基于连续动作空间深度强化学习的多数据融合室内定位方法

陈雪晨[1]，易嘉旋[1]，王霭祥[2]，邓晓衡[1]
（1. 中南大学电子信息学院，湖南 长沙 410004；2. 中南大学计算机学院，湖南 长沙 410083）

摘　要：基于智能手机的室内定位在研究和工业领域都引起了相当大的关注。然而在复杂的定位环境中，定位的准确性和鲁棒性仍然是具有挑战性的问题。考虑到行人航位推算（PDR, pedestrian dead reckoning）算法被广泛配备在最近的智能手机上，提出了一种基于双延迟深度确定性策略梯度（TD3, twin delayed deep deterministic policy gradient）的室内定位融合方法，该方法集成了 Wi-Fi 信息和 PDR 数据，将 PDR 的定位过程建模为马尔可夫过程并引入了智能体的连续动作空间。最后，与 3 个最先进的深度 Q 网络（DQN, deep Q network）室内定位方法进行实验。实验结果表明，该方法能够显著减少定位误差，提高定位准确性。

关键词：Wi-Fi；行人航位推算；室内定位；双延迟深度确定性策略梯度；深度强化学习
中图分类号：TN915.08
文献标志码：A
doi：10.11959/j.issn.2096−3750.2024.00358

Multi-data fusionaided indoor localization based on continuous action space deep reinforcement learning

CHEN Xuechen[1], YI Jiaxuan[1], WANG Aixiang[2], DENG Xiaoheng[1]
1. School of Electronic Information, Central South University, Changsha 410004, China
2. School of Computer Science and Engineering, Central South University, Changsha 410083, China

Abstract: Significant attention has been paid to indoor localization using smartphones in both research and industry. However, the accuracy and robustness of localization remain challenging issues, particularly in complex indoor environments. In light of the prevalent incorporation of pedestrian dead reckoning (PDR) devices in contemporary smartphones, an advanced indoor localization fusion method, anchored in the twin delayed deep deterministic policy gradient (TD3) framework, was proposed. In this approach, a seamless integration of Wi-Fi information and PDR data was achieved. The localization process of PDR was modeled as a Markov process, and a comprehensive continuous action space was introduced for the agent. To evaluate the performance of the proposed method, experiments were conducted and this approach was compared with three state-of-the-art deep Q network (DQN) based indoor localization methods. The experimental results demonstrate that the proposed method significantly reduces localization errors and enhances overall localization accuracy.

Key words: Wi-Fi, pedestrian dead reckoning, indoor localization, twin delayed deep deterministic policy gradient, deep reinforcement learning

0　引言

随着移动的互联网和物联网（IoT, internet of things）技术的发展，智能设备在智慧城市领域发挥着至关重要的作用，并提供各种智能服务，如智能交通系统、智能电网、动作感应游戏[1]等。随着

收稿日期：2023−10−23；**修回日期：**2024−01−24
通信作者：邓晓衡，dxh@csu.edu.cn
基金项目：国家自然科学基金项目（No. 62172441）；四川省重点研发计划（No. 2023YFG0120）
Foundation Items: The National Natural Science Foundation of China (No. 62172441), The Key Research and Development Program of Sichuan Province (No. 2023YFG0120)

智能手机的广泛使用，基于智能手机的室内定位在各种应用中越来越重要，例如购物中心，机场或医院的导航，老年人或残疾人的定位辅助等。室内位置服务巨大的社会和商业价值近年来引起了学术界和工业界的广泛关注。然而，应用于室内场景的基于位置服务仍然面临准确性和鲁棒性的挑战。

行人航位推算（PDR, pedestrian dead reckoning）旨在利用智能手机上的各种传感器来估计行人的步长、步数、方向和其他信息，从而实现室内定位。然而，PDR的准确性受到室内环境中各种因素的影响，例如磁场干扰可能导致估计方向与真实方向偏离，并导致位置误差的累积。近年来，学者对基于PDR的定位方法不断改进。例如，Wu等[2]提出了一种基于步态分析的PDR算法。该方法通过分析惯性测量单元数据的特征来确定步进模式，实现了运动方向的分类，获得了比传统PDR更好的定位效果。Ghaoui等[3]提出了一种基于PDR和粒子滤波（PF, particle filter）的人体运动似然网格和地板图滤波系统。由于Wi-Fi广泛存在于室内环境中，智能手机可以通过收集Wi-Fi信号来实现室内定位，而无须重新部署设备或进行额外布线。这种低成本、高准确性和易操作的技术使得利用Wi-Fi进行室内定位成为可行的选择。基于Wi-Fi的室内定位的常见方法是将其视为分类或回归问题，使用机器学习或深度学习求解[4-8]。Yu等[9]使用改进的概率线性判别分析将受干扰的接收信号强度从原始空间变换到潜在空间。然后，利用贝叶斯规则计算测试点与参考点相似度的后验概率，对后验概率最高的K个参考点进行加权估计位置。Liu等[10]提出了一种使用k-means聚类无线电映射来平衡定位精度和计算复杂度的加权K最近邻（WKNN, weighted k-nearest neighbor）定位策略。但是Wi-Fi信号容易波动，仅依靠Wi-Fi定位精度受限。因此，学者们提出了通过融合PDR和Wi-Fi信息，利用神经网络进行室内定位的方法[11-15]。例如，在文献[11]中Zhang等[11]把Wi-Fi和PDR信息拼接后输入长短期记忆（LSTM, long short-term memory）网络中，再用卡尔曼滤波器估计位置。PDR信息在行人行走初期能提供较精确的估计，而在后期Wi-Fi能够辅助PDR进行定位，并且神经网络可以从原始数据中学习更多抽象和高级的特征，这都有助于实现高精度的室内定位。

近年来，深度强化学习（DRL, deep reinforcement learning）在无人驾驶[16-19]、机器人技术[20-23]和游戏[24-26]等领域引起了广泛关注，并取得了显著的成功。DRL将深度学习的感知能力与强化学习的决策能力相结合，从而能够根据输入信息直接控制决策，更接近人类的思维方式。为了解决室内定位问题，研究人员提出了一种新颖的方法，将其建模为马尔可夫决策过程，并应用深度强化学习进行求解。具体如下，PDR方法根据先前位置和相应的行人信息计算当前位置。方向的确定对定位效果至关重要，研究人员将确定方向的过程建模为马尔可夫过程，并利用深度强化学习来解决该问题，其目标是找到最优策略以最小化定位误差。例如，Dou等[27]提出了一种利用深度Q网络（DQN, deep Q network）的新型室内定位技术。该方法采用二分法逐步缩小搜索区域，适用于2D、3D和多楼层环境的定位。实验结果显示，在各种场景下定位误差减小，并且与常规网格方法相比，训练时间显著缩短。Mohammadi等[28]提出了一种半监督学习方法，通过设计由变分自编码器（VAE, variational autoencoder）和DQN组成的网络来解决室内定位问题。所提出的方法与监督学习方法在数据集标签更少的情况下相比显示出更好的定位效果。同样，Li等[29]提出了一种基于DQN和室内地图信息的方向校正方法，用于室内行人跟踪。然而，这些方法只考虑了利用DQN解决Wi-Fi或PDR室内定位，忽视了基于数据融合的定位的效果。另一方面，这些方法中考虑的是离散动作空间，行动方向受限，难以有效接近目标位置。因此，本研究将Wi-Fi接收信号强度指示（RSSI, received signal strength indicator）和PDR信息进行融合，并利用基于双延迟深度确定性策略梯度（TD3, twin delayed deep deterministic policy gradient）的深度强化学习算法解决连续控制问题。

本文的贡献可以总结如下。

1) 将每个步骤中智能体的方向选择建模为马尔可夫决策优化问题，其目标是确定最优策略以最小化定位误差。融合了Wi-Fi数据和PDR信息，并将Wi-Fi RSSI、PDR估计方向和智能体目标位置的变化定义为状态转移。此外，当智能体跟踪行人轨迹时，将每个步骤智能体选择的方向视为动作决策。

2) 利用TD3深度强化学习算法，使智能体能够从连续的动作空间中选择任意方向移动。动作网

络会根据智能体当前状态输出一个确定的方向作为动作。智能体通过执行动作与室内环境交互来学习到达目标位置的策略。据了解，这是首次提出了针对智能体连续动作的室内定位方法。实验结果表明，该方法可以提高定位准确性。

1　相关介绍

本文所提方法主要基于深度强化学习，下面就相关概念和基本知识予以介绍。

强化学习的5个基本要素包括环境、智能体、状态、动作和奖励。强化学习起源于马尔可夫过程，其中智能体选择动作与环境进行交互，并根据动作获得奖励。这种持续的交互形成了一系列的$(s_k,a_k,r_k,s_{k+1},\cdots)$，$s_k$，$a_k$，$r_k$，$s_{k+1}$分别表示第$k$时刻的状态、动作、奖励和下一个状态，这构成了顺序决策过程。强化学习的目标是学习一个最优策略π，以最大化期望奖励。使用折扣奖励式(1)，可以计算随时间累积的期望奖励

$$R_k^{\gamma}=\sum_{i=k}^{N}\gamma^{i-k}R_{i+1} \tag{1}$$

其中，N表示最大步数，R_k^{γ}表示第k步的折扣奖励，它考虑了该步对未来奖励的影响。γ是奖励折扣因子，其取值范围介于0到1之间。较大的γ值表示更加关注未来的奖励。

Q-learning算法是一种基于价值的强化学习算法。其目标是通过在状态中选择最佳动作来最大化Q值，从而最大化期望奖励。该算法通过迭代确定动作价值函数，该函数表示在给定状态下采取动作的价值。动作价值函数通过估计从该状态采取特定行动获得的未来奖励总和的期望来计算。动作价值函数式如式(2)所示。

$$Q_{\pi}(s_k,a_k)=E_{\pi}[R_k^{\gamma}|s_k,a_k] \tag{2}$$

其中，$Q_{\pi}(s_k,a_k)$表示在步骤k的状态s_k中采取动作a_k所获得的期望折扣奖励。可以通过式(1)把动作价值函数转换为式(3)形式

$$Q_{\pi}(s_k,a_k)=E_{\pi}\left[\sum_{i=k}^{N}\gamma^{i-k}R_{i+1}|s_k,a_k\right] \tag{3}$$

最优策略π^*可以最大化式(3)，而最优动作价值函数可以通过贝尔曼方程转换为式(5)。

$$Q_{\pi^*}(s_k,a_k)=E\left[R_{k+1}+\gamma\max Q_{\pi^*}(s_{k+1},a_{k+1})\right] \tag{4}$$

$$Q_{\pi^*}(s_k,a_k)=\sum_{s_{k+1}}p(s_{k+1},r|s_k,a_k)\left[r+\gamma\max Q_{\pi^*}(s_{k+1},a_{k+1})\right] \tag{5}$$

其中，Q表用于记录每个状态下所有动作的折扣奖励，而Q-learning算法通过应用贝尔曼方程来计算动作价值函数。然而，在实际问题中，状态和动作通常是连续的，使用Q表存储大量的Q值是不可行的。为了解决这个问题，引入了深度神经网络来拟合最优的动作价值函数，即深度Q-learning，也称为DQN[30]。研究表明，深度神经网络通过神经元和激活函数的连接组合可以拟合任何函数，因此其非常适合拟合动作价值函数、策略函数等。DQN利用深度神经网络强大的特征提取能力，能够从大量高维状态中提取特征并映射到动作上。经过良好的训练，可以得到一个能够近似表达最优动作价值函数的神经网络。DQN算法包括两个Q网络：参数为θ的Q网络和参数为θ'的目标Q网络，其中动作价值函数的表达式为$Q(s,a;\theta)$。尽管DQN算法在游戏控制、决策制定、机器人导航等各个应用领域取得了巨大的成功，但其一个缺点是只适用于离散动作空间，无法有效解决连续动作空间的问题。因此，为了应对连续控制的挑战，Silver等[31]提出了深度确定性策略梯度（DDPG, deep deterministic policy gradient）算法。DDPG算法将行动者评论家（AC, actor-critic）机制与神经网络相结合，该算法中包含4个神经网络。其中一个是评价网络，用于输出动作价值，相当于动作价值函数。该神经网络主要学习动作价值函数，以逼近从样本中计算得到的动作价值，并且其更新方法类似于DQN算法中的Q网络。另一个是动作神经网络，用于在连续动作空间中输出具体的行动。此外，还有相应的目标网络。动作网络根据策略梯度的方法进行更新，主要学习策略函数，以提高输出高价值动作的概率。

然而，DDPG算法存在过度估计的问题。因此，文献[32]中提出了DDPG的一个变种，称为TD3。TD3引入了3个关键技术来应对这些问题。

1) TD3提出了一种对目标策略进行平滑正则化的方法。在计算动作价值的目标值时，该方法通过给下一个状态的动作添加噪声扰动来实现。这有助于价值评估更加准确和可靠。

2) TD3采用了双网络的方法。它使用两个评价网络，其中较小价值的网络用于计算目标值。这种方法有助于解决网络过度估计的问题，并提高价值评估的准确性。

3) TD3提出了延迟更新机制。在评价网络进行

多次更新后，更新一次动作网络，以确保动作网络的训练更加稳定。

这3个关键技术使得TD3成为DDPG的一种更加有效和稳定的变体。

2　基于双延迟深度确定性策略梯度（TD3）的室内定位融合方法

2.1　数据融合和马尔可夫过程建模

在室内场景中，环境被定义为带有Wi-Fi信号的建筑楼层的室内布局。智能体是强化学习的核心，充当学习者或决策制定者的角色。它观察环境，根据当前环境的状态选择行动，并在迭代过程中进入下一个状态。智能体的目标是通过最大化一系列动作选择所获得的奖励来达到目标位置。随着训练的进行，智能体学习到一个优化的策略。接下来将详细描述马尔可夫过程的状态、动作和奖励。

首先，将智能体的状态信息定义为Wi-Fi信息和PDR数据的融合表示。智能体状态表示为$(x, y, \mathbf{RSSI}_{k-1}, d_k, \mathbf{RSSI}_k)$。

1) x和y是二维空间中的横纵坐标，代表智能体的当前位置。智能体的位置通过动作网络输出的方向和IMU估计的移动距离基于前一个位置来求得。

2) $\mathbf{RSSI}_{k-1}$是一个L维向量，代表行人轨迹中第$k-1$位置接收到的所有Wi-Fi信号的RSSI。L表示可以接收到的Wi-Fi数量。例如，如果环境中有20个不同的Wi-Fi信号，那么$\mathbf{RSSI}_{k-1}$就是一个20维向量。但是，在无法测量到RSSI的情况下，使用−100 dB进行替代。

3) d_k是根据传感器记录的信息推断出的行人从第$k-1$位置步行到第k位置的行走方向。

4) $\mathbf{RSSI}_k$记录了行人轨迹在第k个位置的RSSI，而第k个位置是智能体在第k个状态中的目标位置。

其次，本文将智能体动作建立为连续区间表示。在当前状态下，智能体依次从动作空间中选择动作，动作被定义为智能体前进的方向，其取值范围为连续区间$(-\pi, \pi)$。与基于DQN的室内定位方法不同，后者仅限于离散的动作空间，而智能体能从连续的动作空间中选择任意方向。

奖励被定义为智能体在执行动作后从环境中获得的反馈，表示智能体动作的质量。因此，智能体越接近目标位置，获得的奖励就越高。相反，如果智能体与室内障碍物（如墙壁）发生碰撞，它将受到惩罚。奖励函数的定义如式(6)所示。

$$r_k = \begin{cases} 0, \text{智能体发生碰撞} \\ \exp\left(-\mathrm{dis}\left(\hat{P}_k, P_k\right)\right), \text{智能体没有发生碰撞} \end{cases} \tag{6}$$

智能体在位置$\hat{P}_{k-1}$执行动作a_k，移动到位置$\hat{P}_k$。P_k是智能体在位置$\hat{P}_{k-1}$的目标位置。

2.2　基于TD3的室内定位算法

在本小节中，将具体阐述如何利用TD3算法解决第2.1节中引入数据融合的马尔可夫过程。该算法的神经网络架构包括3个主要部分：动作网络和目标动作网络构成第一部分，评价网络部分被划分为两个子部分——评价1网络和评价2网络，它们都有相应的目标网络。采用全连接层来构建评价2网络与评价1网络。动作网络、评价1网络和评价2网络的参数分别用ϵ、ϑ_1和ϑ_2来表示，相应的目标网络参数表示为ϵ'、ϑ_1'和ϑ_2'。在每个步骤中，智能体根据自身位置获得状态信息，包括Wi-Fi信号的RSSI和由PDR估计的行进方向。动作网络利用状态信息输出动作，智能体执行该动作，并利用PDR估计的移动距离移动到下一个位置，得到下一个状态。过程迭代进行，从而得到行人轨迹的估计。基于TD3的室内定位融合算法架构如图1所示。

在训练阶段，将智能体初始化在行人轨迹的起始位置，并旨在引导其尽可能接近后续目标位置。将融合了Wi-Fi RSSI和PDR信息的状态s_k输入动作网络后，动作网络根据智能体的状态输出相应的动作。如果动作网络经过充分的训练并且接近最优策略函数，则可以最大化累积奖励。然而，在训练的早期阶段，策略函数往往不是最优的，可能导致局部最优的动作选择。因此，通过向动作a_k加入高斯噪声并进行随机采样，来鼓励对未知动作的探索，同时利用已知动作的经验。这种方法平衡了探索和利用的关系，并有助于有效学习最优策略。

$$a_k = \pi\left(s_k; \epsilon\right) + \delta, \delta \sim \mathrm{N}\left(0, \sigma^2\right) \tag{7}$$

在采样之后，智能体执行动作并利用PDR估计的移动距离进行状态转移，下一个状态记为s_{k+1}。随后，智能体计算定位误差并根据其获得相应的奖励，记为r_k。为了打破连续样本的相关性并帮助网络训练，仍然采用类似DQN的经验重放方法对样本进行随机化处理。将样本元组$\left(s_k, a_k, r_k, s_{k+1}\right)$存储在经验池$D$中，并在每次迭代中从经验池$D$中随机抽取$M$个样本，记为$(s_t, a_t, r_t, s_{t+1})$。在每次迭代中，以下网络参数需要更新：评价1网络、评价2网络、

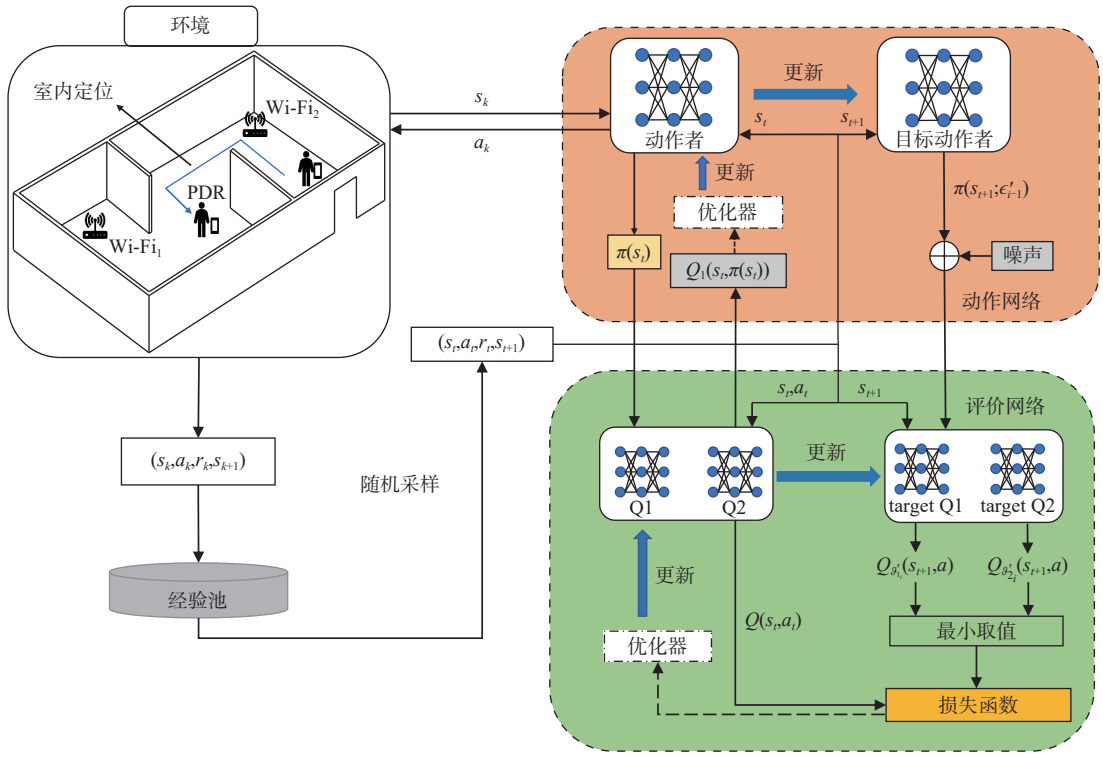


图1　基于TD3的室内定位融合算法架构

动作网络、目标评价1网络、目标评价2网络和目标动作网络，它们的参数分别表示为ϑ_{1_i}、ϑ_{2_i}、ϵ_i、ϑ'_{1_i}、ϑ'_{2_i}和ϵ'_i。首先更新的是评价1网络和评价2网络，它们的更新方式相同。根据TD error来更新评价1网络的参数。$L_i(\vartheta_{1_i})$是评价1网络的损失函数。

$$L_i(\vartheta_{1_i})=\frac{1}{M}\sum(y_t-Q_1(s_t,a_t;\vartheta_{1_i}))^2 \tag{8}$$

计算$y_t=r_t+\gamma\min\sum_{j=1,2}Q_{\vartheta'_j}(s_{t+1},a)$，$y_t$作为目标$Q$值，而$a=\pi(s_{t+1};\epsilon'_{i-1})+\delta_1$，其中$\delta_1$~N$(0,\sigma^2)$表示一个服从正态分布的噪声项。接着计算参数$\vartheta_{1_i}$的梯度$\nabla_{\vartheta_{1_i}}L_i(\vartheta_{1_i})$，然后使用随机梯度下降法迭代更新动作网络的参数$\vartheta_{1_i}$。

接下来是对动作网络的更新。评价网络的准确性对动作网络的训练非常重要，训练良好的评价网络能指导动作网络输出价值高的动作。为了稳定动作网络的训练，会在多次训练评价网络后进行一次动作网络的训练。利用策略梯度计算损失值，梯度$\nabla_{\epsilon_i}J(\epsilon_i)$计算如式(9)。

$$\nabla_{\epsilon_i}J(\epsilon_i)=\frac{1}{M}\nabla_{a_{\epsilon_i}}Q_{\vartheta_{1_i}}(s_i,a_{\epsilon_i})\nabla_{\epsilon_i}\pi_{\epsilon_i}(s_i) \tag{9}$$

然后，使用随机梯度下降方法迭代更新动作网络的参数ϵ_i。最后，为了模型训练的稳定性，采用延迟更新的方式来更新目标网络。如式(10)～式(12)所示。

$$\vartheta'_{1_i}=\tau\vartheta_{1_i}+(1-\tau)\vartheta'_{1_{i-1}} \tag{10}$$

$$\vartheta'_{2_i}=\tau\vartheta_{2\,i}+(1-\tau)\vartheta'_{2_{i-1}} \tag{11}$$

$$\epsilon'_i=\tau\epsilon_i+(1-\tau)\epsilon'_{i-1} \tag{12}$$

在训练阶段，提出的算法详见算法1。

算法1　基于TD3的室内定位算法

网络和目标网络初始化相同权重，初始化经验池D;

for 每个轮次

　for 每个行人轨迹

　　获取智能体的初始状态s_0;

　　for 每个步骤i

　　根据式(7)采样动作a_k;

　　执行动作a_k，转移到下一个状态s_{k+1}，并计算奖励r_k;

　　将经验元组(s_k,a_k,r_k,s_{k+1})存储到经验池D中;

　　从经验池D随机抽取M个样本;

　　根据式(8)，通过随机梯度下降更新评价1网络和评价2网络;

　　if i mod $f==0$

用梯度策略更新动作网络；
用软更新法更新目标网络；
end if
end for
end for
end for

3　实验与分析

本节将提出的室内定位融合方法与多种不同方法进行对比。实验结果表明了集成多种数据和采用连续动作空间的室内定位方法在室内定位方面具有显著效果。

3.1　实验设置

本文在中南大学的实验楼和教学楼中进行了室内定位实验。实验楼楼层面积约为1 309 m^2（77 m×17 m），教学楼的楼层面积约为1 533 m^2（42.0 m×36.5 m）。参数设置如下：奖励折扣因子γ设置为0.9，学习率设置为0.000 1。为了探索和目标平滑正则化，在动作和目标动作上都添加了服从N~$(0, 0.1)$分布的噪声δ和δ_1。经验池的大小设置为1.0×10^6，每对评价网络进行3次训练，就对动作网络进行一次训练并软更新3个目标网络。在网络训练过程中，使用的批量大小为16。网络架构都是由全连接层组成的，动作网络的输出角度被限制在$(-\pi, \pi)$范围内，以获得最佳结果。

在定位实验中，本研究在数据集方面主要关注Wi-Fi RSSI和PDR信息的收集。为了确保训练的模型能够良好地泛化，需要获取大量的行人轨迹数据。由于Wi-Fi RSSI数据和PDR数据的收集是一项耗时且劳动密集的过程，本文将实验楼楼层的走廊划分为相等大小的网格，每个网格的尺寸为0.6 m×0.6 m。本研究在每个网格中使用手机APP[33]测量100次RSSI，之后对生成的虚拟PDR信息添加高斯噪声来模拟现实世界的环境干扰。为了生成虚拟行人轨迹，随机选择一个初始位置，并允许智能体继续移动。每当智能体移动到一个网格中，随机抽取该网格的RSSI测量值作为智能体的Wi-Fi测量值[34]。实验数据集包含大量的轨迹数据，每个轨迹数据包含不同的位置坐标、对应的RSSI和PDR信息。

最后，在上述场景中对提出的方法进行了测试。为了捕捉与行人相关的信息，如方向和步长，在行走阶段本文同样使用APP进行采样。在进行Wi-Fi信号测量时，从先前收集的数据点中随机采样。

3.2　实验结果

本文与多种室内定位方法进行了对比实验，分别是Wi-Fi-DQN、PDR-DQN[27]、Wi-Fi-PDR-DQN、WKNN-Weighted-Euclidean[35]和Adaptive-WKNN[36]。不同方法的平均奖励曲线如图2所示。从图2中可以观察到，在模型按照梯度下降策略进行更新时，TD3方法的平均奖励值会出现波动。然而，随着模型的不断更新，平均奖励逐渐收敛到一个稳定的值。奖励值是对智能体采取的动作的反馈。在训练阶段，奖励值持续增加，表明智能体的策略不断优化，使其选择更接近目标位置的动作。此外，TD3方法的平均奖励值高于其他方法，说明提出的方法在训练中表现更好。图2所展示的结果证实了提出方法的稳定性。

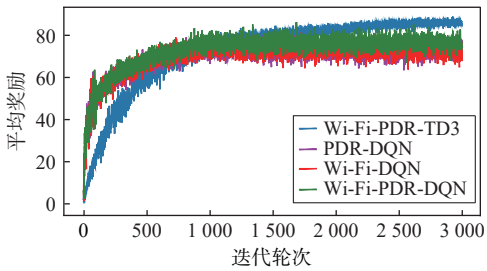


图2　不同方法的平均奖励曲线

为了评估提出的方法的定位效果，在特定楼层上进行多次测试，其中一次测试了约75 m长的行人轨迹。起初，记录了实际的行人轨迹，包括路径上各个位置的坐标，以及相应的PDR信息和RSSI值。在测试过程中，还记录了使用不同方法获得的轨迹的位置坐标。最后，使用式(13)计算轨迹上每个位置的定位误差。

$$\text{Error}_k = \sqrt{(x_{\hat{P}_k} - x_{P_k})^2 + (y_{\hat{P}_k} - y_{P_k})^2} \tag{13}$$

其中，$x_{\hat{P}_k}$和$y_{\hat{P}_k}$表示算法估计的第k个位置$\hat{P}_k$的横坐标和纵坐标，x_{P_k}和y_{P_k}表示轨迹中第k个位置P_k的实际横坐标和纵坐标。轨迹的平均定位误差通过式(14)计算。

$$\text{average Error} = \frac{1}{N}\sum_{k=1}^{N}\text{Error}_k \tag{14}$$

其中，N是行人轨迹上的位置数量。

在实验楼不同方法的拟合轨迹对比如图3所示。由于PDR只能预测行人的移动信息，估计当前位置也需要上一个位置的信息。因此，每种方法

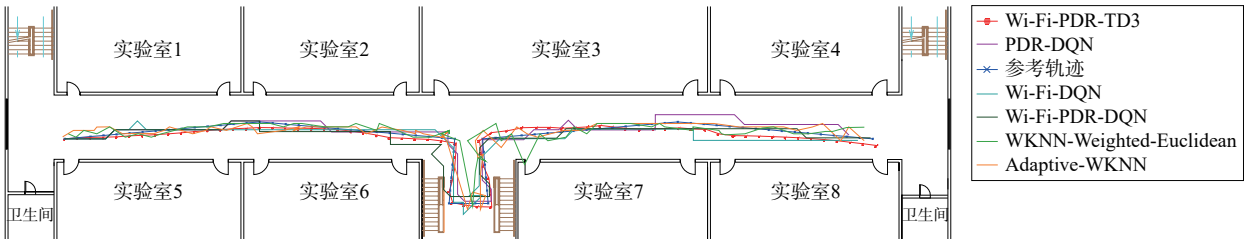


图3　在实验楼不同方法的拟合轨迹对比

都从轨迹的真实初始位置进行测试。从图3中可以看出，尽管智能手机上的传感器收集到的行人信息可能存在偏差，但经过训练的PDR-DQN智能体仍然可以选择更好的移动方向。因此，PDR-DQN方法的结果与由蓝色线表示的目标轨迹相匹配，但由于离散的动作空间限制，该方法的平均定位误差为0.74 m。本文提出的Wi-Fi-PDR-TD3方法不仅具有IMU信息，还将Wi-Fi的RSSI作为辅助信息。智能体可以从$(-\pi,\pi)$区间中选择任意方向移动。该方法对行人轨迹有更好的跟踪效果，平均定位误差为0.64 m。与PDR-DQN相比，Wi-Fi-PDR-TD3方法将定位精度提高了13.5%。Wi-Fi-DQN、Wi-Fi-PDR-DQN、WKNN-Weighted-Euclidean和Adaptive-WKNN的平均定位误差分别为0.86 m、0.68 m、0.96 m和0.86 m。使用累积分布函数（CDF, cumulative distribution function）来表示不同方法的定位误差分布，在实验楼不同方法的定位误差累积概率分布如图4所示，从图中可以看出，所提出的Wi-Fi-PDR-TD3方法具有更好的定位效果，因为其累积分布曲线高于其他方法。

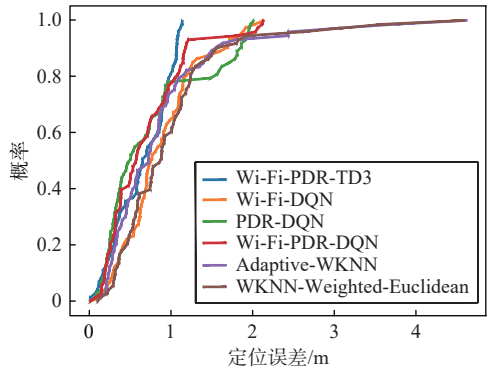


图4　在实验楼不同方法的定位误差累积概率分布

为了验证方法的鲁棒性，本文在教学楼进行了实验。该实验用多种方法对一条长度约为80 m的行人轨迹进行拟合。在教学楼不同方法的拟合轨迹对比如图5所示。本文的方法取得了最好的定位效果，平均定位误差为0.85 m。而PDR-DQN、Wi-Fi-DQN、Wi-Fi-PDR-DQN、WKNN-Weighted-Euclidean和Adaptive-WKNN的平均定位误差分别为0.95 m、1.02 m、0.96 m、1.32 m和1.13 m。在教学楼不同方法的定位误差累积概率分布如图6所示。本文的方法相较于PDR-DQN提高了10.5%的定位精度。经过分析，PDR在行人行走的初期能够提供较为精确的方向估计，但随着行人行走，估计的方向出现了较大误差。较大误差的方向作为状态会提供错误信息让智能体难以学习，本文利用数据融合，把Wi-Fi也加入智能体状态中。尽管Wi-Fi也有测量误

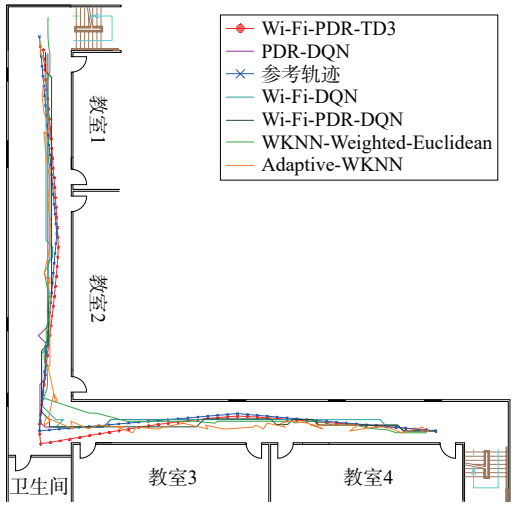


图5　在教学楼不同方法的拟合轨迹对比

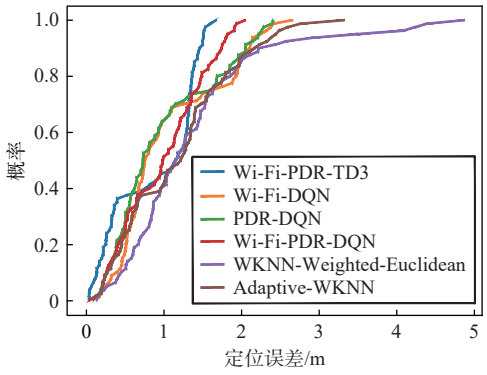


图6　在教学楼不同方法的定位误差累积概率分布

差，但相较于PDR后期的估计信息，有着较高的精度。利用神经网络强大的特征提取能力，融合了PDR和Wi-Fi信息的状态能提供正确的信息让智能体学习。并且引入了连续动作空间，智能体可选择的方向更加多样。结合这两个方面，本文的方法相比其他方法提高了定位精度。

4　结束语

本文提出了一种新的室内定位方法，该方法通过结合Wi-Fi RSSI和PDR的信息，将室内定位问题建模为马尔可夫过程，并采用了深度强化学习算法TD3进行定位。为了提高定位准确度并消除智能体的方向约束，引入了连续的运动空间，允许智能体选择任意角度的方向作为动作。该方法与传统PDR的区别在于通过神经网络的输出确定方向，而不是仅仅依赖IMU导出的方向。虽然IMU可以直接测量方向，在行人行走初期可以提供精度较高的估计，但它仍然容易受到外部影响，导致累积误差。这种敏感性就是仅依靠PDR往往无法实现高精度室内定位的原因。相反，将IMU估计的方向作为神经网络的输入，然后对方向进行调整校正。经过充分的训练后，网络的输出方向得到优化，以更好地靠近目标位置。从本质上讲，神经网络充当方向校正的作用。实验结果表明，本文的方法在跟踪行人轨迹和定位准确度方面表现更加出色。

参考文献：

[1] LIU M, WANG H J, YANG Y, et al. RFID 3-D indoor localization for tag and tag-free target based on interference[J]. IEEE Transactions on Instrumentation and Measurement,2019, 68(10): 3718-3732.

[2] WU R J, PIKE M, CHAI X Q, etal. GA-PDR: using gait analysis for heading estimation in PDR based indoor localization system[C]// Proceedings of the IECON 2023- 49th Annual Conference of the IEEE Industrial Electronics Society. Piscataway: IEEE Press, 2023: 1-6.

[3] GHAOUI M A, VINCKE B, REYNAUD R. Human motion likelihood representation map-aided PDR particle filter[J]. IEEE Sensors Journal, 2023, 23(1): 484-494.

[4] SUN W, XUE M, YU H S, et al. Augmentation of fingerprints for indoor WiFi localization based on Gaussian process regression[J]. IEEE Transactions on Vehicular Technology, 2018, 67(11): 10896-10905.

[5] LI Z, RAO X P. Toward long-term effective and robust device-free indoor localization via channel state information[J]. IEEE Internet of Things Journal, 2022, 9(5): 3599-3611.

[6] GAO B, YANG F, CUI N, et al. A federated learning framework for fingerprinting-based indoor localization in multibuilding and multi-floor environments[J]. IEEE Internet of Things Journal, 2023, 10(3): 2615-2629.

[7] KIM M, HAN D, RHEE J K K. Multiview variational deep learning with application to practical indoor localization[J]. IEEE Internet of Things Journal, 2021, 8(15): 12375-12383.

[8] ZOU H, CHEN C L, LI M X, et al. Adversarial learning-enabled automatic WiFi indoor radio map construction and adaptation with mobile robot[J]. IEEE Internet of Things Journal, 2020, 7(8): 6946-6954.

[9] YU D, LI C G. An accurate WiFiindoor positioning algorithm for complex pedestrian environments[J]. IEEE Sensors Journal, 2021, 21(21): 24440-24452.

[10] LIU S Y, DE LACERDA R, FIORINA J. WKNN indoor Wi-Fi localization method using k-means clustering based radio mapping[C]// Proceedings of the 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring). Piscataway: IEEE Press, 2021: 1-5.

[11] ZHANG M Y, JIA J, CHEN J, et al. Indoor localization fusing WiFi with smartphone inertial sensors using LSTM networks[J]. IEEE Internet of Things Journal, 2021, 8(17): 13608-13623.

[12] LIANG Q, LIU M. An automatic site survey approach for indoor localization using a smartphone[J]. IEEE Transactions on Automation Science and Engineering, 2020, 17(1): 191-206.

[13] LIU R, MARAKKALAGE S H, PADMAL M, et al. Collaborative SLAM based on WiFi fingerprint similarity and motion information[J]. IEEE Internet of Things Journal, 2020, 7(3): 1826-1840.

[14] ZHAO Y H, ZHANGZ X, FENG T Y, et al. GraphIPS: calibration-free and map-free indoor positioning using smartphone crowdsourced data[J]. IEEE Internet of Things Journal, 2021, 8(1): 393-406.

[15] DU X Q, LIAO X W, LIU M M, et al. CRCLoc: a crowdsourcing-based radio map construction method for WiFi fingerprinting localization[J]. IEEE Internet of Things Journal, 2022, 9(14): 12364-12377.

[16] ZHOU Y, MA X Y, HU S T, et al. QoE-driven adaptive deployment strategy of multi-UAV networks based on hybrid deep reinforcement learning[J]. IEEE Internet of Things Journal, 2022, 9(8): 5868-5881.

[17] ZHAO Y J, MA Y, HU S L. USV formation and path-following control via deep reinforcement learning with random braking[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(12): 5468-5478.

[18] OUBBATI O S, ATIQUZZAMAN M, BAZ A, et al. Dispatch of UAVs for urban vehicular networks: adeep reinforcement learning approach[J]. IEEE Transactions on Vehicular Technology, 2021, 70(12): 13174-13189.

[19] CHU N H, HOANGD T, NGUYEN D N, et al. Joint speed control and energy replenishment optimization for UAV-assisted IoT data collection with deep reinforcement transfer learning[J]. IEEE Internet of Things Journal, 2023, 10(7): 5778-5793.

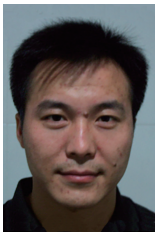
[20] LIU Z, LIU Q M, TANG L, et al. Visuomotor reinforcement learning for multirobot cooperative navigation[J]. IEEE Transactions on Automation Science and Engineering, 2022, 19(4): 3234-3245.
[21] ZHU W, GUO X, OWAKI D, et al. A survey of sim-to-real transfer techniques applied to reinforcement learning for bioinspired robots[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(7): 3444-3459.
[22] CHEN L, WANG Y N, MIAO Z Q, et al. Transformer-based imitative reinforcement learning for multirobot path planning[J]. IEEE Transactions on Industrial Informatics, 2023, 19(10): 10233-10243.
[23] HAN J I, LEE J H, CHOI H S, et al. Policy design for an ankle-foot orthosis using simulated physical human-robot interaction via deep reinforcement learning[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering: a Publication of the IEEE Engineering in Medicine and Biology Society, 2022, 30: 2186-2197.
[24] BENADDI H, IBRAHIMI K, BENSLIMANE A, et al. Robust enhancement of intrusion detection systems using deep reinforcement learning and stochastic game[J]. IEEE Transactions on Vehicular Technology, 2022, 71(10): 11089-11102.
[25] OH I, RHO S, MOON S, et al. Creating pro-level AI for a real-time fighting game using deep reinforcement learning[J]. IEEE Transactions on Games, 2022, 14(2): 212-220.
[26] XU P, YIN Q Y, ZHANG J G, et al. Deep reinforcement learning with part-aware exploration bonus in video games[J]. IEEE Transactions on Games, 2022, 14(4): 644-653.
[27] DOU F, LU J, XU T Y, et al. A bisection reinforcement learning approach to 3-D indoor localization[J]. IEEE Internet of Things Journal, 2021, 8(8): 6519-6535.
[28] MOHAMMADI M, AL-FUQAHA A, GUIZANI M, et al. Semisupervised deep reinforcement learning in support of IoT and smart city services[J]. IEEE Internet of Things Journal, 2018, 5(2): 624-635.
[29] LI Q, LIAO X W, GAO Z Z. An enhanced direction calibration based on reinforcement learning for indoor localization system[C]// Proceedings of the 2020 IEEE Wireless Communications and Networking Conference (WCNC). Piscataway: IEEE Press, 2020: 1-6.
[30] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518: 529-533.
[31] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529: 484-489.
[32] FUJIMOTO S, VAN HOOF H, MEGER D. Addressing function approximation error in actor-critic methods[EB/OL]. 2018: arXiv: 1802.09477.http://arxiv.org/abs/1802.09477.pdf
[33] QIAN Y X, CHEN X C. An improved particle filter based indoor tracking system via joint Wi-Fi/PDR localization[J]. Measurement Science and Technology, 2021, 32(1): 014004.
[34] LI Y, HU X, ZHUANG Y, et al. Deep reinforcement learning (DRL): another perspective for unsupervised wireless localization[J]. IEEE Internet of Things Journal, 2020, 7(7): 6279-6287.
[35] LEI P, LI Y, YUAN L, et al. An improved wifi fingerprint location method for indoor positioning[C]//Proceedings of the 2022 China Automation Congress (CAC). Piscataway: IEEE, 2022: 423-427.
[36] LIU S, LACERDA R D, FIORINA J. Performance analysis of adaptive k for weighted k-earest neighbor based indoor positioning[C]// Proceedings of the 2022 IEEE 95th Vehicular Technology Conference: (VTC2022-Spring). Piscataway: IEEE, 2022: 1-5.

[作者简介]

陈雪晨（1984– ），女，中南大学电子信息学院副教授，主要研究方向为无线通信理论及系统、室内智能定位。

易嘉旋（1999– ），男，中南大学电子信息学院硕士生，主要研究方向为室内智能定位、无线通信、人工智能。

王霭祥（2000– ），男，中南大学计算机学院硕士生，主要研究方向为联邦学习、无线通信、室内定位。



邓晓衡（1974– ），男，中南大学电子信息学院教授、院长，主要研究方向为无线网络与边缘计算、物联网与大数据、智能车联网、分布式计算与系统。