{\overline{\mathsf{KL}_F}}_{(b)} \Bigg) \quad (15)$$

where $\overline{(\cdot)}$ implies the normalization with respect to $\max_F(\cdot)$ and $\epsilon \in [0, 1]$ is the weight parameter. Since our criterion is the weighted sum of (a) and (b), we force the range of both (a) and (b) to be $[0, 1]$ by normalizing $\{\mathsf{LL}_F - \mathsf{LL}_0\}_{F=F_{\min}}^{F_{\max}}$ and $\{\mathsf{KL}_F\}_{F=F_{\min}}^{F_{\max}}$. Our selection criterion in (15) reflects two factors: the effective amount of information, i.e., (a), and classification capability, i.e., (b), from taking the F largest powers and their temporal locations. Note that we compute $\mathsf{LL}_F - \mathsf{LL}_0$ to exactly quantify the increase in information upon taking the F

TABLE I: Numerical values of the key parameters used in our feature size selection steps. ψ_F^2, $\lambda_n^{(F)}$, and $P_{\mathsf{th}}^{(F)}$ are in the unit of 10^{-7}.

F	4	5	6
ψ_F^2	5.23	4.39	3.89
$\{\lambda_n^{(F)}\}_{n=2}^{F-1}$	12.13, 7.27	12.97, 8.11, 5.07	13.46, 8.60, 5.56, 2.45
$\mathsf{LL}_F - \mathsf{LL}_0$	4.651	5.099	5.326
$P_{\mathsf{th}}^{(F)}$	10.98	7.91	5.79
$\{p_n^{(F)}\}_{n=2}^{F-1}$	0.92, 0.57	0.99, 0.83, 0.52	0.99, 0.95, 0.72, 0.35
$\{\mathsf{P}_f^{(F)}\}_{f=3}^{F}$	0.44, 0.53	0.09, 0.49, 0.42	0.01, 0.20, 0.55, 0.24
(a) in (15)	0.7213	0.7857	0.7912
KL_F	16.48	16.64	16.28

largest measurements. To account for the chance that our F measurements include only f ones that are actually useful, we multiply $\frac{f}{F}$ and the acquisition probability $\mathsf{P}_f^{(F)}$ to $\overline{\mathsf{LL}_F - \mathsf{LL}_0}$.

Example. We provide a numerical example of our feature size selection using the setting of 15dB SNR and LOS condition. For brevity, we set $N_{\mathsf{b}} = 10$, $[F_{\mathsf{min}}, F_{\mathsf{max}}] = [3, 8]$, and $\nu = 2$. From the given setting, we obtain $\bar{\varepsilon}^{\mathrm{ord}} = [53.9, 26.8, 17.4, 12.5, 9.46, 6.35, 5.22, 4.06, 3.76, 2.55] \times 10^{-7}$, where the first five entries contain the signal. In Table I, we provide some of the numerical values computed for the given example with $\epsilon = 0.5$. Using the last two rows of Table I, we evaluate our criterion values for $F \in [3, 8]$ to be $\{0.79, 0.76, 0.89, 0.88, 0.87, 0.85\}$ and determine $F^\star = 5$ based on (15). As shown in Table I, the given $\bar{\varepsilon}^{\mathrm{ord}}$ provides a steady increase in $\mathsf{LL}_F - \mathsf{LL}_0$ from $F = 4$ to $F = 6$. However, a larger F also increases the chance of taking measurements from the noise-only temporal bins, which contributes to the decrease in $p_n^{(F)}$ and results in a negligible increase in the effective amount of information. Hence, our selection criterion determines $F^\star = 5$ to be the number of measurements to be taken for our features.

V. Numerical Evaluation

We conduct a set of numerical experiments to evaluate our P-NN. We consider a rectangular sensor space of $d_{\mathsf{x}} = 6$ m, $d_{\mathsf{y}} = 3$ m, and $d_{\mathsf{z}} = 2$ m with $M = 12$ sensors and a cylindrical target space of $d_{\mathsf{r}} = 10$ m and $d_{\mathsf{h}} = 4$ m. We consider the residential model for UWB channel [11], for which we generate L randomly located channel clusters using Poisson distribution of mean $\overline{L} = 3$ and set $K_l = 6$ for all l. For Nakagami distributions, we assume $\mu_{m,l,k}$ follows log-normal distribution of mean 0.67 dB and variance 0.28 dB and $\Omega_{m,l,k} = \beta_{m,l,k}$, $\forall m, l, k$ [11]. We set $\kappa = 1.5$ ns, $\Gamma = 25$ ns, $\gamma = 5$ ns, and $\xi = 2$ and consider $\overline{P}_m = -45$ dBm and $\overline{d}_m = 1$ m for all sensors [11]. We assume both S_m^{s} and S_l^{c} follow zero-mean log-normal distribution with 3 dB variance [11]. We assume $W = 2$ GHz, $T_{\mathsf{f}} = 200$ ns, and $T_{\mathsf{g}} = 2$ ns to have $N_{\mathsf{b}} = 100$. For each sensor m, we define SNR as $\mathbb{E}[\beta_{m,0,0}]/\sigma_m^2$, where the expectation is over the target space. For NLOS channel, we set $\alpha_{m,0,k} = 0$ for all m and k to remove the LOS path. For the KL divergence estimation, we use $u = 30$.

To compare our P-NN with existing algorithms, we consider CNN-LE [6] and NN-LCS [9] as the baseline. Note that CNN-LE and NN-LCS respectively use PDP and TOA/RSS as their features. For the training phase, $D = 30,000$ target locations were randomly generated, and a pair of $\mathcal{D}^{(i)}$ and ρ_i was obtained for each location. We used Adam optimizer of learning rate 0.001, and the training was performed over 50 epochs with the random batch size 256. For the testing phase, a set of $\mathcal{D}$ and ρ pairs were generated from $6,000$ random target locations, and the performance of each WP algorithm was measured by comparing each predicted output $\widehat{\rho}$ with ρ. An visual illustration of our training and testing sets is provided in Fig. 5. For statistical significance, the result was averaged over 20 simulation runs, and five independent scenarios were used.

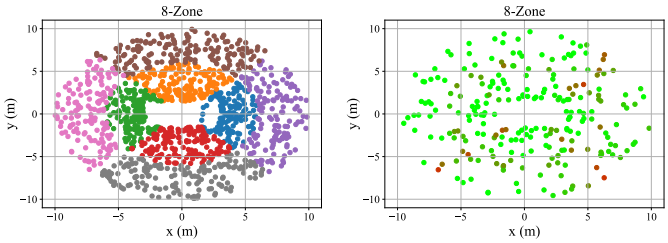


Fig. 5: An illustration of training (left) and testing (right) sets in a 2D plane. For the training set, same color implies the same classification zone. For the testing set, redder color indicates lower classification accuracy.

TABLE II: Zone classification rates (in percent) of P-NN with different values of F. The rates achieved using $F^\star$ in (15) are indicated in bold. We set $\epsilon = 0.3$(or 0.5) for the LOS(or NLOS) channel scenarios.

Scenario #	SNR	F = 4	F = 5	F = 6	F = 7	F = 8	F = 9	F=10
LOS #2	20dB	91.31	**92.24**	91.62	91.52	91.21	91.09	90.69
LOS #4		**89.07**	89.97	89.28	89.39	88.71	87.76	87.33
NLOS #2		87.09	87.33	87.62	**87.91**	88.22	87.81	87.71
NLOS #4		75.16	76.76	77.46	79.54	79.82	**80.28**	80.58
LOS #2	5dB	**62.82**	62.11	62.14	61.02	60.44	60.10	59.65
LOS #4		61.00	**60.20**	59.18	58.34	57.90	58.45	57.66
NLOS #2		**46.80**	45.06	44.25	43.17	43.25	42.80	42.81
NLOS #4		**48.52**	46.98	46.59	46.32	45.77	44.70	44.95

Feature size selection: We demonstrate the effectiveness of our feature size selection method given in Sec. IV. In Table II, we provide the performance (in zone classification rate) of our P-NN using different values of F over various channel conditions. For each row, the numerical value in bold indicates the performance obtained using $F^\star$ from our method. We observe that training our P-NN with $F^\star$ can maintain high performance with a relatively lower feature size. This verifies that taking the largest power and time measurements constitutes minimum description features for navigating the performance-complexity tradeoff. Overall, our feature size selection can adaptively determine the dimensions of our features and lead to high WP performance.

Classification performance: Next, we compare the performance of P-NN with the baselines. In Figs. 6 and 7, we provide classification rate vs. SNR plots for LOS and NLOS channels, respectively. For P-NN, we determine $F^\star$ from a range $[4, 10]$. We observe that the performance of NN-LCS in both plots is significantly lower, demonstrating the difficulty of achieving good WP performance from a small-sized feature. Compared to NN-LCS, both CNN-LE and P-NN provide better performance. Especially in low SNR, P-NN outperforms CNN-LE as it discards the measurements from noise-only bins, the power of which become greater with low SNR, and thus prevents them from being used in the NN training. Given that the performance is competitive between CNN-LE and P-NN (i.e.,

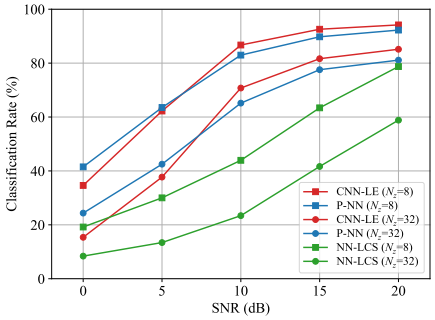


Fig. 6: Performance vs. SNR of different WP algorithms with LOS channels. Feature sizes for CNN-LE and NN-LCS are 1200 and 24, respectively. Feature size for the proposed ranges from 72 to 240.

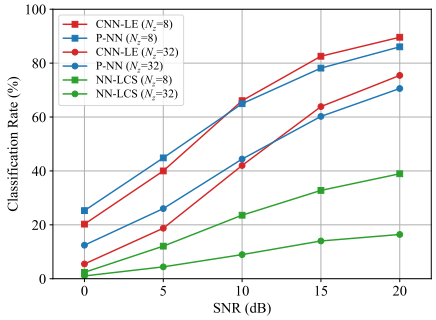


Fig. 7: Performance vs. SNR of different WP algorithms with NLOS channels. Feature sizes for CNN-LE and NN-LCS are 1200 and 24, respectively. Feature size for the proposed ranges from 72 to 240.

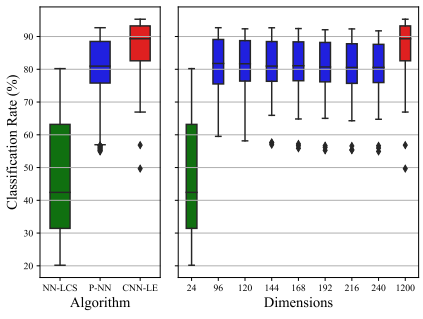


Fig. 8: Classification rates obtained with 10, 15, and 20 dB SNRs by different WP algorithms (left) and the number of dimensions (right). For P-NN, we consider $F \in [4, 10]$.

one outperforms the other depending on the SNR level), our P-NN, which takes only the largest measurements from PDP, takes an advantage in the performance-complexity tradeoff.

Performance-complexity tradeoff: To demonstrate the advantage of our P-NN in the performance-complexity tradeoff, we provide box plots showing the range of classification rates obtained by different WP algorithms and the number of feature dimensions in Fig. 8. We observe that NN-LCS has the lowest dimension, but the performance range is low and has a high variance. CNN-LE exhibits steady and high classification rate, but such a performance is achieved at the cost of utilizing high-dimensional feature. P-NN using our feature set shows the performance similar to the one of CNN-LE at relatively low feature dimensions. This result shows that our feature set can provide WP performance that is much more complexity-efficient.

VI. Conclusions

We have considered a WP scenario for mobile applications and proposed P-NN that utilizes a low-dimensional feature. Our minimum description feature set takes a number of largest power measurements and their temporal positions. For robust performance against varying channel conditions, we have proposed a method of adaptively selecting the feature size by considering the log-likelihood, acquisition probability, and KL divergence. Numerical results have shown that using our feature set achieves positioning performance competitive to the one from using PDP and has a great performance-complexity tradeoff as compared to the baseline algorithms. Potential directions of future works include (i) theoretical optimization on the weight parameter for the feature size selection and (ii) the extension of P-NN to consider coordinates estimation.

Acknowledgment

This work was supported in part by the National Science Foundation (NSF) under grants EEC1941529, CNS2146171, CNS2212565, CNS2225577, and CNS2225578.

References

[1] "IEEE standard for low-rate wireless networks–amendment 1: Enhanced ultra wideband (UWB) physical layers (PHYs) and associated ranging techniques," *IEEE Std 802.15.4z-2020 (Amendment to IEEE Std 802.15.4-2020)*, pp. 1–174, 2020.

[2] S. Gezici and H. V. Poor, "Position estimation via ultra-wide-band signals," *Proc. IEEE*, vol. 97, no. 2, pp. 386–403, 2009.

[3] F. Mazhar, M. Khan, and B. Sällberg, "Precise indoor positioning using UWB: A review of methods, algorithms and implementations," *Wireless Pers. Commun.*, vol. 97, 12 2017.

[4] I. Guvenc and C.-C. Chong, "A survey on TOA based wireless localization and NLOS mitigation techniques," *IEEE Commun. Surv. & Tut.*, vol. 11, no. 3, pp. 107–124, 2009.

[5] J. Fayyad, M. A. Jaradat, D. Gruyer, and H. Najjaran, "Deep learning sensor fusion for autonomous vehicle perception and localization: A review," *Sensors*, vol. 20, no. 15, 2020.

[6] D. T. A. Nguyen, H.-G. Lee, E.-R. Jeong, H. L. Lee, and J. Joung, "Deep learning-based localization for UWB systems," *Electronics*, vol. 9, no. 10, 2020.

[7] D. T. A. Nguyen, J. Joung, and X. Kang, "Deep gated recurrent unit-based 3D localization for UWB systems," *IEEE Access*, vol. 9, pp. 68 798–68 813, 2021.

[8] J. Fontaine, M. Ridolfi, B. Van Herbruggen, A. Shahid, and E. De Poorter, "Edge inference for UWB ranging error correction using autoencoders," *IEEE Access*, vol. 8, pp. 139 143–139 155, 2020.

[9] Z. Zheng, S. Yan, L. Sun, H. Shu, and X. Zhou, "NN-LCS: Neural network and linear coordinate solver fusion method for UWB localization in car keyless entry system," *Sensors*, vol. 23, no. 5, 2023.

[10] D.-H. Kim, A. Farhad, and J.-Y. Pyun, "UWB positioning system based on LSTM classification with mitigated NLOS effects," *IEEE Internet of Things J.*, vol. 10, no. 2, pp. 1822–1835, 2023.

[11] A. F. Molisch, K. Balakrishnan, D. Cassioli, C.-C. Chong, S. Emami, A. Fort, Johan, Karedal, J. Kunisch, H. G. Schantz, U. G. Schuster, and K. Siwiak, "IEEE 802.15.4a channel model-final report," 2004.

[12] H. Urkowitz, "Energy detection of unknown deterministic signals," *Proc. IEEE*, vol. 55, no. 4, pp. 523–531, 1967.

[13] D. Dardari, C.-C. Chong, and M. Win, "Threshold-based time-of-arrival estimators in UWB dense multipath channels," *IEEE Trans. Commun.*, vol. 56, no. 8, pp. 1366–1378, 2008.

[14] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 387–392, 1985.

[15] A. Giorgetti and M. Chiani, "Time-of-arrival estimation based on information theoretic criteria," *IEEE Trans. Signal Process.*, vol. 61, no. 8, pp. 1869–1879, 2013.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[17] A. Ghasempour and M. Martínez-Ramón, "Electric load forecasting using multiple output gaussian processes and multiple kernel learning," in *IEEE Symp. Ind. Electron. & Appl. (ISIEA)*, 2023, pp. 1–6.

[18] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. 19, no. 6, pp. 716–723, 1974.

[19] J. Marcum, "A statistical theory of target detection by pulsed radar," *IRE Trans. Inf. Theory*, vol. 6, no. 2, pp. 59–267, 1960.

[20] F. Perez-Cruz, "Kullback-Leibler divergence estimation of continuous distributions," in *IEEE Int. Symp. Inf. Theory*, 2008, pp. 1666–1670.