DOI: 10.19650/j.cnki.cjsi.J2209304

一种基于生成对抗网络与模型泛化的机器人推抓技能学习方法*

吴培良[1,2],刘瑞军[1,2],李 瑶[1,2],陈雯柏[3],高国伟[3]

(1.燕山大学信息科学与工程学院 秦皇岛 066004; 2.河北省计算机虚拟技术与系统集成重点实验室 秦皇岛 066004;
3.北京信息科技大学自动化学院 北京 100192)

摘 要:杂乱环境中机器人推动与抓取技能自主学习问题被学者广泛研究,实现二者之间的协同是提升抓取效率的关键,本文提出一种基于生成对抗网络与模型泛化的深度强化学习算法 GARL-DQN。首先,将生成对抗网络嵌入到传统 DQN 中,训练推动与抓取之间的协同进化;其次,将 MDP 中部分参数基于目标对象公式化,借鉴事后经验回放机制(HER)提高经验池样本利用率;然后,针对图像状态引入随机(卷积)神经网络来提高算法的泛化能力;最后,设计了 12 个测试场景,在抓取成功率与平均运动次数指标上与其他 4 种方法进行对比,在规则物块场景中两个指标分别为 91.5%和 3.406;在日常工具场景中两个指标分别为 85.2%和 8.6,验证了 GARL-DQN 算法在解决机器人推抓协同及模型泛化问题上的有效性。

关键词: 推抓技能学习;生成对抗网络;DQN;模型泛化

中图分类号: TH701 TP242 **文献标识码**: A **国家标准学科分类代码**: 520.201

Robot pushing and grasping skill learning method based on generative adversarial network and model generalization

Wu Peiliang[1,2], Liu Ruijun[1,2], Li Yao[1,2], Chen Wenbai[3], Gao Guowei[3]

(*1. School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China;*
2. The Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province, Qinhuangdao 066004, China;
3. School of Automation, Beijing Information Science and Technology University, Beijing 100192, China)

Abstract: Autonomous learning of robot pushing and grasping skills in the cluttered environment has been widely studied. The cooperation between them is the key to improving grasping efficiency. In this article, a deep reinforcement learning algorithm GARL-DQN based on the generative adversarial network and model generalization is proposed. Firstly, the generated adversarial network is embedded into the traditional DQN to train the coevolution between pushing and grasping. Secondly, some parameters in MDP are formulated based on the goal object, and the hindsight experience replay mechanism (HER) is used for reference to improve the sample utilization of the experience pool. Then, according to the image state, a random (convolution) neural network is introduced to improve the generalization ability of the algorithm. Finally, 12 test cases are designed and compared with the other four methods in terms of grasp success rate and average motion times. In the regular block cases, two indicators are 91.5% and 3.406, respectively. In the daily tool scene, two indicators are 85.2% and 8.6, respectively. These results show the effectiveness of the GARL-DQN algorithm in solving the problems of robot pushing and grasping cooperation and model generalization.

Keywords: pushing and grasping skill learning; generate adversarial network; DQN; model generalization

收稿日期:2022-02-16 Received Date: 2022-02-16

*基金项目:国家重点研发计划(2018YFB1308300)、国家自然科学基金区域联合基金(U20A20167)、北京市自然科学基金(4202026)、河北省自然科学基金(F202103079)项目资助

0 引　　言

近年来,机器学习方法的快速发展使得机器人在杂乱场景中学习自主操作目标物体成为可能。现阶段,部分学者利用强化学习的决策能力与环境交互实现自监督训练。抓取操作与非抓取操作(例如推动)技术通常被分别研究,但为了从其协作效应中获得收益,部分学者开始探讨建立机器人推抓技能的学习策略以达到协同收益最大化,该类方法为复杂场景中机器人操作提供了较好的解决方案。

机器人抓取技能学习研究通常基于数据驱动和分析型两个方面。Deng 等[1] 设计了一种结合吸盘与夹持器的新型机械手,与深度 Q 网络(deep Q network, DQN)结合来完成抓取任务,但其机械手的特殊性限制了应用场景。Liang 等[2] 基于 YCB 数据集提出一种轻量级抓取评估模型,用于解决直接从稀疏点云定位的抓取位置问题。卢笑等[3] 提出了一种基于深度强化学习的两阶段显著性目标检测方法,用于复杂场景下的显著性目标检测速度和精度,同时提出一种分治的训练策略。在行人检测数据集上验证了该算法的泛化能力。Arsalan 等[4] 将抓取生成问题公式化为可变自动编码器,使用抓取评估网络细化抓取动作的采样,使用真实机器人在大型数据集中进行了实验验证。葛俊彦等[5] 设计了一种基于三维目标检测的机器人抓取方法,弥补了二维图像识别引导机器人抓取任务中对视角要求较高的缺陷,但其未考虑目标物体遮挡情况。Hang 等[6] 在薄片物体抓取场景中,利用欠驱动机械手的可重构性完成抓取规划。Lin 等[7] 在迁移学习框架下研究视觉先验知识与抓取操作的关系,探讨了将模型参数从视觉网络迁移到动作预测网络的过程,提高了抓取成功率。李秀智等[8] 提出一种双网络架构的机器人最优抓取姿态检测算法,用于在非结构化场景中抓取问题,可高效精确地计算出目标物体的最优抓取区域,但其使用改进 YOLO 模型,时空复杂度较高。

推动作为一种基本的运动元素,极大地拓展了机器人的操作技能。Sarantopoulos 等[9] 提出一种分割 DQN 用于学习最优推动策略,将其划分为一组子网并对应推动原语操作,最终提高了收敛速度与策略质量。Huang 等[10] 提出了一个深度交互预测网络(deep interactive prediction network, DIPN),用于预测机器人机械手在混乱中推动物体时发生的复杂交互。Jochen 等[11] 回顾总结了机器人推动动作的贡献,在专注于预测被推动物体运动的同时也分析了推动动作的上层应用。Zhou 等[12] 实现了给定位置下控制机械手的接触模式和瞬时物体运动,通过仿真与机器人实验实现高效计算。Shome 等[13] 将 RGB-D 数据输入带有末端夹持器的机械臂中,使其可以利用环境来推动多个立方体对象以打包。Song 等[14] 通过沿同一路径同时推动几个相似的对象来降低训练时长,通过实验证明其可以在减少末端执行器轨迹长度的同时解决困难的重排任务。

在抓取任务中,融入有效的推动可以使得机械臂尽快将目标物体分离出来,为抓取提供必要的操作空间。Chen 等[15] 采用具有连续输出的推动策略与基于规则的抓取策略进行抓取检测,以抓取环境中的多个目标物体。Song 等[16] 基于两阶段对物体进行抓取,首先将物体被从不同角度进行推动,移动到工作台边缘后再进行抓取,实现了滑动推动与抓取的结合。Zeng 等[17] 提出了一种基于无监督的深度强化学习框架,对于任意时刻获取的工作空间的状态图像,通过两个相同的前馈神经网络分别训练推抓动作以达到协同。Yang 等[18] 基于 Zeng 等的训练场景,在得到推抓置信度图后采取二分类协同策略,实现了不可见目标物体的探索。Marios 等[19] 通过横向推动来学习分离对象策略,分别计算目标物体不同方向的离散系数再进行抓取。Novkovic 等[20] 提出一种基于强化学习的主动交互式感知系统,用于场景探索与对象搜索,有效地解决了目标寻找任务。Kurenkov 等[21] 基于模仿学习对机器人的推动进行路径规划以形成闭环操作,直至目标物体在杂乱场景中完全显现出来,使其在缺少杂波条件的前提下实现杂乱环境中对目标物体的精准抓取。

传统强化学习中将训练好的模型用于新任务中往往不尽如人意。Al-Shanoon 等[22] 提出了一种深度强化学习的机器人抓取策略,使用较少时间和有限的简单对象进行训练,在真实场景中有效地泛化到新对象中。但对于不同任务场景的成功率并未得到验证。Xu 等[23] 提出了一种具有高样本效率、以目标为条件的分层强化学习公式,来学习推动与抓取策略,最终实现对目标物体的抓取。但非端到端的算法模型在不同层之间的子目标不同,影响最终目标函数收敛。

针对杂乱环境中机器人的推抓技能学习任务,本文提出一种基于生成对抗网络与模型泛化的深度强化学习算法(GARL-DQN)。首先,将生成对抗网络嵌入到传统 DQN 中,将推动网络作为生成器来辅助抓取,抓取网络作为判别器判断当前状态是否可以抓取;其次,使用优先级经验回放机制对不同价值的经验赋予不同的概率;最后,通过引入针对图像状态的随机(卷积)神经网络来提高 GARL-DQN 算法的泛化能力。

1 问题描述与求解框架

1.1 推动与抓取任务描述与符号表示

作为典型的非结构化场景,杂乱场景的数学建模困难。本文将该环境中的推抓协同问题建模为一个条件化

的马尔科夫决策过程(Markov decision process, MDP)。此 MDP 由元组 $(S,A,R,\boldsymbol{P},\gamma,\boldsymbol{p}(s_0))$ 构成,S 为 RGB-D 相机采集到的任意时刻图像构成的状态集合;A 为包含推抓动作的有限动作集;R 为针对执行动作设计的奖励函数集合:$S\times A\rightarrow R$;$\boldsymbol{P}(s_{t+1}|s_t,a_t)$为由当前状态转移到下一时刻状态的概率构成的状态转移矩阵;$\gamma\in(0,1]$为未来奖励的折扣因子;$\boldsymbol{p}(s_0)$为与初始状态有关的分布函数。本文研究面向目标的抓取任务,故将传统 MDP 中决策、奖励以及 Q 值分别基于目标 g 表示为:$\boldsymbol{\pi}(s_t|g)$,$R(s_t,a_t,g)$,$Q_\pi(s_t,a_t,g)$。

1.2 问题的求解框架

本文提出一种机器人自监督学习方法 GARL-DQN,用于训练杂乱场景中机器人推抓之间的协同。首先,将两个 RGB-D 相机采集到的当前环境中的图像状态信息 s_t 送入经验池 B 中,并通过重力方向正投影构建 RGB、Depth 以及 Mask 目标掩码高度图。其次,将上述高度图经过特征提取网络进行特征提取,将提取到的特征经过随机网络层处理,以提高该算法的泛化能力。然后,将特征输入到推动与抓取网络中用于生成推动与抓取功用性图。最后,将抓取网络作为判别器,推动网络作为生成器,评估当前状态是否可以对目标物体执行抓取,以便在推动与抓取之间进行选择。两个网络交替训练,提高训练速度。基于 GARL-DQN 的深度强化学习机器人操作技能框架如图 1 所示。

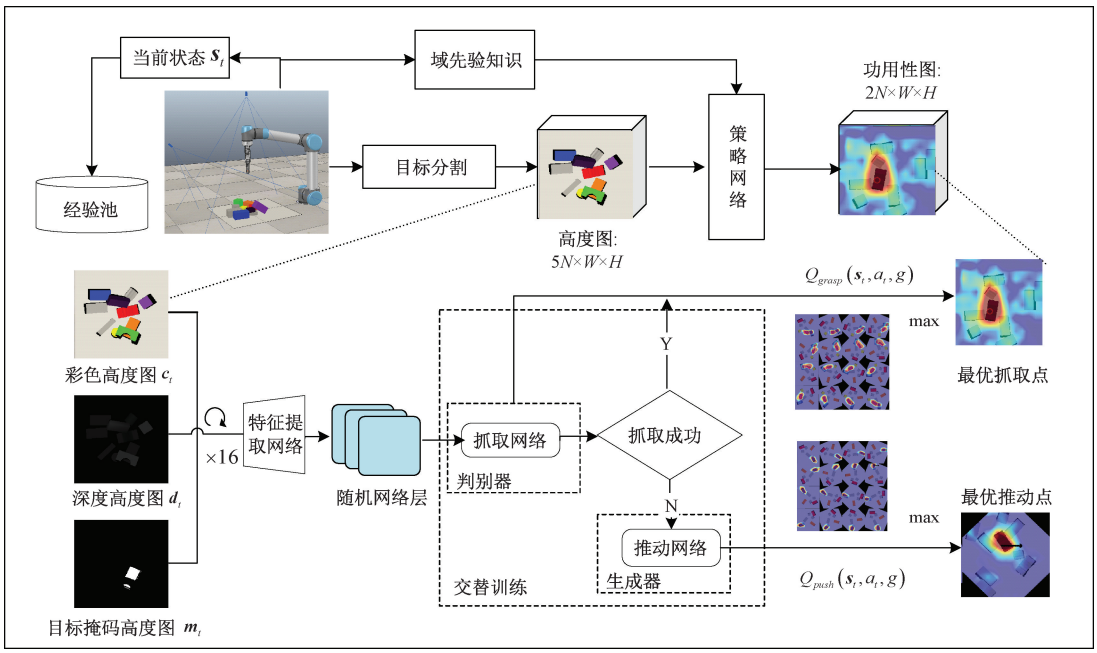

图 1 基于 GARL-DQN 的深度强化学习机器人操作技能框架

Fig. 1 Deep reinforcement learning robot operation skill framework based on GARL-DQN

2 GARL-DQN 算法设计

2.1 GARL-DQN 的泛化模型建模

模型泛化与迁移学习密切相关,用于从源任务转移知识来提高目标任务的性能。然而,强化学习与监督学习不同的是,将源任务上预先训练的模型进行微调以适应目标任务往往是无益的。因此,本小节构建了一种随机卷积网络来增强 GARL-DQN 算法的泛化能力。

引入一个随机网络 $\boldsymbol{f}$,将其先验参数初始化为 $\boldsymbol{\varphi}$,强化学习算法模型的初始状态 s_0 经过网络层 $\boldsymbol{f}$ 处理后得到 $\hat{s}_0=\boldsymbol{f}(s_0;\boldsymbol{\varphi})$。在该任务中,会将初步提取后的当前状态 $\hat{s}_0$ 输入到一个卷积神经网络中,同时保证输入与输出特征维度相同。在每一轮迭代后,该网络都会重新初始化网络 $\boldsymbol{f}$ 的权重,使其可以在有噪声的特征空间上学习。网络参数的选取方式由式(1)中的混合分布表示。其中 $\boldsymbol{I}$ 为卷积核,$\alpha\in[0,1]$ 为常数,n_{in} 与 n_{out} 为输入输出维度,N 表示正态分布。

$$P(\boldsymbol{\varphi})=\alpha\cup(\boldsymbol{\varphi}=\boldsymbol{I})+(1-\alpha)N\left(\boldsymbol{0};\sqrt{\frac{2}{n_{in}+n_{out}}}\right) \tag{1}$$

机器人抓取任务的目标是获得一个最优动作价值函数,通过最小化 TD 损失函数来优化价值网络 Q 的参数 $\boldsymbol{\theta}$,将当前时刻状态与下一时刻状态随机化后分别表示为 $\hat{s}_t=\boldsymbol{f}(s_t;\boldsymbol{\varphi})$ 和 $\hat{s}_{t+1}=\boldsymbol{f}(s_{t+1};\boldsymbol{\varphi})$,可得损失函数计算公式如式(2)所示。

$$L_{value}^{random}(\boldsymbol{\theta}_t) = E_{(s_t,a_t,r_t,s_{t+1},g)\sim B}\left[\frac{1}{2}(R(\hat{s}_{t+1}) + \gamma \max_{a_{t+1}} Q_{target}(\hat{s}_{t+1},a_{t+1};\boldsymbol{\theta}_{t-1}) - Q_{predict}(\hat{s}_t,a_t;\boldsymbol{\theta}_t))^2\right] \quad (2)$$

其中,B 表示经验池,用于提高样本利用率。将不同时刻状态分布之间的特征匹配损失(feature matching loss, FML)作为额外损失,用来限制值网络对于原始特征图以及随机化处理后的特征图采取动作的相似程度,计算方式如式(3)所示。

$$L_{FM}^{random} = E_{(s_t,a_t,r_t,s_{t+1},g)\sim B}[\,\|Q(f(s_t;\varphi),a_t;\boldsymbol{\theta}) - Q(s_t,a_t;\boldsymbol{\theta})\|^2] \quad (3)$$

将总损失定义为二者之和,其中 $\beta > 0$ 是超参数,计算公式如式(4)所示。

$$L^{random} = L_{value}^{random} + \beta L_{FM}^{random} \quad (4)$$

2.2　GARL-DQN 抓取网络目标重标记策略

为了实现推抓之间的协同,在训练环境中,机器人通过 RGB-D 相机采集到当前时刻的图像状态信息,分别经过视觉特征提取网络与随机卷积网络 f 提取特征,作为抓取网络算法的输入。该算法是异策略算法,将目标策略与行为策略分开训练,在保证探索的同时求得全局最优解。

将面向目标的抓取网络表示为 ϕ_g,在训练场景中随机指定目标物体 g 并将抓取奖励表示为 R_g,R_g 的定义方式如下:

$$R_g = \begin{cases} 1, & \text{成功抓取目标物体} \\ 0, & \text{未成功抓取目标物体} \end{cases} \quad (5)$$

$R_g = 0$ 可分为如下两种情况:若机器人未抓取到任何物体则认为是失败的回合,不存入经验池中;若机器人抓取到非目标物体或者为移动遮挡物所做的抓取动作(如图 2 所示),则基于“目标重标记”策略,将非目标物体标记为 g' 并将样本元组 (s_t,a_t,r_t,s_{t+1},g) 转换为 $(s_t,a_t,r_t',s_{t+1},g')$ 存储到经验池中以提高样本效率,其中 $r':=R(s_t,a_t,g')$。经过训练,抓取网络会将抓取 Q_g 值稳定在一个特定值 Q_g^* 上,以此作为抓取判定阈值。

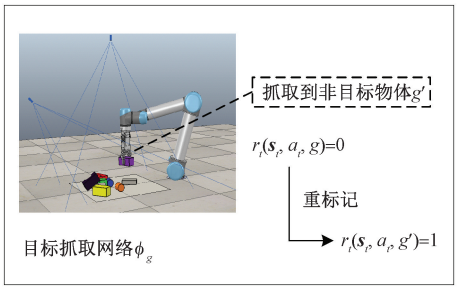

图 2　抓取网络目标重标记策略

Fig. 2　Goal relabel policy of Grasp Net

2.3　GARL-DQN 的推动网络建模

本文将推动动作作为抓取动作的辅助动作,目标为减小目标物体周围的“空间占有率”。但本文的目标为减少机器人的总运动次数,故应尽可能地降低推动动作的频率。考虑到机器人推抓之间的相互作用复杂且耦合度较高,故将基于目标的推动网络 ϕ_p 作为生成器,使得动作价值函数 Q 值不断逼近抓取网络学习到的阈值 Q_g^*,由抓取网络作为判别器来判断当前状态是否适合抓取。推动网络训练目标如式(6)所示,$D_g(\hat{s}_t)$ 的定义如式(7)所示,η 的定义如式(8)所示。采取最优动作得到的下一时刻状态如式(9)所示,其中 T 为状态转移函数,即 $\hat{s}_{t+1} = T(\hat{s}_t,a_t)$。

$$\min_{\phi_p} V(D_g,G_p) = E[\log(1 - D_g(G_p(\hat{s}_t,g)))] \quad (6)$$

$$D_g(\hat{s}_t) = \begin{cases} 1, & \eta > 0 \\ -\eta, & \eta < 0 \end{cases} \quad (7)$$

$$\eta = \max_{a_t} \phi_g(\hat{s}_t,a_t,g) - Q_g^* \quad (8)$$

$$\hat{s}_{t+1} = G_p(\hat{s}_t,g) = T(\hat{s}_t,\operatorname{argmax}_{a_t}\phi_p(\hat{s}_t,a_t,g)) \quad (9)$$

综上,推动是为了使得抓取 Q 值最大化以达到阈值 Q_g^*。基于以上分析,将推动奖励函数设置为:

$$R_p = \begin{cases} 0.5, & Q_g^{\text{improved}} > 0.1 \text{ 并检测到环境变化} \\ -0.5, & \text{环境未变化} \\ 0, & \text{其他} \end{cases} \quad (10)$$

其中 $Q_g^{\text{improved}} = Q_g^{\text{after pushing}} - Q_g^{\text{befor pushing}}$,检测到环境变化是指目标物体周围的环境结构发生变化,并且目标物体的空间占有率减少量 $o_g^{\text{decreased}} > 0.1$,$o_g^{\text{decreased}}$ 表示由高度图计算的目标物体周围非目标物体占有像素的减少量。同时,为了提高样本利用率,在推动网络中也引入目标重标记机制。在目标物体被其他物体遮挡的情况下,如果机器人抓取了非目标物体,则将整个回合中一系列推动动作的目标设置为被抓取的物体,表明该推动序列有利于该非目标物体的抓取。目标推动网络建模如图 3 所示。

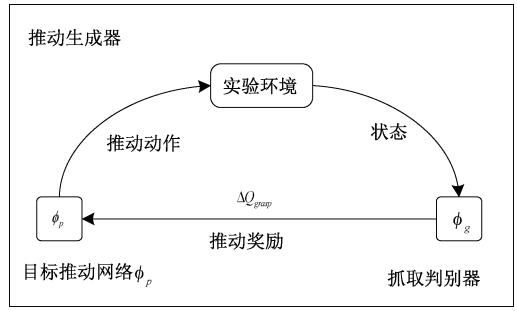

图 3　基于目标的推动网络模型

Fig. 3　Push Net model based on goal

2.4 GARL-DQN 的生成对抗网络建模

本小节给出抓取网络 ϕ_g 与推动网络 ϕ_p 之间的生成对抗网络建模,使得该算法可以更好地拟合出动作参数,学习到最优的推抓位置参数 $p(x,y,z)$ 与角度参数 ω。基于两个网络之间的零和博弈,将目标设置为一个状态的收益分布而不是收益均值,将平均回报向量转化为回报分布函数。将动作价值函数 $Q_\pi(\hat{s},a)$ 表示为 $Z_\pi(\hat{s},a)$ 随机变量,建立期望值与期望函数之间的关系:$Q_\pi(\hat{s},a)=E(Z_\pi(\hat{s},a))$,将定义在分布上的贝尔曼算子表示为 f^π,最终得到贝尔曼方程如式(11)与(12)所示。

$$f^\pi Z_\pi(\hat{s},a)=R(\hat{s},a)+\gamma Z_\pi(S',A') \tag{11}$$

$$Z_\pi(\hat{s},a)=R(\hat{s},a)+\gamma Z_\pi(S',A'),\forall \hat{s}\in S \quad \forall a\in A \tag{12}$$

生成器即推动网络(push net, PN) $PN:Z\rightarrow X$ 为一种映射,该映射从高维噪声空间 $Z=\mathbb{R}^{d_z}$ 找到状态特征,并将其转化为一个输入状态空间 X,且基于该输入状态空间定义如式(13)所示的目标分布 f_X,生成器的任务为拟合观测数据与 f_X 之间的潜在分布。

$$f_X(\boldsymbol{x})=r(\hat{s},a)+\gamma \max_a G(\boldsymbol{z}|(\hat{s}',a)) \tag{13}$$

判别器即抓取网络(grasp net, GN) $GN:X\rightarrow\{0,1\}$ 为该时刻的状态打分,以此判断当前状态是来自真实数据分布 f_X 或是生成器 PN,在本任务中抽象为基于抓取标准阈值对当前状态进行打分。两个网络交替训练更新参数。

本文目标为最小化输出与真实分布之间的距离。一方面,推动网络的目标为产生最优状态-动作值分布的现实样本,即 $Z_{\pi^*}(\hat{s},a)$ 的估计值。另一方面,抓取网络旨在将真实样本 $f(Z(\hat{s},a))$ 与从推动网络输出的样本 $Z(\hat{s},a)$ 进行对比,判断当前时刻状态是否达到抓取阈值。在每个回合中,推动网络接收当前时刻状态 $\hat{s}$ 作为输入,在对分布 $Z(\hat{s},a)$ 估计中的每个动作返回一个样本 $PN(\boldsymbol{z}|(\hat{s},a))$,执行最优动作 $a^*=\max_a PN(\boldsymbol{z}|(\hat{s},a))$。然后,机器人接收环境奖励值并转换到状态 $\hat{s}'$,将元组 $(\hat{s},a,r,\hat{s}')$ 保存到经验池 B 中。每次更新时,从经验池均匀采样,并根据公式更新抓取网络和判别网络。

$$\min_{PN}\max_{GN} \underset{x\sim f_X(x)}{E}[GN(\boldsymbol{x})]+\underset{x\sim PN(z)}{E}[-GN(\boldsymbol{x})] \tag{14}$$

将 $\boldsymbol{x}=r+\gamma\max_a G(\boldsymbol{z}|(\hat{s}',a))$ 定义为真实样本。鉴别网络 GN 的目标是区分上述真实分布值与生成网络 PN 所产生的输出之间的差异,即判断当前状态是否适合抓取。目标函数如式(15)所示。式中的 $\boldsymbol{\omega}_{GN},\boldsymbol{\omega}_{PN}$ 分别为抓取网络与推动网络的权重矩阵,根据 $\boldsymbol{\omega}^{(t+1)}\leftarrow\boldsymbol{\omega}^{(t)}-\alpha_t\nabla_{\boldsymbol{\omega}^{(t)}}L(\boldsymbol{\omega}^{(t)})$ 进行更新。

$$L(\boldsymbol{\omega}_{GN},\boldsymbol{\omega}_{PN})=\begin{cases}\underset{(\hat{s},a,r,\hat{s}')\sim B}{E}[GN_{\boldsymbol{\omega}_{GN}}(\boldsymbol{x}|(\hat{s},a))]-\\ \underset{(\hat{s},a)\sim B}{E}\ \underset{\substack{X\sim PN_{\boldsymbol{\omega}_{PN}}(z|(\hat{s},a))\\ z\sim N(0,1)}}{E}[GN_{\boldsymbol{\omega}_{GN}}(X|(\hat{s},a))]\\ \underset{(\hat{s},a)\sim B}{E}\ \underset{\substack{X\sim PN_{\boldsymbol{\omega}_{PN}}(z|(\hat{s},a))\\ z\sim N(0,1)}}{E}[-GN_{\boldsymbol{\omega}_{GN}}(X(\hat{s},a))]\end{cases} \tag{15}$$

本节将 GAN 思想与传统强化学习算法 DQN 进行结合,实现了推动网络与抓取网络之间的零和博弈,并在算法中加入模型泛化思想,旨在提高强化学习算法在不同任务上的泛化能力。GARL-DQN 操作技能学习伪代码算法流程如算法 1 所示。

算法 1 GARL-DQN 操作技能学习算法流程

输入:MDP 回合次数 M, 判别网络 Grasp Net(GN)和更新次数 n_g, 生成网络 Push Net(PN)和更新次数 n_p,学习率 α,梯度惩罚系数 λ,批量大小 m,先验分布 $P(\boldsymbol{\varphi})$.

输出:执行动作 a_t,Q_{t+1}^*

初始化容量为 N 的经验池 B,GN、PN 网络参数,初始分布 Z,初始状态 $\hat{s}_0$ 及动作 a_0.

$t\leftarrow 0$

For episode = 1 to M do

从先验分布 $P(\boldsymbol{\varphi})$ 抽取随机网络参数 $\boldsymbol{\varphi}$

For time = 1 to $T_{\max}$ do

采样 $\boldsymbol{z}\sim N(0,1)$

$a_t\leftarrow\max_a PN(\boldsymbol{z}|(\hat{s}_t,a))$

采样 $\hat{s}_{t+1}\sim P(\cdot|(\hat{s}_t,a_t))$

将样本元组 $(\hat{s}_t,a_t,r_t,\hat{s}_{t+1},g)$ 放在经验池 B 中

{更新判别网络 Grasp Net (GN)}

For step = 1 to n_g do

在经验池 B 中进行最小批次采样 $\{\hat{s},a,r,\hat{s}',g\}_{i=1}^m$

采样 $\{\boldsymbol{z}\}_{i=1}^m\sim N(0,1)$

定义 $y_i=\begin{cases}r_i, & \hat{s}' \text{ 为结束状态}\\ r_t+\gamma\max_{a_i} PN(\boldsymbol{z}_i|(\hat{s}'_i,a_i)), & \text{其他}\end{cases}$

随机抽取一批样本 $\{\boldsymbol{\varepsilon}\}_{i=1}^m\sim N(0,1)$

$\tilde{\boldsymbol{x}}_i\leftarrow\boldsymbol{\varepsilon}_i y_i+(1-\boldsymbol{\varepsilon}_i)\max_{a_i} PN(\boldsymbol{z}_i|(\hat{s}'_i,a_i))$

$L^{(i)}\leftarrow GN(PN(\boldsymbol{z}_i|(\hat{s}_i,a_i^*))|(\hat{s}_i,a_i^*))-GN(y_i|(\hat{s}\hat{s}_i,a_i^*))+\lambda(|\nabla_x GN(\tilde{\boldsymbol{x}}_i|(\hat{s}_i,a_i^*))|-1)^2$

$\boldsymbol{\omega}_{GN}\leftarrow \text{Adam}(-\nabla_{\boldsymbol{\omega}_{GN}}\frac{1}{m}\sum_{i=1}^m L^{(i)},\alpha)$

在时间步长 T 内更新动作价值网络 $Q(f(s_t;\varphi),a_t;\boldsymbol{\theta})$

计算置信度 Q 值 $Q_{t+1}^*=R_{a_t}(s_t,s_{t+1})+\gamma\max_a(s_{t+1},a;\boldsymbol{\omega})$

计算误差期望值

End For

{更新生成网络 Push-Net(PN)}

For step = 1 to n_p do

采样 $\{z^{(i)}\}_{i=1}^{m} \sim N(0,1)$

$\omega_{PN} \leftarrow \mathrm{Adam}(-\nabla_{\omega_{PN}} \frac{1}{m}\sum_{i=1}^{m} L^{(i)}, \alpha)$

在时间步长 T 内更新动作价值网络 $Q(f(s_t;\varphi), a_t;\theta)$

计算误差期望值

End For

优化关于 θ 的损失函数 $L^{random} = L_{value}^{random} + \beta L_{FM}^{random}$

End For

End For

3　实验与结果

3.1　实验环境搭建

为验证本文算法对于机器人目标物体抓取任务的性能,在 V-REP 3.5.0 动力学仿真软件中模拟该任务的实验场景。该软件内部的运动学模块可准确地模拟真实机器人的运动轨迹,同时具有重力等物理引擎可模拟真实物体属性。使用 RGB-D 相机采集工作空间状态信息,该相机可以提供 RGB 图像以及每个像素的深度信息,并将深度值快速转换为点云信息用于 3D 感知。

仿真实验环境如图 4 所示。工作空间中配置了装有 RG2 夹具的 UR5 机械臂模型,并在工作空间正上方与斜上方 45°的位置均安装 RGB-D 相机,该相机会在每次机械臂执行完动作后采集图像信息,提供完整且大小为 640×480 的深度信息。

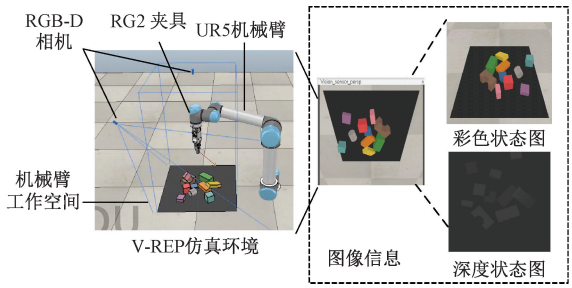

图 4　仿真环境场景

Fig. 4　Simulation environment scene

本文使用的硬件配置为 3.6 GHz Intel Core i9-9900k CPU 和 NVDIA 2070S GPU,操作系统为 Ubuntu18.04 LTS,V-REP 的版本为 3.5.0 的教育版本,采用 0.4 版本的 PyTorch 框架来训练网络模型。

3.2　训练阶段

为验证推动与抓取操作之间的协同性,工作空间中被随机初始化为 m 个随机目标块和 n 个不同形状的基本块,目标块形状与颜色随机匹配,在前 1 000 回合中基本块的个数为 3,后 1 500 回合训练中基本块个数为 8。并将该算法与如下基线方法进行比较。

RAND:不经过监督训练而采取随机像素点抓取。

Grasp-Only:是一种贪婪的确定性抓取策略,它使用单个 FCN 网络进行抓取,该网络使用二分类(来自试错)的监督。在此策略下的机器人仅执行抓取动作。

VPG[17]:在输入中通过添加目标掩码来学习面向目标的推动与抓取策略,是一种使用并行结构作为目标不可知任务预测推动与抓取的功用性图的方法,在目标掩码中根据最大 Q 值执行推动或抓取动作。

GIT[18]:一种深度强化学习方法,使用目标分割网络提取特征来增强机器人感知,基于 DQN 二分类器进行机器人推动与抓取训练。

将机器人执行动作的最大阈值设置为 30,当动作数超过阈值或所有目标物体均被成功抓取时,则重置环境。计算每 50 次迭代中的平均成功抓取率($\frac{n_{success}}{n_{total}}$)并绘制曲线。该过程中目标物体被随机指定,训练 2 500 次后绘制训练性能对比如图 5 所示。

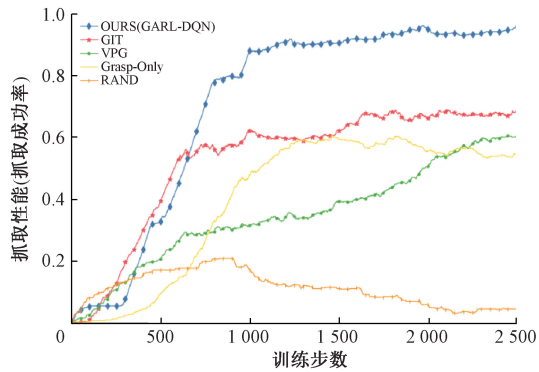

图 5　训练阶段抓取成功率对比

Fig. 5　Comparison of grasp success rate of train stage

从图 5 中可以看出,RAND 面对目标抓取任务时忽略环境而采取随机策略执行动作,抓取成功率极低。Grasp-Only 算法能够完成任务,但没有采取辅助抓取动作,忽略了杂乱环境对目标掩码的影响,导致成功率较低。VPG 方法采用 DQN 强化学习训练推动动作改变环境结构,使目标更好地暴露在工作空间中便于视觉感知。但 VPG 算法对推动动作的训练效率比较低,仅为了改变环境结构而去推动,故成功率在 60% 左右。GIT 使用简单的二分类器对动作进行拟合预测提高动作协同效率,成功率在 60%~70% 之间。本文使用基于 GARL-DQN 的深度强化学习算法,在传统的 DQN 算法框架的基础上融入生成对抗思想,训练推抓网络之间的零和博弈实现协同,进而提高了抓取成功率,使其稳定在 90% 左右。

3.3 测试阶段

测试阶段设置了两种实验场景,与上述 4 种方法进行对比。规则物块场景中目标物体被其他基本块紧紧包围,目标块与训练时相同,用于验证推抓之间的协同;日常工具场景中物体为训练过程中从未见过的物体,用于验证算法的泛化能力。

1)规则物块场景下的算法效率验证

本节设计了如图 6 所示的 8 个测试案例,每个场景包含一个目标物体。对每个案例进行 30 轮实验,若机器人在 5 次内成功抓取目标物体,则记为一轮成功案例,旨在保证抓取成功率的同时,减少平均运动次数。与上述 4 种方法对比如图 7 和 8 所示。由于每个测试场景中目标物体分布不同,故本文算法表现略显不同,表 1 中展示了不同方法的平均性能对比。平均移动次数定义为 $\frac{\sum_{1}^{n}(\text{推}+\text{抓})\text{成功次数}}{n(\text{重复实验次数})}$。

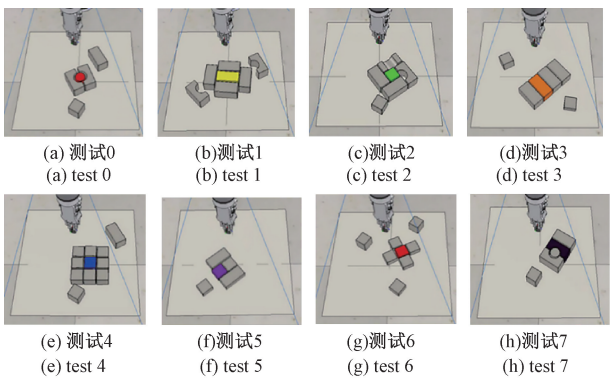

图 6 规则物块的 8 种测试案例

Fig. 6 Eight test cases of regular object blocks

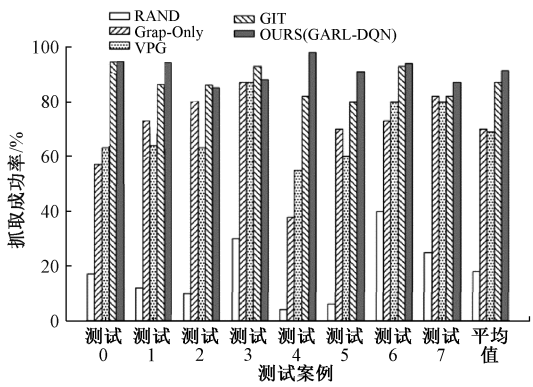

图 7 规则物块的抓取成功率对比

Fig. 7 Comparison of grasp success rate of regular object blocks

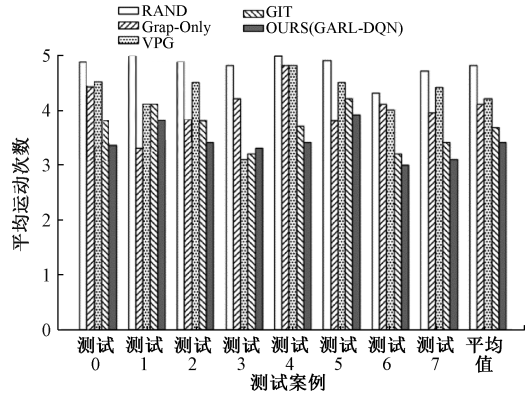

图 8 规则物块的平均运动次数对比

Fig. 8 Comparison of average motion times of regular object blocks

表 1 规则物块案例平均表现

Table 1 Average performance of regular object blocks cases

方法	抓取成功率/%	平均运动次数
RAND	17.5	4.775±0.60
Grasping-only	35.0	4.325±0.98
VPG	70.0	4.025±0.83
GIT	87.5	3.675±0.90
Ours(GARL-DQN)	**91.5**	**3.406±0.50**

2)日常工具场景下的模型泛化能力验证

本节设置了如图 9 所示的 4 个测试案例,每个场景中包含不同高度和形状复杂的日常工具,场景中每个物体被依次设置为目标物体,直接应用训练阶段训练好的模型进行测试,用于验证 GARL-DQN 算法的泛化能力。抓取阈值设置为目标物体的数量。表 2 展示了本方法与其他 4 种方法的平均性能对比。

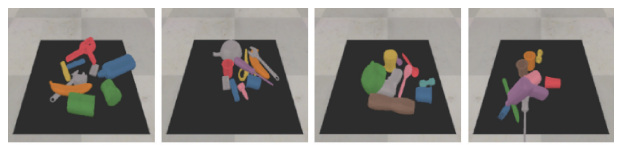

图 9 日常工具的 4 种测试案例

Fig. 9 Four test cases of daily tools

表 2 日常工具案例平均表现

Table 2 Average performance of daily tool cases

方法	抓取成功率/%	平均运动次数
RAND	15.5	15.14
Grasping-only	34.2	12.63
VPG	52.4	10.81
GIT	61.3	9.85
Ours(GARL-DQN)	**85.2**	**8.60**

3)测试阶段结果分析

由规则物块构建的 8 种测试场景模拟了真实世界中目标物体被紧紧包围的场景,由于没有足够空间供抓取,用于检验 GARL-DQN 算法中的推抓协同。表 1 展示了 8 个测试案例的表现,RAND 和 Grasping-Only 两种方法在每个测试案例中都具有较高的运动次数和较低的成功率,抓取成功率在 10%~35%之间,但运动次数在 4.3 以上。VPG 方法对于每个测试案例有不同的表现,可以体现出来推动动作对抓取动作的影响,减少了运动次数,抓取成功率在 60%~75%之间,运动次数在 4.0 左右。GIT 采用二分类器来训练推抓之间的协同作用,每个测试案例的抓取成功率都有提高,同时减少了运动次数,抓取成功率在 85%以上,运动次数在 3.6 左右。而本文采取基于生成对抗网络的 GARL-DQN 训练框架,以 3.4 次的平均运动次数实现了 91.5%的抓取成功率,性能达到最优。

表 2 中展示了日常工具场景中不同算法的表现,用于验证 GARL-DQN 算法的泛化能力。RAND 和 Grasping-Only 两种方法策略的完成率很低,即使能够完成任务,其平均抓取成功率也保持在 15%~30%之间。总体成功率仍然较低,对于日常工具场景的泛化能力依然较弱。RADN 随机选择动作,忽略了杂波环境对目标的影响,从而导致在面对目标运动时出现过多错误动作。Grasping-Only 对于目标物体采取仅抓取策略,虽然对目标周围的杂波环境有一定改善,但影响较小导致该算法成功率较低。VPG 方法仅依靠预测动作的 Q 值选择动作,不能有效判断目标所处的杂波状态,有较多错误抓取动作和冗余推动动作,导致抓取成功率仅在 50%左右,较规则物块场景成功率有明显降低,原因在于其仅依赖 DQN 无法实现良好的算法迁移,当机器人面对新环境时,无法很好地将模型应用在新场景中,故导致抓取率降低。同时,平均运动次数将近 11 次,即无法在一轮中实现全部目标物体的抓取。GIT 使用动作二分类器来协调机器人的推抓动作,不再根据最大 Q 值进行动作选择,目标物体抓取成功率达到了 61%,平均运动次数为 9.85。但其推抓之间的协同性较差,易使得推动动作成为冗余动作。本文将 DQN 与生成对抗网络以及模型泛化思想结合起来形成 GARL-DQN 算法,较好的解决了推动与抓取之间的协同作用,使得平均运动次数减少到 8.6 次,基本在一个回合中抓取到所有目标物体。同时,引入随机网络使得模型泛化能力大大提高,在日常工具场景中的抓取成功率也可达到 85%以上,均优于其他算法。

4　结　　论

众所周知,在服务机器人共融式进化过程中,自主认知与操作杂乱场景下的工具是其必备的学习技能。本文将生成对抗网络和模型泛化思想与基于值函数的 DQN 强化学习算法相结合形成 GARL-DQN 算法,实现了推动与抓取之间协同,同时,提高了强化学习模型的泛化能力。在训练阶段提高了推抓学习性能,在测试阶段提高了推抓成功率,减少平均运动次数,通过实验验证了所提出方法的有效性。但本文算法仍存在一定不足,针对提高模型泛化能力的问题,本文引入的随机卷积网络较为简单,之后的工作中会对最优逼近网络进行探索,使得性能进一步提升。另外,本文在仿真环境中加入的日常工具模型数量有限且模型固定,之后的工作会对于真实场景中多种工具建立更完善的模型并加以实验验证。

参考文献

[1] DENG Y H, GUO X F, WEI Y X, et al. Deep reinforcement learning for robotic pushing and picking in cluttered environment [C]. IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2019, 2019: 619-626.

[2] LIANG H Z, MA X J, LI S, et al. PointNetGPD: Detecting grasp configurations from point sets[C]. 2019 International Conference on Robotics and Automation, ICRA 2019, 2019: 3629-3635.

[3] 卢笑,曹意宏,周炫余,等. 基于深度强化学习的两阶段显著性目标检测[J]. 电子测量与仪器学报, 2021, 35(6) 34-42.
LU X, CAO Y H, ZHOU X Y, et al. Two-stage salient object detection with deep reinforcement learning[J]. Journal of Electronic Measurement and Instrumentation, 2021, 35(6) 34-42.

[4] ARSALAN M, CLEMENS E, DIETER F. 6-dof graspnet: Variational grasp generation for object manipulation [C]. IEEE/CVF International Conference on Computer Vision, 2019: 2901-2910.

[5] 葛俊彦,史金龙,周志强,等. 基于三维检测网络的机器人抓取方法[J]. 仪器仪表学报, 2021, 41(8): 146-153.
GE J Y, SHI J L, ZHOU ZH Q, et al. Robot grasping method based on three-dimensional detection network[J]. Chinese Journal of Scientific Instrument, 2021, 41(8): 146-153.

[6] HANG K Y, MORGAN A S, DOLLAR A M. Pre-grasp sliding manipulation of thin objects using soft, compliant,

or underactuated hands [J]. IEEE Robotics and Automation Letters, 2019, 4(2): 662-669.

[7] LIN Y C, ZENG A, SONG S R, et al. Learning to see before learning to act: Visual pre-training for manipulation [C]. IEEE International Conference on Robotics and Automation, 2020: 7286-7293.

[8] 李秀智,李家豪,张祥银,等. 基于深度学习的机器人最优抓取姿态检测方法[J]. 仪器仪表学报,2020, 41(5): 108-117.
LI X ZH, LI J H, ZHANG X Y, et al. Detection method of robot optimal grasp posture based on deep learning[J]. Chinese Journal of Scientific Instrument, 2020,41(5):108-117.

[9] SARANTOPOULOS L, KIATOS M, DOULGERI Z, et al. Split deep q-learning for robust object singulation[C]. IEEE International Conference on Robotics and Automation, 2020: 6225-6231.

[10] HUANG B C, SHUAI D H, ABDESLAM B, et al. Dipn: Deep interaction prediction network with application to clutter removal [C]. 2021 IEEE International Conference on Robotics and Automation, ICRA 2021, 2021: 4694-4701.

[11] JOCHEN S, CLAUDIO Z, RUSTAM S. Let′s push things forward: A survey on robot pushing[J]. Frontiers in Robotics and AI, 2019, 1(189): 1-11.

[12] ZHOU J J, MASON M T, PAOLINI R, et al. A convex polynomial model for planar sliding mechanics: Theory, application, and experimental validation [J]. International Journal of Robotics Research, 2018, 37(2-3): 249-265.

[13] SHOME R, TANG W N, SONG C, et al. Towards robust product packing with a minimalistic end-effector [C]. IEEE International Conference on Robotics and Automation, 2019: 9007-9013.

[14] SONG C, BOULARIAS A. Object rearrangement with nested nonprehensile manipulation actions [C]. IEEE International Conference on Intelligent Robots and Systems, 2019: 6578-6585.

[15] CHEN Y W, JU Z J, YANG C G. Combining reinforcement learning and rule-based method to manipulate objects in clutter [C]. International Joint Conference on Neural Networks, 2020: 1-6.

[16] SONG C, BOULARIAS A. A probabilistic model for planar sliding of objects with unknown material properties: Identification and robust planning [C]. IEEE/RSJ International Conference on Intelligent Robots and Systems, 2020: 5311-5318.

[17] ZENG A, SONG S, WELKER S, et al. Learning synergies between pushing and grasping with self-supervised deep reinforcement learning[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems, 2018: 4238-4245.

[18] YANG Y, LIANG H Y, CHOI C. A deep learning approach to grasping the invisible[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 2232-2239.

[19] MARIOS K, SOTIRIS M. Robust object grasping in clutter via singulation [C]. Proceedings of IEEE International Conference on Robotics and Automation, 2019: 1596-1600.

[20] NOVKOVIC T, PAUTRAT R, FURRER F, et al. Object finding in cluttered scenes using interactive perception[C]. 2020 IEEE International Conference on Robotics and Automation, 2020: 8338-8344.

[21] KURENKOV A, TAGLIC J, KULKARNI R, et al. Visuomotor mechanical search:Learning to retrieve target objects in clutter [C]. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2020: 8408-8414.

[22] AL-SHANOON A, LANG H X, WANG Y, et al. Learn to grasp unknown objects in robotic manipulation [J]. Springer Science and Business Media Deutschland Gmbh, 2021, 14(4): 571-582.

[23] XU K, YU H X, LAI Q E, et al. Efficient learning of goal-oriented push-grasping synergy in clutter [J]. Institute of Electrical and Electronics Engineers Inc, 2021, 6(4): 6337-6344.

作者简介

吴培良(通信作者),2004 年于燕山大学获得学士学位,2010 年于燕山大学获得博士学位,现为燕山大学教授、博士生导师,主要研究方向为机器人学习、多智能体系统。E-mail: peiliangwu@ysu.edu.cn。

Wu Peiliang (Corresponding author) received his B. Sc. degree and Ph. D. degree both from Yanshan University in 2004 and 2010. He is currently a professor and a Ph. D. advisor at Yanshan University. His main research interests include robot learning and multi-agent system.

刘瑞军,2019 年于河北科技大学获得学士学位,现为燕山大学硕士研究生,主要研究方向为家庭服务机器人工具操作技能学习,强化学习。

E-mail:lrj15733151101@ stumail. ysu. edu. cn

Liu Ruijun received her B. Sc. degree from Hebei University of Science and Technology in 2019. She is currently a master student at Yanshan University. Her main research interest is family service robot tool operation skills learning and reinforcement learning.

陈雯柏,1997 年于东北大学获得学士学位,2004 年于燕山大学获得硕士学位,2011 年于北京邮电大学获得博士学位,现为北京信息科技大学教授,博士生导师,研究方向为机器感知与模式识别。

E-mail:chenwb@ bistu. edu. cn

Chen Wenbai received his B. Sc. degree from Northeastern University in 1997, received his M. Sc. degree from Yanshan University in 2004, and received his Ph. D. degree from Beijing University of Posts and Telecommunications in 2011. He is currently a professor and a Ph. D. advisor at Beijing Information Science and Technology University. His main research interests include machine perception and pattern recognition.