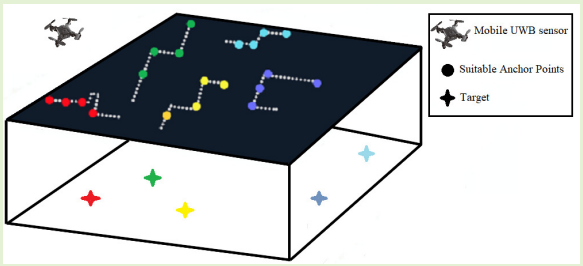
Improving Indoor Localization Using Mobile UWB Sensor and Deep Reinforcement Learning

Leyla Nosrati, Samaneh Hoseini Semnani, Mohammad Sadegh Fazel, Sajed Rakhshani and Mohammad Ghavami

***Abstract*— Indoor localization has posed a significant challenge for researchers. Current methodologies predominantly rely on ultra-wideband (UWB) technology and selection of an appropriate number of anchor points to achieve precise accuracy at the centimeter level. However, the efficacy of these approaches can be compromised when anchor points are incorrectly positioned due to factors such as multi-path effects. Such misalignment can lead to signal power attenuation, thereby diminishing the overall accuracy of localization. In this paper, we propose a novel solution to address this issue. Our approach involves the utilization of deep reinforcement learning (DRL) to train a mobile UWB sensor in the identification of suitable anchor points. By leveraging DRL, we aim to mitigate the loss of transmitted signal power associated with unsuitable anchor placement. Subsequently, we conduct an evaluation to compare the performance of intelligently selected anchor points against two alternative strategies: anchor points selected with predefined constant positions and those chosen randomly. We employ the convolutional neural network (CNN) algorithm for this comparative analysis. Specifically, we utilize the received UWB signal time vector as input and predict the 2D target position using a CNN regressor to estimate the target location. Our simulation results demonstrate a significant improvement in localization accuracy when employing the DRL approach for anchor point selection. Specifically, the mean absolute error (MAE) achieved is approximately 0.09 m which represents a significant improvement compared to manual or random selection of anchor points, which provide MAEs of about 0.45 m and 1.20 m, respectively.**



***Index Terms*— Ultra-wideband, Indoor localization, Machine learning, Mobile sensor, Power attenuation**

I. Introduction

INDOOR target localization holds immense significance across diverse sectors such as healthcare, energy management, and search and rescue operations. While the Global Positioning System (GPS) efficiently determines outdoor locations, its functionality is severely limited indoors due to signal attenuation by obstacles like walls. To overcome this challenge, technologies like Wi-Fi and ultra-wideband (UWB) are employed. Wi-Fi, standardized under IEEE 802.11, operates in industrial, scientific, and medical applications and is used to provide networking capabilities and indoor localization [1]. Conversely, UWB is a wireless technology [2], designed for high-speed data transfer over short distances (less than 20 meters), wide bandwidth, the ability to penetrate walls with minimal power spectral densities, approximately -41.3 dBm/MHz, and offers precision within centimeters for indoor localization [3]. Localization techniques typically utilize anchor points, to triangulate the target's location by measuring distances. However, the efficacy of these techniques depends on the strategic placement of anchor points. Issues such as multi-path effects or non-line-of-sight (NLOS) propagation can compromise accuracy if anchor points are improperly positioned.

In this study, we propose a novel approach utilizing deep reinforcement learning (DRL) algorithm to train a mobile UWB sensor (i.e., the agent) to identify suitable anchor points. This algorithm allows agent to optimize behavior through trial-and-error interactions with the environment. Here, if an action leads to a satisfactory state or better than the previous state of the agent, the tendency to produce that action is reinforced, otherwise the agent is punished. Additionally, the proposed DRL algorithm processes high-dimensional inputs, such as the sensor's current position and the average received power along its path. After identifying suitable anchor points, the UWB signal time vector is collected at these positions. Subsequently, a convolutional neural network (CNN) utilizes the received UWB signal time vector to estimate 2D target position. This methodology facilitates intelligent anchor point selection, effectively tackling the challenges of indoor localization.

L. Nosrati, S. Hoseini Semnani, and M. S. Fazel are with the Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, 84156-83111, Iran (e-mail: l.nosrati@alumni.iut.ac.ir; samaneh.hoseini@iut.ac.ir; fazel@iut.ac.ir).

S. Rakhshani is with the Medical Image and Signal Processing Research Center, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, 81746-73461, Iran (e-mail: sajedrakhshani@msn.com).

M. Ghavami is with the South Bank Applied BioEngineering Research (SABER), School of Engineering, London South Bank University, 103 Borough Road, London, SE1 0AA, U.K (e-mail: ghavamim@lsbu.ac.uk).

II. Related Work

In previous similar investigations, authors often use Wi-Fi and UWB technologies and machine learning based techniques for indoor localization [4]–[8]. In [4], the authors introduced a feature-based localization method employing a Deep Long Short-Term Memory (DLSTM) algorithm for UWB localization. Their findings indicated that by utilizing extracted features from user distance and applying the DLSTM algorithm, they were able to achieve a mean localization error of 0.05 m. In [5], the authors proposed Deep Autoencoder-Backpropagation (DAE-BP) algorithm leveraging the Time Difference of Arrival (TDOA) value extracted from received UWB signals. They demonstrated that the DAE-BP algorithm achieved a Mean Square Error (MSE) of 0.03 m. In [6], the authors improved the performance of their indoor localization system by incorporating predicted NLOS conditions and ranging error information. They combined this information with Weighted Least Squares (WLS) location estimation algorithm and achieved a Mean Absolute Error (MAE) of 0.3 m. In [7], the authors utilized a CNN algorithm to estimate the position of UWB device. They employed Red, Green, and Blue images extracted from the received UWB signal as input for the CNN algorithm. This approach resulted in an estimated position of the transmitter with a RMSE of approximately 0.5 m. In [8], the authors introduced a CNN-based approach for indoor localization utilizing Wi-Fi signals. Their simulations showcased that by leveraging Received Signal Strength Indicator (RSSI) values, an impressive accuracy of 95.76% in indoor localization can be attained.

Reinforcement learning has achieved great attention recently in the indoor localization problem [9]–[13]. In [9], the authors formulated the problem of indoor localization as a Markov decision process (MDP) by leveraging Wi-Fi technology and RSSI values. They proposed a Deep Q-Network (DQN) model and through simulation, demonstrated that a significant portion of the target can be localized within a distance of less than 0.2 m. In [10], the authors introduced a deep reinforcement learning approach for indoor localization using Wi-Fi technology and RSSI fingerprints. They employed a top-down searching strategy to enhance accuracy. Through simulations, they showed that their model successfully localized 75% of targets within indoor environments with an error of 0.55 m. In [11], the authors introduced a method for UWB node selection based on the MSE metric. They conducted simulations to assess this method's effectiveness and illustrated that utilizing UWB node groups with the lowest MSE in combination with the WLS method allowed for accurate localization of the mobile station within a NLOS environment, achieving an accuracy of less than 0.5 m. In [12], the authors introduced a new metric for UWB node selection known as Geometric Dilution of Precision (GDOP). They conducted simulations to assess the effectiveness of the GDOP metric in indoor localization. Their findings demonstrated that in challenging environments, utilizing the GDOP metric allowed for the localization of the mobile station with an accuracy of approximately 0.25 m. In [13], the authors introduced a framework called Deep Q-Learning Energy optimized LoS/NLoS for UWB node selection, aiming to balance localization accuracy and UWB beacon battery life. This framework utilized a deep Q-learning algorithm to train the mobile user in identifying a suitable set of UWB beacons with LOS links. The Target node was then localized using the TDOA framework, achieving a localization error of less than 0.4 m within the building.

Most studies on indoor localization with UWB devices have traditionally relied on fixed UWB beacons installed on building walls or anchor points with predetermined positions, which may pose limitations in scenarios involving unexpected infrastructure damage or large areas requiring numerous nodes [14]. In critical applications like healthcare, search and rescue operations, precise and rapid target localization is paramount. Thus, there is a pressing need for efficient approaches to address these challenges. In this paper, we propose a novel solution that leverages the high flexibility and mobility of mobile sensors as aerial anchor points. By employing a mobile UWB sensor trained to intelligently identify suitable anchor points, our approach enhances indoor localization accuracy.

III. Main Contributions and Organization

The main contributions of this paper are as follows:

1) Previous studies have primarily utilized fixed UWB beacons as receiver nodes to localize transmitter nodes indoors. However, in many scenarios, individuals may be unwilling to wear localization devices. In this paper, we present a novel approach to effectively and efficiently localize target indoors using a mobile UWB sensor deployed as a transceiver positioned outside the building. Our approach operates under the assumption that the target does not carry a device attached to its body.
2) In this paper, we approached the problem of effective UWB sensor placement by formulating it as an MDP. Unlike previous works such as [9] and [13], the key distinction lies in how we define the components of the MDP. Our proposed solution involves the interaction of a mobile UWB sensor with its environment, considering factors such as reflected direct and indirect rays from target and walls. The objective is to identify four suitable anchor points with minimal UWB signal attenuation.
3) To the best of our knowledge, there is currently no existing method that integrates Proximal Policy Optimization (PPO) and CNN algorithms, especially in the context of addressing various multi-path effects or NLOS propagation conditions inherent in UWB signals. Our approach aims to fill this gap by leveraging the combined power of PPO and CNN algorithms to enhance the accuracy of indoor localization under challenging signal propagation conditions.
4) In this paper, we demonstrate that after intelligently identifying anchor points, utilizing the received UWB signal time vector alongside the CNN algorithm leads to a significant enhancement in the performance of the indoor localization system.

The rest of this paper is organized as follows: section IV elaborates on how the mobile UWB sensor learns to intelligently identify four anchor points using the PPO algorithm.

In section V, we delve into the investigation of the indoor localization problem. The simulation results are presented in section VI. Lastly, section VII provides concluding remarks for the paper.

IV. Mobile UWB Sensor Effective Placement Problem

In this section, part IV-A begins by defining the problem of effective placement of the mobile UWB sensor as an MDP. Subsequently, part IV-B details the implementation of the PPO algorithm to address this MDP.

A. Mobile UWB Sensor Effective Placement Problem as an MDP

In our scenario, we envision a mobile UWB sensor situated outside a building, while a target is located inside. Initially, both the mobile UWB sensor and the target assume random positions. The primary objective of the mobile UWB sensor is to navigate towards anchor points where the transmitted signal power experiences minimal attenuation. We have imposed constraints on the motion of the mobile UWB sensor, restricting it to four orthogonal directions within a 2D surface, with a fixed step value for each motion. Notably, we don't impose any limitations on the sensor's flight path during the process of identifying anchor points. To reduce computational complexity in our approach, we assume that the target is positioned on the floor. This assumption enables us to estimate the 2D position of the target in our results. Therefore, to formulate the problem of effectively placing the mobile UWB sensor as an MDP under these conditions, we define the observation space, action space, and the reward function as follows:

1) Observation space: The mobile UWB sensor relies on two types of observations for decision-making in the environment. The first observation is its current and previous normalized 2D positions and the second is the received average power measurements at these positions.
2) Action Space: The action space for the mobile UWB sensor is discretized due to the high time cost associated with recording observations in continuous actions. Each action corresponds to the sensor's motion at each time step and is represented as a discrete movement in the environment. The action space consists of four possible actions of $(d_i, 0)$, $(-d_i, 0)$, $(0, d_i)$, $(0, -d_i)$. Here, each action vector represents the displacement in the x- and y-coordinates of the sensor, where d_i denotes the motion step value at time step i.
3) Reward Function: At each time step, the mobile UWB sensor receives a reward as follows:

$$R = \left(\left(\frac{r_1 + r_f}{th_1} \right) * r_{p_1} + r_{p_2} \right) + r_2 \tag{1}$$

where, $r_1 = p_i - p_{i-1}$ represents the difference between the normalized received average power in the current (p_i) and previous (p_{i-1}) time step. This term encourages the mobile UWB sensor to move towards points with higher received average power, facilitating the identification of suitable anchor points. The r_f function defines the reward value based on the number of suitable anchor points found by the mobile UWB sensor. It varies from 0.25 to 1 depending on the conditions specified as follows:

$$r_f = \begin{cases} 0.25 & \text{if } p_i > th_2 \ \& \ length(\boldsymbol{pos}) = 0 \\ 0.50 & \text{if } p_i > th_2 \ \& \ length(\boldsymbol{pos}) = 1 \ \& D > 1 \\ 0.75 & \text{if } p_i > th_2 \ \& \ length(\boldsymbol{pos}) = 2 \ \& D > 1 \\ 1 & \text{if } p_i > th_2 \ \& \ length(\boldsymbol{pos}) = 3 \ \& D > 1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

the $\boldsymbol{pos}$ matrix stores the 2D positions of these anchor points. It has dimensions of 2×4, accommodating the x and y coordinates of up to four anchor points. The $length()$ function counts the number of suitable anchor points stored in the $\boldsymbol{pos}$ matrix. We define D as the shortest distance between the anchor points stored in the $\boldsymbol{pos}$ matrix. For example, if p_i exceeds the threshold th_2 and the length of the $\boldsymbol{pos}$ matrix is 3 (indicating that three suitable anchor points have been found) and D is greater than 1, the mobile UWB sensor earns a reward of 1. In this case, its position is recorded as the desired position (suitable anchor points) in the $\boldsymbol{pos}$ matrix. Otherwise, it earns a reward of zero. In essence, the mobile UWB sensor is rewarded more as it finds more suitable anchor points. Therefore, the r_f value gradually increases from 0.25 to 1.

We considered the constraint that the round-trip distance of the mobile UWB sensor from the target should be less than 20 meters, which is a limitation of the UWB technology due to the low emission levels allowed [2]. To enforce this constraint, we incorporated penalties of the r_{p_1} and r_{p_2} as follows in the training process, which encourages the mobile UWB sensor to maintain its distance from the target and stay within the boundaries of the building.

$$r_{p_1} = \begin{cases} 0 & \text{if } \ x \ , \ y = 1 \parallel \ x \ , \ y = 0 \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

$$r_{p_2} = \begin{cases} -1 & \text{if } \ x \ , \ y = 1 \parallel x \ , \ y = 0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where x, y represent the current normalized 2D position of the mobile UWB sensor.

Finally, if p_i exceeds the threshold th_3 and the length of the $\boldsymbol{pos}$ matrix is equal to 4, then r_2 is added to the overall reward (R function) as follows:

$$r_2 = \begin{cases} 1 & \text{if } \ p_i > th_3 \ \& \ length(\boldsymbol{pos}) = 4 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

The thresholds th_1, th_2, and th_3 are chosen according to the desired value for the reward function. To achieve the maximum reward of 2 in each time step, we assume that the sum of r_1 and r_f should not exceed the th_1 threshold. Then, we divide the sum of these two values by the th_1 threshold. On the other hand, our goal is to reach the th_3 threshold, which in our scenario is the minimum normalized received average power acceptable for localization. To reach this threshold, we teach the mobile UWB sensor in two steps to achieve this value. In other words, we consider the th_2 threshold lower than the

final desired value and use it as an intermediate step towards reaching th_3. Here is the Algorithm 1 demonstrating how to calculate the reward function in each time step:

Algorithm 1 Calculation of the reward function in each time step

Input: p_i as the normalized received average power in i^{th} time step, D as the shortest distance between the anchor points in the $\boldsymbol{pos}$ matrix, th_1 , th_2 and th_3 as the desired thresholds

Output: R as reward function, $\boldsymbol{pos}$ as matrix of suitable anchor points

if $p_i < th_2$ **then**
$R = r_1/th_1$
else if $p_i > th_2$ & $length(\boldsymbol{pos}) = 0$ **then**
Store current mobile UWB sensor position into $\boldsymbol{pos}$,
$R = (r_1 + 0.25)/th_1$
else if $p_i > th_2$ & $length(\boldsymbol{pos})! = 0$ & $D < 1$ **then**
$R = r_1/th_1$
else if $p_i > th_2$ & $length(\boldsymbol{pos}) = 1$ & $D > 1$ **then**
Store current mobile UWB sensor position into $\boldsymbol{pos}$,
$R = (r_1 + 0.5)/th_1$
else if $p_i > th_2$ & $length(\boldsymbol{pos}) = 2$ & $D > 1$ **then**
Store current mobile UWB sensor position into $\boldsymbol{pos}$,
$R = (r_1 + 0.75)/th_1$
else if $p_i > th_2$ & $length(\boldsymbol{pos}) = 3$ & $D > 1$ **then**
Store current mobile UWB sensor position into $\boldsymbol{pos}$,
$R = (r_1 + 1)/th_1$
else if $p_i > th_3$ & $length(\boldsymbol{pos}) = 4$ & $D > 1$ **then**
$R = ((r_1 + 1)/th_1) + 1$
end if

B. PPO Implementation

The PPO algorithm [15] is implemented to train the mobile UWB sensor to find four suitable anchor points efficiently. PPO was chosen for its better convergence and performance in reinforcement learning environments, as well as its compatibility with both discrete and continuous action spaces. To expedite the agent's training in the simulation environment, multiprocessing training is utilized, with training occurring concurrently in four parallel environments. The Actor-Critic method [16] is employed in PPO algorithm, utilizing two deep neural networks for the Actor and Critic, respectively. The Actor network selects actions, while the Critic network evaluates the actions chosen by the Actor. According to Fig.1, both networks take a two-dimensional matrix as input, with dimensions of 80×3. Each column of this matrix contains three measurements: the received average power, and the length and width of the mobile UWB sensor's position. The Critic network's output is a single neuron providing an estimate of future total rewards, while the Actor network's output consists of four neurons, each returning a probability value for an action in the action space. Over time, the Actor learns to take better actions, while the Critic becomes more adept at evaluating those actions.

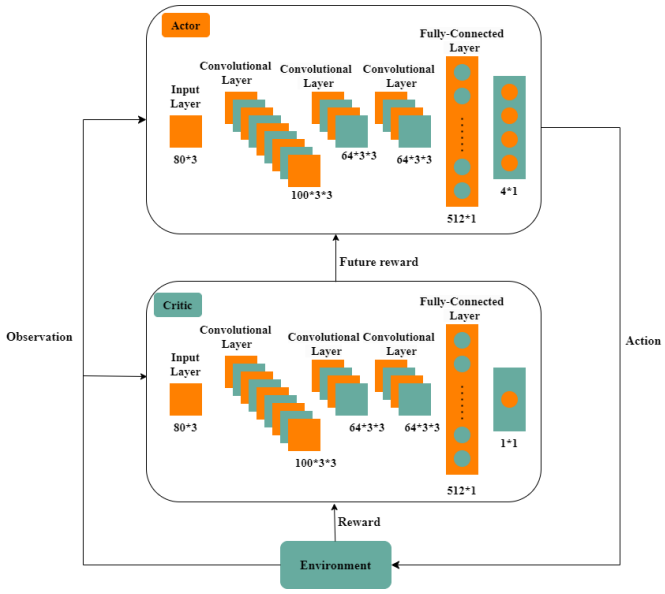


Fig. 1. The diagram of the proposed PPO framework.

One of the critical observations of the agent to find the suitable anchor points is the received average power measurements. To calculate these values, we use the modified Saleh Valenzuela (SV) channel (UWB channel) model based on the IEEE 802.15.4a channel model, as described in the Appendix. To train the agent to interact with the environment and maximize rewards, we define multiple episodes, each consisting of states and actions starting from the initial state and ending with the terminal state. At the beginning of each episode, we randomly select a target inside the building and the initial locations of the mobile UWB sensor outside the building. Each episode concludes if the mobile UWB sensor finds four suitable anchor points, exceeds the maximum number of steps, or reaches the wall of the building. The maximum number of episodes and steps per episode are set to 17,500 and 128, respectively. Additionally, the values for the discount factor, learning rate, and clipping rate are 0.99, 0.00025, and 0.2, respectively.

V. Indoor Localization Problem

Indoor localization aims to accurately determine the position of a target within an indoor environment. The proposed model, depicted in Fig. 2, utilizes a mobile UWB sensor acting as an aerial anchor. This sensor flies to four predetermined positions outside a building with dimensions of $20 \times 20 \times 2.5\text{m}^3$. These positions are strategically chosen using the trained PPO algorithm. This selection process enhances the accuracy of determining the 2D position of a target located inside the building.

To tackle the indoor localization problem, this paper leverages machine learning algorithms, notably the CNN, as outlined in the Appendix. Machine learning techniques are preferred due to their capacity to achieve accurate results with low energy consumption and cost, circumventing the necessity for intricate mathematical formulations. The flowchart in Fig. 3

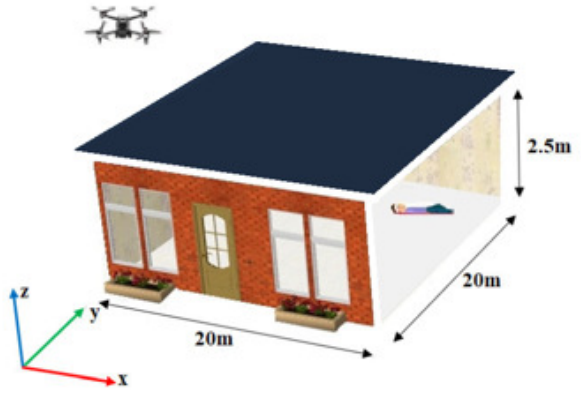


Fig. 2. The desired model for indoor localization.

delineates the phases involved in solving the indoor localization problem. In the initial phase, the PPO algorithm undergoes training to intelligently select four suitable anchor points. This process entails defining action and observation spaces alongside a reward function. Subsequently, in the second phase, the target is placed at M distinct known positions, and the mobile UWB sensor transmits and receives signals at the predetermined anchor points for each position. The received UWB signal time vector is then measured using Algorithm 2. Following this, measurements for the M target positions are randomly partitioned into training, validation, and testing datasets, with proportions of 60%, 20%, and 20% respectively. The training data undergoes preprocessing to ensure a clean dataset, which is then employed to train the CNN regressor. This CNN model elucidates the relationship between UWB signal measurements and their corresponding targets. Upon completion of CNN training, the target's position is estimated by analyzing the testing data.

Algorithm 2 Calculate the received UWB signal time vector

Input: p_t, c, d, w, f_c, x_t, x_{val}
Output: The received UWB signal time vector
Calculate $\beta_{0,l}$ based on equation (11)
Calculate T_l based on equation (12)
Calculate $\boldsymbol{RAY}$ as the vector to save the received UWB signal time of paths in first cluster as follows:
$\boldsymbol{RAY}(\mathbf{0},\mathbf{1}) = T_l$
$k = 0$
while $\boldsymbol{RAY}(\boldsymbol{k},\mathbf{1}) < T_{l+1}$ **do**
$k = k + 1$
$\boldsymbol{RAY}(\boldsymbol{k},\mathbf{1}) = \tau_{k,l}$
end while

VI. Simulation Result

In section VI-A, we assess the PPO algorithm's efficacy in identifying four anchor points during both training and testing phases. In VI-B, we compare intelligently selected anchor points with constant or random choices in indoor localization. Finally, in VI-C, we investigate indoor localization using methodologies from [6], [7], and [13] to validate the efficiency of our proposed method and dataset for target position estimation.

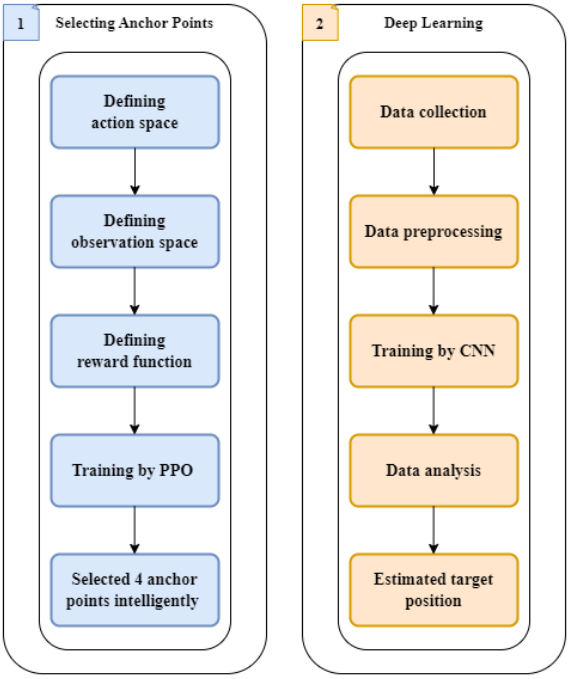


Fig. 3. Flowchart for solving the indoor localization problem.

A. Evaluation of the PPO Algorithm

Reinforcement Learning diverges from conventional machine learning methods in its treatment of the division between training and testing datasets. Unlike supervised learning, where training data is typically fixed and predetermined, in reinforcement learning, the agent dynamically acquires data through its interactions with the environment. This implies that during the training phase, the agent learns from the outcomes of its actions and the responses provided by the environment, rather than relying on a static dataset. During the testing phase, which aims to evaluate the performance of reinforcement learning algorithms, a smaller number of episodes compared to the training phase are typically used, and the environment is reset between episodes. Averaging the rewards obtained per episode in the testing phase yields a more reliable estimate of the agent's overall performance.

In a successful training phase, the average reward received should increase and then converge to a certain value in the final episodes. According to Fig. 4, it's evident that the agent has undergone a successful training phase with 17500 episodes. Additionally, converging to a reward value of 2 indicates the agent's success in finding the four suitable anchor points, as per the reward function defined in Section IV. To evaluate the PPO algorithm in the testing phase, a smaller number of episodes than the training phase are used. With 25 episodes and 128 time steps in each episode, the mobile UWB sensor achieves success. Success is defined by the agent's ability, with a mean reward of 1.85 and a mean step count of 23 per episode, to locate the four suitable anchor points.

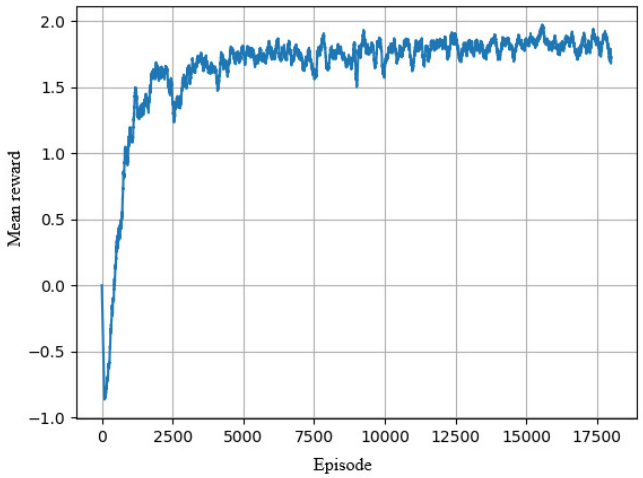


Fig. 4. Growth of the received average reward in the training process.

TABLE I
ACCURACY OF SELECTING ANCHOR POINTS

Selecting the anchor points	MAE [m]
Intelligent	0.09
Constant	0.45
Random	1.2

B. Performance Comparison

We introduce two anchor point selection modes for comparison:

- **Fixed anchor point selection:** Four predetermined positions of the mobile UWB sensor serve as anchor points for each target position inside the building.
- **Random anchor point selection:** Four positions of the mobile UWB sensor are randomly selected as anchor points for each target position inside the building.

To evaluate the performance of these modes, we employ the indoor localization model depicted in Fig. 2. The mobile UWB sensor acts as an aerial anchor, flying to four known positions outside the building, either at fixed positions or randomly selected ones. We utilize the proposed CNN algorithm flowchart for estimating the target position in the indoor environment. The mean absolute error (MAE) criterion is used for comparison, defined as:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{p} |s_i' - s_i|}{p} \tag{6}$$

where p represents the number of test samples, and s and s' indicate the actual and predicted target positions, respectively. Table I showcases the simulation results for the intelligent, constant, and random modes. Based on these results, a notable enhancement in indoor localization is observed when utilizing the PPO algorithm for intelligently selecting anchor points.

C. Validation of the Proposed Method with Related Works

In [6], authors proposed a method for estimating target positions in indoor environments using a combination of CNN algorithms and WLS estimators, with fixed anchor points. To assess its performance, a dataset was utilized to calculate the measured range (d') and the actual distance (d) between the mobile UWB sensor and the target. These parameters were computed using the modified SV channel model:

$$\mathrm{d}' = \frac{c * t'}{2} \tag{7}$$

$$\mathrm{d} = \frac{c * t}{2} \tag{8}$$

where c is the speed of light, t' is the time length of the first cluster, and t is the arrival time of the first cluster leader. The received UWB signal time was used as input to the CNN-based regression model to estimate ranging error, which was then employed in the WLS estimator for target position estimation. This result is shown in the **WLS+CNN** row in Table II.

Similar to [6], in [7], the authors utilized anchor points with predefined constant positions. We employed the CNN model proposed in [7] to estimate target positions, utilizing the received UWB signal time vector as input. Simulation results, presented in the **CNN** row of Table II, illustrate that intelligently selecting anchor points results in reduced localization errors.

Furthermore, we validated our method against [13], which assumes synchronized UWB anchor points and employs a TDOA framework for target position estimation. In this technique, the signal distance is computed by measuring the time it takes for the signal to reach the anchor points. In our scenario, assuming that the target is located in the range of i^{th} and j^{th} anchors, the TDOA information is defined by T_{ij} as follows [13]:

$$T_{ij} = \tau_i - \tau_j = \tag{9}$$

$$\frac{\sqrt{(x_t - x_i)^2 + (y_t - y_i)^2} - \sqrt{(x_t - x_j)^2 + (y_t - y_j)^2}}{c}$$

where τ_i, τ_j are the time length of the first cluster of i^{th} and j^{th} anchors. Also, (x_t ,y_t), (x_i ,y_i), and (x_j ,y_j), are the 2D position of target, i^{th} anchor, and j^{th} anchor, respectively. The best localization performance is achieved when the position estimation ranges used are within the line-of-sight range. However, due to multipath effects, TDOA measurements can be biased, leading to several meters of localization error. This result is shown in the **TDOA+PPO** row in Table II.

In conclusion, while intelligently selecting anchor points improves indoor localization accuracy, it's crucial to choose a suitable algorithm for processing received UWB signals. Combining intelligent anchor point selection with CNN-based signal processing yields significant improvements compared to fixed or random anchor point selection methods. Notably, our approach outperforms [6] and [7] in terms of accuracy.

TABLE II
ACCURACY OF LOCALIZATION METHODS

Ref/year	Localization algorithm	MAE [m]
[6]/2018	WLS+CNN	0.3
[7]/2020	CNN	0.5
[13]/2022	TDOA+DQN	0.3
This work	**TDOA+PPO**	**1.1**
This work	**CNN**	**0.2**
This work	**WLS+CNN**	**0.11**
This work	**CNN+PPO**	**0.09**

VII. CONCLUSION

This paper introduces a novel approach leveraging DRL to enhance indoor localization accuracy. Specifically, we employ the PPO algorithm to strategically position a mobile UWB sensor as an MDP, aiming to minimize the loss of transmitted signal power and achieve high localization accuracy. Our simulation results validate the effectiveness of the PPO algorithm in identifying four suitable anchor points with minimal steps. By integrating these anchor points with the CNN algorithm, our proposed method outperforms fixed or randomly-selected anchor points, as well as other related works, in terms of localization error reduction.

Furthermore, we acquired the necessary dataset through UWB channel simulation to address the indoor localization problem. However, in our future work, we aim to enhance the validity and efficiency of our proposed method by providing datasets from real-world scenarios. Additionally, to assess the scalability of PPO across diverse environments, it is imperative to tackle challenges such as defining various reward functions, incorporating a broader range of observations and actions, and designing more complex environments with obstacles.

ACKNOWLEDGMENT

This paper is extracted from the thesis approved and defended in the Faculty of Electrical and Computer Engineering of the Isfahan University of Technology. We would like to express our sincere gratitude to Dr. Nader Karimi, Associate Professor at Isfahan University of Technology for providing us with his laboratory facilities.

APPENDIX

A. UWB channel

The modified Saleh Valenzuela (SV) channel model (UWB channel), based on the IEEE 802.15.4a channel model is described as follows:

The impulse response of the modified SV channel model is represented as:

$$h(t) = \sum_{l=0}^{L} \sum_{k=0}^{K} \beta_{k,l} e^{j\theta_{k,l}} \delta(t - T_l - \tau_{k,l}) \tag{10}$$

where, L and K are the number of clusters and multipath components within each cluster, respectively. T_l is the arrival time of the first path of the l^{th} cluster. $\tau_{k,l}$, $\beta_{k,l}$, and $\theta_{k,l}$ are the arrival time, amplitude, and phase shift of the k^{th} path within the l^{th} cluster, respectively. $\delta(.)$ is the Dirac delta function. The leader power of the l^{th} cluster $(\beta_{0,l})^2$ and the arrival time of the first path of the l^{th} cluster (T_l) are calculated as follows:

$$20\log_{10}\beta_{(0,l)} = p_t^{dBm} + 10\log_{10}\left(\frac{c^2}{(4\pi d f_c)^2}\frac{1}{(1-(\frac{w}{2f_c})^2)}\right) \tag{11}$$

and

$$T_l = \frac{2\left\|x_{val} - x_t\right\|_2}{c} \tag{12}$$

where, p_t is the transmitter power, c is the speed of light, and d is the distance between the transceiver and the related virtual node. w and f_c indicate the bandwidth and central frequency, respectively. x_t and x_{val} are the position of the transceiver and virtual node, respectively.

The average power of the k^{th} path within the l^{th} cluster $(\beta_{k,l})^2$ decreases linearly with a time constant of γ. Moreover, the Nakagami distribution models the small-scale fading part of $\beta_{k,l}$. The shape parameter m is modeled as a log-normally distributed random variable with a mean of σ_m and standard deviation of μ_m. The channel phase shift is uniformly distributed between $[0, 2\pi]$. It is important to note that considering only the amplitude of the received signal in our approach eliminates the need to account for the phase shift of the channel component. It should be noted that the received average power of the UWB signal is determined based on the average power of the entire k path within the first cluster.

B. Regression with CNN

A Convolutional Neural Network (CNN) typically comprises an input layer, hidden layers, and an output layer. The hidden layers consist of convolution layers, pooling layers, and fully connected (FC) layers. Convolution layers employ filters to extract specific features from different regions of the input data. Pooling layers reduce the number of parameters while retaining crucial information. Subsequently, all data are flattened into a vector and fed into FC layers, akin to traditional neural networks. The Rectified Linear Unit (ReLU) function is commonly used in hidden layers to mitigate the vanishing gradient problem. However, as the network outputs continuous values for target positions, the sigmoid continuous activation function is utilized in the output layer. The Adaptive Moment Estimation (Adam) algorithm facilitates parameter updates, with actions such as incorporating more training data, reducing network capacity, and applying L_2 regularization to control overfitting. Optimal filter numbers and sizes, pooling sizes, and neuron counts in dense layers are determined to prevent overfitting. Moreover, the values for hyper-parameters like learning rate, batch size, and number of epochs are 0.001, 32, and 100, respectively.

LEYLA NOSRATI received the B.Sc. degree from the University of Bonab, Bonab, Iran, in 2017, and the M.Sc. degree in electrical engineering from the Isfahan University of Technology, Isfahan, Iran, in 2020. Her research interests include indoor localization, signal and image processing, computer vision, Ultra Wideband (UWB) technology, and unmanned aerial vehicle (UAV) communications.

SAMANEH HOSEINI is an assistant professor in the Department of Electrical and Computer Engineering at the Isfahan University of Technology. She is leading the Intelligent Self-Coordinating Teams (ISCT) laboratory and the Swarm Intelligence Robotics Team. Prior to that, she was a Postdoctoral fellow at the University of Toronto (Canada) under supervision of Prof. Hugh Liu and Prof. Anton de Ruiter. She obtained her B.Sc. and M.Sc. degrees in Computer Engineering from the Kashan and Isfahan Universities, Iran, in 2005 and 2007, respectively. After graduation with honor, she moved to Canada to pursue his academic studies and received her Ph.D. degree in Electrical and Computer Engineering from the University of Waterloo (Canada) in 2015. Her current research includes Deep Learning, Swarm Intelligence, Multi-agent Systems, and Robotics.

MOHAMMAD SADEGH FAZEL received the B.Sc. degree in electrical engineering (Electronics) from the Isfahan University of Technology (IUT), Isfahan, Iran, the M.Sc. degree in electrical engineering (Telecommunication Systems) University of Tehran, Tehran, Iran, and the Ph.D. degree in Wireless Communication Systems, Centre for Communication Systems Research from the University of Surrey, Guildford, Surrey, UK. In 2011, he joined the Department of Electrical and Computer Engineering at IUT as an Assistant Professor. His research interests include physical layer of wireless communication systems, massive MIMO systems, beamforming and NOMA, signal processing, and telecommunication in general.

SAJED RAKHSHANI received his B.Sc degree from the Shiraz University of Technology in 2014, and obtained his M.Sc degree in Communication Engineering from the Graduate University of Advanced Technology, Kerman, Iran, in 2017. His research interests include pattern recognition, heuristic optimization algorithms, deep neural networks, digital signal processors, and medical signal processing. He is currently pursuing his educations by researching at the Medical Image and Signal Processing research center at Isfahan university of medical sciences.

MOHAMMAD GHAVAMI is currently a Professor of telecommunications with the London South Bank University. Prior to this appointment, he was with King's College London, from 2002 to 2010, and with the Sony Computer Science Laboratories, Tokyo, from 2000 to 2002. He has authored the books, namely the Ultra Wideband Signals and Systems, and the Adaptive Antenna Systems, and has published over 180 technical papers mainly related to UWB and its medical applications. He holds three US and one European patents. He won the esteemed European Information Society Technologies Prize, in 2005, and two invention awards from Sony. He has been the Guest Editor of the IET Proceedings Communications, the Special Issue on Ultra Wideband Systems, and the Associate Editor of the Special Issue of the IEICE Journal on UWB Communications.

REFERENCES

[1] Z. Chen, H. Zou, J. Yang, H. Jiang, and L. Xie, "WiFi fingerprinting indoor localization using local feature-based deep LSTM," IEEE Systems Journal, vol. 14, no. 2, pp. 3001-3010, 2019.

[2] M. Ghavami, L. Michael, and R. Kohno, Ultra wideband signals and systems in communication engineering. John Wiley & Sons, 2007.

[3] S. Monica and G. Ferrari, "Accurate indoor localization with UWB wireless sensor networks," in 2014 IEEE 23rd International WETICE Conference, 2014: IEEE, pp. 287-289.

[4] A. Poulose and D. S. Han, "Feature-Based Deep LSTM Network for Indoor Localization Using UWB Measurements," in 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 298-301, 2021.

[5] X. Ye and Y. Zhang, "Research on UWB positioning method based on deep learning," in 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), pp. 1505-1508, 2020.

[6] K. Bregar and M. Mohorˇciˇc, "Improving indoor localization using convolutional neural networks on computationally restricted devices," IEEE Access, vol. 6, pp. 17429-17441, 2018.

[7] D. T. A. Nguyen, H.-G. Lee, J. Joung, and E.-R. Jeong, "Convolutional Neural Network-based UWB System Localization," in 2020 International Conference on Information and Communication Technology Convergence (ICTC): IEEE, pp. 488-490

[8] J.-W. Jang and S.-N. Hong, "Indoor localization with wifi fingerprinting using convolutional neural network," in 2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN), pp. 753-758, 2018.

[9] F. Dou, J. Lu, T. Xu, C.-H. Huang, and J. Bi, "A Bisection Reinforcement Learning Approach to 3-D Indoor Localization," IEEE Internet of Things Journal, vol. 8, no. 8, pp. 6519-6535, 2020.

[10] F. Dou, J. Lu, Z. Wang, X. Xiao, J. Bi, and C.-H. Huang, "Top-down indoor localization with Wi-fi fingerprints using deep Q-network," in 2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), 2018: IEEE, pp. 166-174.

[11] A. Albaidhani, A. Morell, and J. L. Vicario, "Anchor selection for UWB indoor positioning," Transactions on emerging telecommunications technologies, vol. 30, no. 6, p.e3598, 2019.

[12] A. Albaidhani and A. Alsudani, "Anchor selection by geometric dilution of precision for an indoor positioning system using ultra-wide band technology," IET Wireless Sensor Systems, vol. 11, no. 1, pp. 22-31, 2021.

[13] Z. Hajiakhondi-Meybodi, A. Mohammadi, M. Hou, and K. N. Plataniotis, "DQLEL: Deep Q-Learning for Energy-Optimized LoS/NLoS UWB Node Selection," IEEE Transactions on Signal Processing, vol. 70, pp. 2532-2547, 2022.

[14] P. Perazzo, L. Taponecco, A. A. D'amico, and G. Dini, "Secure positioning in wireless sensor networks through enlargement miscontrol detection," ACM Trans. Sensor Netw., vol. 12, no. 4, pp. 132, Nov. 2016.

[15] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2017.

[16] J. Peters and S. Schaal, "Natural actor-critic," Neurocomputing, vol. 71, no. 7-9, pp. 1180–1190, 2008.