基于视觉推理的机器人多物体堆叠场景抓取方法

张翰博, 兰旭光*, 周欣文, 田智强, 张扬, 郑南宁

西安交通大学电信学院, 人工智能与机器人研究所, 视觉信息处理与应用国家工程实验室, 西安 710049

* E-mail: xgfan@mail.xjtu.edu.cn

收稿日期: 2018-05-28; 接受日期: 2018-08-06; 网络版发表日期: 2018-11-20

国家自然科学基金(批准号: 91748208, 61573268)和国家重点研发计划(编号: 2017YFB1302200)资助项目

摘要 基于视觉的机器人抓取方法是研究智能机器人抓取问题的重要思路. 本文提出了一种基于机器视觉推理的适用于多物体堆叠场景的机器人抓取方法, 算法包含了场景理解和抓取规划两个步骤. 在第一步的场景理解中, 本文的算法包含两个主要部分: 视觉操作关系推理和抓取部位检测. 在视觉操作关系推理中, 本文提出了一种基于深度卷积网络的视觉操作关系网络(Visual Manipulation Relationship Network, VMRN), 以对物体和操作关系进行实时推理. 在视觉操作关系网络中, 通过设计物体对池化层, 实现了物体检测和操作关系推理的端对端训练, 提升了算法的速度和性能. 在第二步的抓取部位检测中, 本文提出了基于有向锚点框的全卷积视觉抓取部位检测网络, 实现了对物体抓取部位的实时检测, 并在康奈尔抓取数据集上取得了目前最高的精度. 在抓取规划中, 通过结合场景深度信息和抓取部位检测结果, 获取当前被抓取物体的抓取点和对应的抓取向量, 并通过坐标系变换将Kinect坐标系的抓取向量映射为机器人坐标系的抓取向量, 完成当前抓取. 实验结果表明, 本文提出的机器人抓取方法能够在多物体堆叠环境下按照正确顺序完成抓取任务, 并成功抓取目标物体.

关键词 机器人抓取, 视觉操作关系推理, 视觉抓取部位检测, 视觉操作关系数据集

1 引言

机器人抓取是机器人研究领域的的一个重要部分, 是机器人与环境交互和完成复杂任务的基础^[1]. 当人类在面临抓取任务时, 通常通过视觉系统进行环境的感知, 定位目标物体, 寻找合适的抓取部位, 并完成抓取. 对人类而言, 抓取动作是一种本能, 但对机器人来说, 由于抓取任务涉及到环境感知和建模、动作规划和执行控制等多个方面, 因此机器人抓取是一个非常具有挑战性的研究方向.

通过结合机器人逆运动学方法, 如何对任务环境

进行有效的感知, 使机器人能够获取目标物体合适的抓取部位, 成为抓取任务中需要解决的重要问题. 近期的研究中^[1~3], 把机器人在抓取任务中的环境感知看作是一类检测问题: 在给定包含目标物体的场景图像中, 检测出物体上最适合机器人抓取的部位.

随着深度卷积网络的发展, 在机器视觉的诸多领域中, 深度学习方法相较于传统算法, 都取得了很大的进展^[4~7]. 原因在于深度神经网络通过监督学习的方式对参数进行训练和调整, 最终使深度网络能够提取有利于特定任务的特征, 以取代手工设计的特征. 这在一定程度上克服了手工设计特征在不同任务中的局

引用格式: 张翰博, 兰旭光, 周欣文, 等. 基于视觉推理的机器人多物体堆叠场景抓取方法. 中国科学: 技术科学, 2018, 48: 1341–1356
Zhang H B, Lan X G, Zhou X W, et al. Robotic grasping in multi-object stacking scenes based on visual reasoning (in Chinese). Sci Sin Tech, 2018, 48: 1341–1356, doi: [10.1360/N092018-00169](https://doi.org/10.1360/N092018-00169)

限性,如泛化能力不足、特征表征能力差等问题^[8]。特别地,在基于机器视觉的抓取部位检测算法的研究中,深度学习方法取得了前所未有的性能^[9,10]。在康奈尔抓取数据集上,检测精度和检测速度都有很大的提升。

在基于视觉的抓取部位检测算法中,文献[1~3,9,10]报道的方法主要面向的是场景中只包含单一目标物体。当场景中包含多个物体时,抓取部位检测算法将会输出场景中置信度最高的抓取部位,而并不考虑物体间的堆叠关系。在适用于多物体堆叠场景的机器人抓取方法中,文献[11~13]报道的方法主要关注的是包含多个可抓取物体的复杂场景下抓取部位和机器人抓取姿态的检测,缺乏对场景中物体种类的理解和物体与抓取部位间从属关系的判断。更重要的是,这些方法并未考虑抓取顺序对物体稳定性和抓取安全性的影响。因此,若用上述方法执行多物体堆叠场景中对特定物体的抓取任务时,存在以下两个重要的问题:如何完成对场景中特定物体的抓取任务;如何保证抓取过程中物体的安全性,使物体不会因跌落、摔倒等动作被破坏。

文献[14]的工作与本文的工作相似,即均在多物体堆叠场景中完成对指定物体的检测和抓取。与本文方法相比,该工作有两个不同点:(1)每次只检测当前抓取的一个物体,并未对场景的全局信息进行感知,因此,在抓取到指定物体之前,该算法并不能判断场景中是否存在该物体;(2)数据量较小,用于训练网络的原始图片仅有352张,因此训练过程容易发生过拟合现象。

为了使机器人能够以安全可靠的方式完成对特定目标物体的抓取,在面向包含多物体的抓取任务场景时,理解任务场景,并获取正确合理的抓取顺序就显得尤为重要。如图1,左图中杯子应先于书被抓取,右图中手机应先于盒子被抓取。如不采用正确的抓取顺序,则可能会对目标物体造成损坏。因此,我们提出了视觉操作关系的概念。视觉操作关系定义在任意一对物体之间,表示机器人在执行多物体抓取任务时,所应遵循的正确顺序。在本文中,正确的顺序定义为先被抓取的物体不应影响后被抓取物体的物理稳定性。通过融合场景中所有物体对的视觉操作关系,可以获得表征整个场景所有物体间视觉操作关系的操作关系

树。结合操作关系树和物体检测,即可以正确的操作顺序完成对多物体堆叠场景中特定物体的抓取任务。

机器人学中的空间关系推理在一定程度上与本文的视觉操作关系相似。空间关系推理通常以环境的点云数据为输入,通过相关算法进行分析,最终获取场景中物体之间的空间关系。文献[15]采用分割算法处理点云,获取分割后的物体。之后使用支持向量机(SVM)提取物体间的相交点网络(Contact Point Network, CPN),获取物体间的空间关系,建立有层级结构的场景表征,以帮助机器人进行物体操作任务。文献[16]使用点云追踪方式使机器人能够理解和重构物体间的空间关系,完成诸如倒水等复杂任务的执行。文献[17]将空间关系推理应用于行为的语义比对和分析。

但空间关系推理与本文的视觉操作关系推理有两个显著的不同点:(1)空间关系推理强调对物体间空间关系的分析,如上、下、左、右、前、后、靠近、远离等,而视觉操作关系则直接表征物体间的操作顺序;(2)空间关系推理往往需要更加复杂的输入,如点云等,而视觉操作关系基于目前的视觉关系,其输入信息则更为简单,通常以RGB图片为输入。

物体检测和视觉关系推理由来已久。近期的一些研究工作表明,深度卷积网络在进行物体检测^[18~20]和物体间的视觉关系推理方面取得了良好的性能^[21~23]。这表明深度卷积网络能有效地提取物体和物体间的视觉关系的相关特征。基于此,本文希望建立一种基于深度卷积网络的视觉操作关系推理算法,来同时完成对场景中待抓取物体的检测以及物体间的视觉操作关系推理的过程。

因此,我们提出了一种基于深度卷积网络的视觉操作关系网络¹⁾(Visual Manipulation Relationship Network, VMRN),通过视觉操作关系网络,能够同时完成场景中物体检测和视觉操作关系推理,并获取场景的操作关系树,以辅助机器人在多物体堆叠场景下执行抓取任务。

在抓取部位检测方面,我们提出了基于有向锚点框的全卷积抓取部位检测网络²⁾。与之前的检测算法^[2,3,9,10]相比,该网络通过在预设锚点框上加入角度信息,以更适应对多角度抓取部位的检测。实验证明,

1) Zhang H B, Lan X G, Zhou X W, et al. Visual manipulation relationship network. arXiv preprint arXiv:1802.08857, 2018.

2) Zhou X W, Lan X G, Zhang H B, et al. Fully Convolutional Grasp Detection Network with Oriented Anchor Box. arXiv preprint arXiv:1803.02209, 2018.

图 1 操作关系的重要性

我们的抓取部位检测算法性能在康奈尔抓取数据集上优于之前的算法。

本文旨在建立一种基于视觉推理的机器人多物体堆叠场景抓取方法。不同于以往的多物体场景抓取方法中仅考虑抓取部位和抓取姿态，而忽略所抓取物体的种类和安全的抓取方式，本文的抓取算法是在对环境进行感知、理解的基础上，帮助机器人在包含多个物体的堆叠场景中，以安全、可靠的方式完成对指定物体的抓取任务，实现机器人的自主作业。因此，本文通过结合视觉操作关系推理、抓取部位检测和机器人运动规划与控制，首次提出了基于视觉推理的多物体堆叠场景抓取方法。本文的研究结果为智能机器人的发展提供了基础，使机器人能够适应于多物体堆叠场景的抓取任务。

2 基于视觉推理的机器人多物体堆叠场景抓取方法

2.1 总体框架

本文提出的基于视觉推理的机器人多物体堆叠场景抓取方法如图2所示。在一次抓取过程中，首先通过

视觉操作关系网络和抓取部位检测网络完成环境感知过程，得到场景中的物体检测、物体间操作关系推理和抓取部位检测结果；在此基础上，通过中心点匹配算法完成物体与抓取部位的匹配；通过结合场景深度信息，计算当前被抓取物体的抓取点和抓取向量；通过坐标系变换，完成相机坐标系内的抓取点和抓取向量到机器人坐标系内抓取点和抓取向量的映射，完成抓取动作。

2.2 视觉操作关系推理

2.2.1 基于视觉的物体检测与操作关系推理算法

视觉操作关系推理包含了物体检测和操作关系推理。其中，物体检测被定义为以包含多物体的图像为输入，在图像中定位物体的位置，并识别物体的类别。操作关系推理被定义为以包含多物体的图像为输入，推理图像中任意物体之间的操作关系并生成操作关系树的过程。在操作关系树中，节点表示物体，边表示物体间的操作关系，节点间的父子关系表示操作所应遵循的先后顺序，其中子节点表示应先被抓取的物体。特别地，对于操作关系树来说，叶节点表示应当场景中应被最先抓取的物体。图3为操作关系树的示例。

物体间的操作关系可以看做是一类特殊的不包含语义信息的视觉关系。我们使用卷积神经网络构建视觉操作关系推理算法，以同时完成物体检测和操作关系推理的过程。视觉操作关系网络结构如图4(a)所示。检测过程分为两个阶段：第一阶段进行物体检测；第二阶段结合物体检测结果和卷积特征完成视觉操作关系推理。网络包含三个部分：特征提取器 Φ 、物体检测器 Ω 和操作关系推理器 Θ 。其中，特征提取器通过卷积

图 2 基于视觉推理的机器人多物体堆叠场景抓取方法

图 3 操作关系树

图 4 视觉操作关系推理网络和抓取部位检测网络图。(a) 视觉操作关系推理网络; (b) 抓取部位检测网络

网络对输入图像进行卷积特征提取; 物体检测器以多尺度卷积特征为输入, 完成对场景中物体的检测; 操作关系推理器以物体对池化后的物体对特征为输入,

完成对任意物体对的视觉操作关系的分类, 并通过对场景中所有物体对视觉操作关系的融合, 推理出整个场景的操作关系树。

2.2.2 物体对池化

为实现视觉操作关系网络的端对端训练, 我们通过物体对池化实现操作关系推理网络梯度的反向传播. 如图4(a)所示, 物体对池化层位于特征提取器 Φ 和操作关系推理器 Θ 之间, 以物体在图像中的位置信息(如物体检测器的输出结果, 或是数据集中物体位置的标定信息)和输入图像的卷积特征图 $CNN(I; \Phi)$ 为输入, 通过遍历场景中所有的物体对 O_i, O_j (n 个物体对应 $n^2 - n$ 个物体对), 进行物体对特征的自适应池化, 获取固定大小的物体对特征 $CNN(O_i, O_j; \Phi)$, 并将所得结果作为一个mini-batch输出. 随后, 操作关系推理器 Θ 以该mini-batch为输入, 对场景中所有物体对的关系进行分类推理, 并将所得结果整合为场景的操作关系树.

2.2.3 训练方法

在物体检测中, 考虑到机器人作业的实时性要求, 文中使用SSD算法^[19]作为物体检测算法. SSD算法使用多尺度特征图进行物体分类和位置回归. 物体检测结果用5维向量表示: $(cls, x_{min}, y_{min}, x_{max}, y_{max})$. 其中 cls 表示物体种类, (x_{min}, y_{min}) 表示物体框左上顶点的坐标, (x_{max}, y_{max}) 表示物体框右下顶点的坐标. 令 D_{OD} 表示用于训练物体检测器的数据, Ω 表示物体检测器的参数, 物体检测的损失函数定义为

$$L_{OD} = L_{loc} + \alpha L_{conf}, \quad (1)$$

其中, L_{loc} 表示标定物体框和预测位置之间的Smooth-L1损失; L_{conf} 表示物体分类的交叉熵损失; α 用来调节置信度损失所占权重.

在操作关系推理中, 由于任意物体对 (O_i, O_j) 的操作关系可分为三种: O_i 先于 O_j ; O_i 后于 O_j ; O_i 的操作与 O_j 无关. 因此, 操作关系推理过程实质上可以看作是一个三分类问题. 给定任意一对物体的特征 $CNN(O_i, O_j; \Phi)$, 通过操作关系推理网络参数 Θ , 将物体对特征映射到分类结果的概率空间中. 在这个过程中, 对换输入物体对的顺序可能会对分类的结果造成影响. 使用 R 表示物体间可能的操作关系, 则分类概率可通过Softmax函数定义:

$$P(R|O_i, O_j; \Theta) = \frac{e^{h_{\Theta}^R(CNN(O_i, O_j; \Phi))}}{\sum_{i=1}^3 e^{h_{\Theta}^i(CNN(O_i, O_j; \Phi))}}. \quad (2)$$

本文使用多分类交叉熵损失作为操作关系推理网

络的损失函数:

$$L_{rp}(R|O_i, O_j; \Theta) = -\ln(P(R|O_i, O_j; \Theta)). \quad (3)$$

为了增强网络的泛化性能, 用于操作关系推理网络训练的数据 D_{RP} 包含两部分: 在线数据 D_{on} 和离线数据 D_{off} , $D_{RP} = D_{on} \cup D_{off}$. 对于任意一张图像, D_{RP} 是包含物体对特征 $CNN(O_i, O_j; \Phi)$ 和操作关系标定 (O_i, R, O_j) 的集合. 其中, R 表示 O_i 和 O_j 的真实操作关系. 在线数据 D_{on} 来自于物体检测器的输出, 而离线数据 D_{off} 来自于数据集标定. 与之相对应, 操作关系推理网络的损失 L_{RP} 包含来自于在线检测数据的损失 L_{on} 和离线数据的损失 L_{off} 两部分:

$$\begin{aligned} L_{RP}(D_{RP}; \Theta) &= \lambda L_{on} + (1 - \lambda) L_{off} \\ &= \lambda \sum_{D_{on}} L_{rp}(R|O_i, O_j; \Theta) \\ &\quad + (1 - \lambda) \sum_{D_{off}} L_{rp}(R|O_i, O_j; \Theta), \end{aligned} \quad (4)$$

其中, λ 用来调整在线数据和离线数据的损失权重.

结合物体检测损失 L_{OD} 和操作关系推理损失 L_{RP} , 视觉操作关系网络的损失函数定义为

$$L(I; \Phi) = \mu L_{OD}(D_{OD}; \Omega) + (1 - \mu) L_{RP}(D_{RP}; \Theta), \quad (5)$$

其中, μ 用来平衡物体检测损失和操作关系推理损失. 在本文实验中, α 取值为1, λ 和 μ 取值为0.5.

2.3 抓取部位检测

2.3.1 基于有向锚点框的全卷积抓取部位检测算法

在抓取部位检测方面, 我们提出了基于有向锚点框的全卷积抓取部位检测算法(Fully Convolutional Grasp Detection based on Oriented Anchor Box, FCGD). 类似于物体检测算法, 本文中的锚点框定义为在目标检测算法中^[18,19]作为参考的预设框, 抓取部位检测结果通过以预设框为基准的偏差量表示, 包括抓取部位的中心点坐标、长、宽和旋转角度.

考虑到抓取部位检测中角度的重要性, 本文在锚点框中引入了预设的角度项. 输入的图片被分为 $N \times N$ 的网格(通常与用于最终的分类和回归的卷积特征图大小相同), 每个网格和 k 个中心点与该网格中心点重合的有向锚点框相关联. 对应于不同的有向锚点框, 在一个网格区域内, 算法可以预测 k 个不同的抓取框.

预设锚点框的角度在 -90° 到 90° 的范围内均匀分布. 例如在 k 等于6的情况下, 每个网格包含6个旋转角, 分别为 75° , 45° , 15° , -15° , -45° 和 -75° 的有向锚点框. 在图5中我们展示了关联多个锚点框的网格. (x_a, y_a) 是网格以及它与它相关联的锚点框的中心坐标, w_a 和 h_a 是有向锚点框预定义宽和高, θ_a 是预定义的旋转角. 给定上述有向锚点框参数及预测得到的抓取框参数, 算法通过以下公式将预测抓取框及标定抓取框映射为锚点框的偏置量:

$$\begin{aligned} t_x &= (x - x_a) / w_a, & \hat{t}_x &= (\hat{x} - x_a) / w_a; \\ t_y &= (y - y_a) / h_a, & \hat{t}_y &= (\hat{y} - y_a) / h_a; \\ t_w &= \log(w / w_a), & \hat{t}_w &= \log(\hat{w} / w_a); \\ t_h &= \log(h / h_a), & \hat{t}_h &= \log(\hat{h} / h_a); \\ t_\theta &= (\theta - \theta_a) / (90 / k), & \hat{t}_\theta &= (\hat{\theta} - \theta_a) / (90 / k), \end{aligned} \quad (6)$$

x, y, w, h 和 θ 分别表示预测抓取框的中心点横纵坐标、宽、高以及旋转角; 变量 $\hat{x}, \hat{x}_a$ 和 $\hat{x}$ 分别表示预测的抓取框、预设锚点框和标定的抓取框(这种表示方法也应用于 y, w, h, w); t_x, t_y, t_w, t_h 和 t_θ 是预测得到的抓取框相对于有向锚点框的偏置; $\hat{t}_x, \hat{t}_y, \hat{t}_w, \hat{t}_h$ 和 $\hat{t}_\theta$ 是标定抓取框相对于和它匹配到的有向锚点框的偏置. 匹配策略将在后续章节详细介绍.

在物体检测算法中, 数据集的目标物体大小的变化范围很大, 为此需要在物体检测锚点框中引入不同的尺度以适应不同尺度的物体. 而受限于固定的拍摄视点及夹持器的大小, 抓取部位尺度的变化在康奈尔抓取数据集中并不明显. 因此在本文的抓取部位检测算法中, 有向锚点框的大小被统一设置为 54×54 . 这一尺度略大于网格面积, 从而保证有向锚点框可以完全覆盖整个图像. 因此在一个网格内, 有向锚点框的大小或长宽比是固定的, 而不同的有向锚点框有不同的旋转角.

图5 关联多个锚点框的网格

2.3.2 网络结构

抓取部位检测网络由特征提取器和多抓取部位预测器构成, 网络结构如图4(b)所示. 本文采用了50层和101层的残差网络作为抓取部位检测网络的特征提取器. 其中, 50层的残差网络有16个残差模块, 101层的残差网络有33个残差模块. 网络以 320×320 像素的RGB图片作为输入, 通过残差网络, 得到2048通道, 大小为 10×10 的特征图.

多抓取预测器由两个组成部分: 分类层及回归层, 两部分均为卷积核大小为 3×3 的卷积层. 回归层的输出通道数为 $5 \times k$, 表示在特征图每一像素上, 所预测的 k 个抓取框相对于 k 个预设锚点框产生的 $5 \times k$ 个偏置量. 分类层的输出通道为 $2 \times k$, 表示对特定位置的 k 个有向锚点框作为抓取部位和非抓取部位的置信度.

2.3.3 匹配策略

在抓取部位检测网络训练的过程中, 需要进行标定抓取框和预设锚点框的匹配, 来训练网络学会修正较接近正确抓取部位的预设锚点框作为最终的抓取部位. 当匹配成功时, 会给有向锚点框一个正标签. 在本文的训练过程, 当一个有向锚点框满足以下条件时会被赋予正标签: (1) 有向锚点框和任意一个标定抓取框的中心在同一个网格内; (2) 有向锚点框和上述同一标定抓取框的角度差应当低于 $90^\circ/k$. 其中, 第一个条件类似于文献[3]中的点度量标准, 这确保了被匹配的有向锚点框位置距离与匹配到的标定抓取框相近. 第二个条件保证了被匹配的有向锚点框与标定抓取框旋转角相近.

由上可知, 较大的 k 值使得被匹配的有向锚点框与标定框的角度差更小, 这使得在网络训练过程中对旋转角的回归更加容易. 然而, 在加大 k 的同时会引入更多的负样本, 加剧正负样本间的不均衡, 增加了对有向锚点框进行分类的难度. 由于在抓取部位的评判标准[1~3]中, 以预测和标定的抓取框相差小于 30° 为预测正确最为常见, 因此将 k 值设为6可以使预设的抓取锚点框之间的角度相差为 30° , 可以较好地全面覆盖所有的标定抓取框. 基于以上的考虑, k 在实验中被设为6.

在本文的匹配策略中, 通过将标定抓取框中心点的横纵坐标分别除以网格的长宽并对结果向上取整, 由此可以得到被匹配的有向锚点框所在的网格. 通过

将标定抓取框旋转角除以 $180/k$ 并向上取整, 由此可以在与该网格相关联的有向锚点框中找到可匹配到的有向锚点框.

在先前的工作中^[2,3,9], 匹配算法通过计算预设框与标定框Jaccard系数, 并将Jaccard系数大于0.5的预设框赋予正标签. Jaccard系数定义如下:

$$J(G, \hat{G}) = \frac{G \cap \hat{G}}{G \cup \hat{G}}, \quad (7)$$

其中, G 为预测抓取框区域, $\hat{G}$ 为标定抓取框区域. 这种方法并没有考虑到角度的重要性, 且计算复杂度较高. 本文的匹配策略以很低的计算复杂度结合了对旋转角与位置两方面的考虑. 在算法性能测试的过程中, 本文与文献^[2,3,9]相同, 使用Jaccard系数作为评价标准.

2.3.4 训练方法

在抓取部位检测网络训练的过程中, 本文采用 $\text{smooth } L_1$ ^[19]作为回归损失. 基于预测偏置量, 回归损失通过下式定义:

$$L_{\text{reg}}(\{t\}) = \sum_{i \in \text{Positive}}^N \sum_{m \in \{x, y, w, h, \theta\}} \text{smooth}_{L_1}(t_m^{(i)} - \hat{t}_m^{(i)}), \quad (8)$$

其中, N 是与所有标定抓取框匹配的有向锚点框的数量, 即正样本数量; t_m 表示由网络预测的偏置; $\hat{t}_m$ 表示相应的标定框偏置.

被赋予正标签的有向锚点框只是全体有向锚点框中很小的一部分, 作为正样本, 其余均为负样本. 在正负样本极不平衡的情况下, 反向传播所有分类梯度将会导致分类损失函数的不收敛. 因此在实验中, 训练过程只回传所有正样本及部分负样本的分类损失. 在本文实验中, 我们将负样本数量设为正样本的三倍, 并通过负样本作为抓取部位的置信度进行排序, 选择前 $3N$ 个作为负样本. 这样做的目的是增大网络对正负样本的区分度. 对于分类问题, 实验中采用交叉熵损失, 定义如下:

$$L_{\text{cls}}(\{p\}) = - \sum_{i \in \text{Positive}}^N \log(p_g^{(i)}) - \sum_{i \in \text{Negative}}^{3N} \log(p_u^{(i)}), \quad (9)$$

其中, $p_g^{(i)}$ 是正样本作为可抓取部位的置信度; $p_u^{(i)}$ 是负样本作为非抓取部位的置信度.

最终, 抓取部位检测网络的损失函数定义为

$$L(\{p\}, \{t\}) = \frac{1}{N} (L_{\text{cls}}(\{p\}) + \alpha L_{\text{reg}}(\{t\})). \quad (10)$$

分类损失和回归损失都通过正样本数 N 进行归一化, 并用 α 来平衡两部分损失. 在实验中, α 设为10.

3 视觉操作关系数据集

3.1 数据采集

在视觉操作关系数据集中, 由于需要训练神经网络同时完成物体检测和操作关系推理, 因此对于数据的采集, 有以下要求:

(1) 所包含的物体必须是可抓取物体, 且在种类和数量上应尽量丰富;

(2) 标定信息需包括物体在图像中的位置信息、物体种类信息和物体间的操作关系.

因此, 本文中的视觉操作关系数据集包含了31类物体, 共5185张图像, 17688个物体实例和51530对操作关系. 数据集分布和数据集图像实例如图6所示. 数据集的标定格式参考PASCAL VOC, 包含了物体的位置信息(物体框左上角和右下角顶点的像素坐标)、物体的种类信息以及每一个物体在操作关系树中的节点索引值、父节点索引值和子节点索引值.

在网络训练和测试的过程中, 数据集被分为训练集和测试集两部分, 其中训练集包含了4656张图像, 15911个物体实例和46934对操作关系.

3.2 数据标定

在视觉操作关系数据集中, 物体位置和种类的标定参照PASCAL VOC的标定标准, 即物体框应尽量紧贴物体边界. 在操作关系标定中, 遵循如下标准: 当移动物体 O_i 会对其他物体的稳定性产生影响时, 则物体 O_i 不应当先被抓取, 即该物体不应成为操作关系树的叶节点.

如图6(b)左上角图像所示, 红色的盒子放置在绿色盒子上, 橘红色的易拉罐在两个盒子旁边. 如果先移动绿色盒子, 则其会对红色盒子的稳定性产生影响, 所以绿色盒子不应成为操作关系树的叶节点. 但是如果先移动红色盒子或橘红色的易拉罐, 则其余物体的稳定性不会受到影响, 因此其为操作关系树的叶节点. 特别地, 橘红色的易拉罐为操作关系树中的孤立节点, 即其没有父节点和子节点.

图 6 视觉操作关系数据集. (a) 数据集分布; (b) 数据集图像示例

4 算法验证

4.1 实验平台

在视觉操作关系网络和抓取部位检测网络的训练和测试过程中, 本文使用的服务器配备四块GPU, 每块GPU均为配备11 GB显存的GTX 1080Ti, 内存为128 GB, 并配备双E5-2670 v3, 12核心24线程CPU. 训练使用的深度学习框架为Torch7.

在机器人实验中, 本文使用了RethinkRobotics公司的Baxter机器人作为机器人实验平台. 机器人实验环境如图7所示. Baxter机器人拥有两个机械臂, 每个机械臂均为7自由度, 臂长104 cm. 在本文实验中, 使用的夹持器为两指平行夹持器. 实验中使用的图像传感器为Kinect V2, 固定于Baxter机器人头部显示器上方, 俯视角约为75°, 距离桌面垂直高度为1 m. Kinect获取的RGB图像分辨率为1920×1080, 并包含了相同

分辨率的深度图. 主计算机的GPU为配备11 GB显存的GTX 1080Ti, 与网络训练和测试过程中所用GPU相同, 主机内存32 GB, CPU为英特尔酷睿i7-4790, 4核心8线程CPU.

4.2 视觉操作关系推理实验

本文使用VGG网络和ResNet50网络的特征提取层作为视觉操作关系网络的基网络, 分别叫做VGG-VMRN和ResNet-VMRN. 由于在网络随机初始化后, 物体检测的输出结果不稳定, 因此在视觉操作关系网络的训练中, 前10000个迭代只使用了离线训练数据 D_{off} .

在视觉操作关系推理实验中, 本文使用视觉操作关系数据集中的测试集进行视觉操作关系网络性能的评估. 测试集中包含了529张图像, 1777个物体实例和

图 7 机器人实验环境

4596对物体关系. 在测试过程中, 本文使用了以下三种评价标准.

(1) 操作关系推理精度: 使用离线的标定数据代替物体检测结果进行操作关系推理精度的评估. 该情况下, 操作关系推理的精度不受物体检测部分精度影响.

(2) 物体对标准: 以物体对为单位, 进行物体检测和操作关系推理. 当两个物体均检测正确, 且操作关系推理正确时, 算作物体对检测正确. 该部分标准包含两个部分: 物体对准确率和物体对召回率.

(3) 图像标准: 以图像为单位, 进行物体检测和操作关系推理. 当单张图像中所有物体均检测正确, 且操作关系推理正确时, 算作图像检测正确.

为了验证物体对池化和加入在线训练数据对算法性能的影响, 本文在不使用物体对池化层进行梯度反向传播(VMRN-无关系梯度)和单独使用在线数据 D_{on} 或单独使用离线数据 D_{off} 进行网络训练的情况下测试了网络性能.

表 1 视觉操作关系网络实验结果

模型	关系梯度	训练数据	物体检测精度(mAP)	操作关系精度(%)	物体对召回率(%)	物体对准确率(%)	图像精度(%)	检测时间(ms)
VGG-SSD-VAM ^[18]	—	$D_{on} \cup D_{off}$	93.01	88.76	75.50	71.28	46.88	>100
ResNet-SSD-VAM ^[18]	—	$D_{on} \cup D_{off}$	91.72	88.76	74.33	75.19	49.72	
VGG-VMRN	无关系梯度	$D_{on} \cup D_{off}$	93.01	88.36	77.28	73.04	50.66	28
	有关系梯度	D_{on}	94.18	92.80	82.64	77.76	60.49	
	有关系梯度	D_{off}	94.36	92.75	81.55	76.09	58.60	
	有关系梯度	$D_{on} \cup D_{off}$	94.09	93.36	82.29	78.01	63.14	
ResNet-VMRN	无关系梯度	$D_{on} \cup D_{off}$	91.72	90.73	77.68	77.55	53.12	
	有关系梯度	D_{on}	91.81	92.01	79.03	72.55	54.44	
	有关系梯度	D_{off}	92.71	91.86	79.33	74.71	55.95	
	有关系梯度	$D_{on} \cup D_{off}$	92.67	92.19	80.55	76.02	57.28	

作为对比, 本文选用Lu等人^[18]的视觉关系推理方法结合SSD物体检测在视觉操作关系数据集上进行了训练和测试(SSD-VAM). VAM以包含物体对的联合框作为网络输入, 对谓语进行分类作为关系推理结果. 但是由于在视觉操作关系数据集中, 交换物体次序会导致结果变化, 因此本文使用物体对中两个物体的物体框和联合框同时作为VAM的网络输入, 进行操作关系推理.

实验结果如表1所示. 与Lu等人的方法相比, 本文在算法精度和速度上都有较大提升. 原因在于本文中视觉操作关系网络通过物体对池化层进行端对端的训练和测试, 因此卷积特征提取部分能更好地适应操作关系推理任务, 且加快了网络前向传播的速度. 而单独使用 D_{on} 或 D_{off} 和不使用物体对池化层进行操作关系网络梯度反向传播的网络精度均低于标准的视觉操作关系网络, 这证明了本文所提出方法的有效性.

图8展示了视觉操作关系推理实验的结果. 从图中可以看出, 视觉操作关系推理网络能够有效地完成场景中物体检测和操作关系推理的任务. 在检测过程中出现的错误则集中在对物体的漏检、多检和错检以及对操作关系的漏检(对操作关系的多检为冗余信息, 不影响机器人操作顺序).

4.3 抓取部位检测实验

抓取部位检测网络在康奈尔抓取数据集上进行了训练和测试. 康奈尔抓取数据集包含了240个可抓取物体, 共计885张图片. 在每一张图片中, 包含多个表示抓取部位的正样本和负样本, 每个样本由抓取框的四个

图 8 视觉操作关系推理结果示例

顶点表示.

和先前的工作一样, 实验中将康奈尔抓取数据集分为包含708张图片的训练集和177张图片的测试集两部分. 实验采用以下两种方式来对数据集进行拆分.

(1) 逐图片拆分: 将图片随机地分到训练集和测试集. 目的是测试网络对于已经见过的物体在新位置和新姿态下的泛化能力.

(2) 逐物体拆分: 将数据集在物体层面上拆分数据集. 所有包含同一物体的图片都被分到相同的集中. 这种拆分是为了测试网络对新物体的泛化能力.

由于康奈尔抓取数据集数据量较小, 因此在网络训练过程中, 需要通过数据增强扩充数据量. 在本文的实验中, 数据增强过程通过图片平移、裁剪、旋转、翻转以及颜色和锐度变换完成, 具体过程如下:

(1) 对图片在 x 和 y 轴各做50个像素点内的随机

平移;

(2) 在平移后的图片中心裁剪 320×320 像素的区域;

(3) 对裁剪后的图片以顺时针或逆时针方向在 15° 的范围内随机旋转;

(4) 对旋转后的图片随机进行水平或竖直翻转;

(5) 对翻转后的图片进行颜色及锐度变换.

通过上述数据增强过程, 最后得到 320×320 像素的图片, 作为抓取部位检测网络输入.

与之前的工作^[2,3,9,10]一样, 抓取部位检测实验采用矩形标准来评估抓取检测结果. 在矩形标准中, 抓取部位检测结果在同时满足以下两个条件时将被判定为正确的检测:

(1) 预测抓取框与标定抓取框旋转角的差小于 30° ;

(2) 预测抓取框与标定抓取框的Jaccard系数要大于预设阈值.

在网络性能测试过程中, 测试集图片的预处理只包含了中心裁剪过程. 实验中是在不同的Jaccard系数阈值下对提出的模型进行测试. 表2展示了不同模型、不同Jaccard系数阈值下的实验结果. 从表中可以看出, Jaccard系数阈值低于25%时, 虽然基于101层残差网络的抓取部位检测网络在逐图片的拆分上有着与基于50层残差网络的抓取部位检测网络近似的精度, 但使用101层残差网络作为特征提取器能够在Jaccard系数阈值提高的情况下依然保证很高的精度. 50层残差网络与101层残差网络的比较也说明使用更深的网络是除引入多尺度与多长宽比之外另一种提升模型在不同物体上的泛化能力的方式.

实际试验中, 对于每张测试图片网络共有600 ($10 \times 10 \times k$, $k=6$)个预测抓取框, 相当于Chu等人³⁾所提出算法预测量的三分之一. 因此, 本文的抓取部位检测算法在使用更少预测量的情况下, 达到了更高的精度. 从表2中可以看出, 基于50层残差网络的抓取部位检测网络在逐图片拆分下的精度超过目前最优算法1.74%, 基于101层残差网络的抓取部位检测网络在几乎所有标准下性能都达到目前最优. 这表明有向锚点框机制为抓取部位检测提供了一种更准确和高效的方法.

图9展示了抓取部位检测实验的主观结果. 其中, 图9(a)可视化了在逐图片拆分下的正确检测结果, 第一列为数据集的正标定, 第二列为检测结果中置信度最高的可抓取部位, 第三列为检测结果中所有置信度大于0.5的可抓取部位. 图9(b)展示了测试集中存在的所有误检结果, 共4个, 其中第一行是检测结果, 第二

行是数据集标定. 图9(c)展示了多物体下算法的泛化能力. 从检测结果可以看出, 算法预测的合理抓取部位中有一部分数据集中并未标定, 说明本文的抓取部位检测算法的泛化能力较强, 并没有发生过拟合现象. 但是这种预测会导致算法评估时精度的下降, 如图9(b)第1, 2, 4列所示, 合理的抓取部位会被判定为错误的检测. 图9(c)说明本文的抓取部位检测算法在多物体堆叠场景中有较好的泛化能力, 能够完成多物体堆叠场景下的抓取部位检测任务.

4.4 机器人多物体堆叠场景抓取实验

在机器人多物体堆叠场景抓取实验中, 本文使用Baxter机器人完成了对文中所述算法框架的验证. 机器人多物体堆叠场景抓取实验的目的是使机器人能够完成多物体堆叠场景下对特定物体的抓取任务, 该特定物体定义为抓取任务的目标物体.

在实验中, 放置物体的桌面位于Baxter机器人正前方. 由于Baxter机械臂工作空间的限制, 实验中在桌面上划分出约40 cm×40 cm的区域作为工作空间, 每次实验中, 待抓取物体被放置在工作空间内, 以便在大部分情况下, Baxter机器人都能够顺利抓取物体. 受限于Baxter机器人夹持器的开合范围, 每次实验中所有物体在抓取部位的尺度方面差距不能过大, 即要求每次实验中的物体可抓取部位的宽度相似, 且需要在每次实验前选定合适的夹持器设置.

机器人平台实验流程如图2中算法框架所示. 实验中使用的视觉操作关系推理网络为有关系梯度的使用在线数据和离线数据同时训练的VGG-VMRN, 而抓取部位检测网络则使用了ResNet101-FCGD网络. 在多物体堆叠场景中完成对特定物体的抓取过程, 主要包括

表2 抓取部位检测实验结果

模型	预设框设置	逐图片拆分(%)				逐物体拆分(%)			
		20%	25%	30%	35%	20%	25%	30%	35%
Guo等人 ^[9]	1种尺度1种长宽比	93.8	93.2	91.0	85.3	85.1	82.8	79.3	74.1
Guo等人 ^[9]	3种尺度3种长宽比	88.1	86.4	83.6	76.8	90.8	89.1	85.1	80.5
Chu等人 ³⁾	3种尺度3种长宽比	—	96.0	94.9	92.1	—	96.1	92.7	87.6
ResNet50-FCGD	单一尺度有向锚点框	98.87	97.74	94.92	89.83	95.48	94.91	89.26	84.18
ResNet101-FCGD	单一尺度有向锚点框	98.31	97.74	96.61	95.48	97.74	96.61	93.78	91.53

3) Chu F J, Vela P A. Deep Grasp: Detection and Localization of Grasps with Deep Neural Networks. arXiv preprint arXiv:1802.00520, 2018

图9 抓取部位检测主观实验结果. (a) 正确检测示例; (b) 误检示例; (c) 多物体示例

以下步骤:

(1) 使用视觉操作关系网络进行物体检测和操作关系推理: 使用Kinect V2采集包含所有待抓取物体的场景图像, 并截取图像中的工作空间相关区域作为视觉操作关系网络输入. 通过视觉操作关系网络, 获取物体在场景图像中的位置、分类信息及场景操作关系树;

(2) 使用抓取部位检测网络检测图像中潜在的抓取部位: 通过物体检测结果截取场景图像中每个物体所占区域, 并通过填充0值像素点将截取区域填充至正方形作为抓取部位检测网络的输入, 检测每个物体区域内潜在的抓取部位;

(3) 使用中心点匹配完成待抓取物体和对应抓取部位的匹配: 对每一个待抓取物体, 计算其所属图像区域内所有抓取部位中心点和区域中心点距离, 选择距离物体区域中心点最近的抓取部位作为该物体的抓取部位;

(4) 使用操作关系树选择当前抓取物体: 有操作关系树中目标物体节点开始, 使用深度优先遍历算法遍历其子节点, 选择第一个叶节点作为当前被抓取物体, 以该物体对应的抓取部位作为当前抓取部位;

(5) 结合场景深度信息, 计算当前机器人抓取向量, 并执行抓取: 以当前抓取部位中深度信息最小的点为抓取点, 计算抓取点附近的表面法向量作为抓取向量(Kinect坐标系); 通过坐标系变换, 将该抓取向量映射为机器人坐标系内的抓取向量; 通过该抓取向量, 确定机器人抓取姿态; 使用机器人逆运动学, 移动夹持器至预抓取部位(沿抓取向量距离抓取点10 cm处), 最后沿抓取向量方向移动夹持器至抓取点, 抓取当前被抓取物体;

(6) 完成当前抓取动作并判断是否需要进行下一次抓取: 判断当前被抓取物体是否为目标物体. 如果是, 则当前抓取任务完成; 如果不是, 则将当前被抓取物体放置在机器人工作空间外的预设位置, 并返回步骤1.

实验结果如表3所示. 在多物体堆叠场景机器人平台的实验中, 本文进行了200次实验. 其中使用操作关系网络VMRN结合抓取部位检测网络FCGD的抓取算法框架实验120次(单物体场景与不包括操作关系网络算法相同), 包含1, 2, 3个物体的场景实验各40次, 包括了夹持器的两种宽度设置. 两种夹持器宽度设置对应的物体种类如图10所示. 作为对比实验, 对只使用抓取部位检测网络FCGD实验了80次, 实验中以包含多物体的场景图片为输入, 对抓取检测结果中置信度最高的抓取部位进行抓取, 抓取过程中同样使用深度信息进行抓取部位从图片坐标系到机器人坐标系的映射. 抓取过程中, 规定能够按照正确的操作顺序成功抓取预设目标物体(实验中设置为场景操作关系树的叶节点物体)为实验成功.

表 3 机器人实验成功率

物体数量		1	2	3
窄夹持器(无关系网络)	实验次数	—	20	20
	成功次数	—	8	2
	成功率	—	40%	10%
宽夹持器(无关系网络)	实验次数	—	20	20
	成功次数	—	9	4
	成功率	—	45%	20%
总计	实验次数	—	40	40
	成功次数	—	17	6
	成功率	—	42.5%	15%
窄夹持器	实验次数	20	20	20
	成功次数	19	16	12
	成功率	95%	80%	60%
宽夹持器	实验次数	20	20	20
	成功次数	20	17	11
	成功率	100%	85%	55%
总计	实验次数	40	40	40
	成功次数	39	33	23
	成功率	97.5%	82.5%	57.5%

图 10 多物体抓取任务示例

可以看到，视觉操作关系网络的引入能够极大增加机器人按照正确顺序抓取物体的概率。原因在于，在未引入视觉操作关系网络时，物体的抓取顺序随机性较大，能够按照正确顺序完成抓取任务的概率较低。但随着场景中物体个数增多，实验成功率有明显下降。原因主要有以下四个方面：

(1) 场景中存在多个物体时，物体间的遮挡和重叠影响物体检测结果的精度；

(2) 抓取部位和物体的匹配使用简单的中心点匹配方法，当物体较多、重叠面积较大时，匹配正确率很难保证；

(3) 视觉操作关系网络和抓取部位检测网络存在泛化误差；

(4) 抓取成功与否受机器人夹持器的灵活性限制。

表4展示了在视觉关系检测中，包含不同物体个数的图片的测试精度，使用的网络为使用离线数据和在线数据训练的有关系梯度的VGG-VMRN。可以看到，物体个数对于以图片为单位的测试精度有较大影响。原因在于图片中包含的操作关系个数随物体个数的增加呈 $O(n^2)$ 的复杂度增加，因此将整张图片中所有操作关系预测正确的难度也随之增加。在之后的工作中，会通过设计新的网络损失函数，以增加对图片整体的预

表 4 图像精度和图片中物体个数的关系

物体个数	2	3	4	5
精度	58/73	173/267	62/116	43/73
精度(%)	79.45	64.79	54.39	58.90

*测试网络为VGG-VMRN, 有关系梯度, $D_{on} \cup D_{off}$

测准确率, 特别是增加对图片中首先被抓取的物体的预测准确率.

而在抓取部位和物体匹配的过程中, 由于使用的算法为中心点匹配算法, 当增加场景中堆叠的物体数量时, 物体边界框重叠面积会增大, 物体中心点数目变多, 距离相近, 通过边界框中心点区分不同物体的抓取部位的过程中, 精度也会迅速下降. 在之后的工作中, 会寻找新的基于物体特征的抓取部位匹配算法完成该匹配过程.

图10展示了部分成功的实验结果, 抓取任务的成功与否受视觉操作关系网络的精度、抓取部位检测精度、待抓取物体和抓取部位的匹配准确率及机器人执行动作的误差和限制条件等多个因素影响.

5 结论与讨论

本文提出了基于视觉推理的机器人多物体堆叠场景抓取算法. 在本文的算法框架中, 通过视觉操作关系网络进行场景物体检测和操作关系推理; 通过全卷积抓取部位检测网络完成对场景中潜在抓取部位的检测; 通过中心点匹配完成抓取部位和物体的匹配过程; 通过结合深度图完成当前抓取物体抓取向量和抓取点的计算; 通过坐标系变换完成抓取向量和抓取点到机器人的映射, 完成机器人抓取.

文中的视觉操作关系网络首次完成了多物体堆叠场景下物体的操作关系推理任务, 并通过物体对池化

完成了操作关系和物体检测的端对端训练; 抓取部位检测网络通过结合有向锚点框机制, 并在训练中使用角度差匹配策略代替重叠面积匹配, 提升了算法的精度和速度, 是目前在康奈尔抓取数据集上精度最高的算法.

但是, 目前视觉操作关系网络受限于数据集中的物体数量, 在物体检测过程中泛化能力需进一步提升. 因此, 在之后的工作中, 我们拟通过两种方式增强现有网络的性能.

(1) 增加数据量, 构建更加丰富的视觉操作关系数据集. 基于现有的深度学习方法对于数据的依赖性, 构建一个完整的、有代表性的、高性能的数据集是一项非常有意义的工作. 目前本文的视觉操作关系数据集包含超过5000张图片, 31类可抓取物体, 超过17000个物体实例. 但仍在数据多样性和复杂性方面有所不足. 因此在未来的工作中, 我们会在各个方面对数据集进行扩充, 构建一个更加丰富的视觉操作关系数据集, 并支持其他研究者进行相关方面的研究.

(2) 精简网络结构, 通过设计特定网络结构提取与操作关系相关而独立于物体种类的特征等方法, 以提升物体检测和操作关系推理中对于新物体和新场景的泛化能力. 此外, 在抓取部位和物体的匹配方面, 会寻找更加高效和准确的匹配方式代替现有的中心点匹配; 在机器人执行抓取时, 与Sergey等人^[13]类似, 我们希望在以纯RGB为输入的环境中完成抓取任务. 但我们更多的关注在于使机器人能够理解环境, 在对环境中物体进行识别和物体操作关系进行预测的基础上, 规划并完成对特定物体的抓取任务, 以建立一种智能的抓取系统, 而并非简单的在多物体场景中选择合适的抓取姿态, 不考虑抓取的对象种类和安全的可靠的抓取顺序. 因此, 我们拟通过强化学习方式训练视觉伺服的抓取策略, 以纯RGB视觉输入完成反馈执行的过程, 代替现有的逆运动学过程.

附录: 训练参数设置

附表 1 训练参数设置

参数名称	批大小	学习速率	迭代次数	学习速率衰减	动量
VGG-VMRN	8	0-80k:0.001, 80-120k:0.0001	120000	0	0.9
ResNet-VMRN	8	0-80k:0.001, 80-120k:0.0001	120000	0	0.9
FCGD	16	0.0001	80000	0.0001	0.9

参考文献

- 1 Jiang Y, Moseson S, Saxena A. Efficient grasping from RGBD images: Learning using a new rectangle representation. In: 2011 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2011. 3304–3311
- 2 Lenz I, Lee H, Saxena A. Deep learning for detecting robotic grasps. *Int J Robotics Res*, 2015, 34: 705–724
- 3 Redmon J, Angelova A. Real-time robotic grasp detection using convolutional neural networks. In: 2015 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2015. 1316–1322
- 4 Ding C, Tao D. Trunk-branch ensemble convolutional neural networks for video-based face recognition. *IEEE Trans Pattern Anal Mach Intell*, 2018, 40: 1002–1014
- 5 Zhang S, Benenson R, Omran M, et al. Towards reaching human performance in pedestrian detection. *IEEE Trans Pattern Anal Mach Intell*, 2018, 40: 973–986
- 6 Varol G, Laptev I, Schmid C. Long-term temporal convolutions for action recognition. *IEEE Trans Pattern Anal Mach Intell*, 2018, 40: 1510–1517
- 7 Levine S, Finn C, Darrell T. End-to-end training of deep visuomotor policies. *J Mach Learn Res*, 2016, 17: 1–40
- 8 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, 521: 436–444
- 9 Guo D, Sun F, Kong T, et al. Deep vision networks for real-time robotic grasp detection. *Int J Adv Robot Syst*, 2016, 14: 1–8
- 10 Wang Z C, Li Z Q, Wang B, et al. Robot grasp detection using multimodal deep convolutional neural networks. *Adv Mech Eng*, 2016, 8: 1–12
- 11 Viereck U, Pas A, Saenko K, et al. Learning a visuomotor controller for real world robotic grasping using simulated depth images. In: 1st Conference on Robot Learning (CoRL). 2017. 78: 291–300
- 12 Gualtieri M, Pas A, Saenko K, et al. High precision grasp pose detection in dense clutter. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2016. 598–605
- 13 Levine S, Pastor P, Krizhevsky A, et al. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *Int J Robotics Res*, 2018, 37: 421–436
- 14 Guo D, Kong T, Sun F C, et al. Object discovery and grasp detection with a shared convolutional neural network. In: 2016 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2016. 2038–2043
- 15 Rosman B, Ramamoorthy S. Learning spatial relationships between objects. *Int J Robotics Res*, 2011, 30: 1328–1342
- 16 Zampogiannis K, Yang Y Z, Fermüller C, et al. Learning the spatial semantics of manipulation actions through preposition grounding. In: 2015 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2015. 1389–1396
- 17 Ziaetabar F, Aksoy E E, Wörgötter F, et al. Semantic analysis of manipulation actions using spatial relations. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2017. 4612–4619
- 18 Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell*, 2017, 39: 1137–1149
- 19 Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector. *Comput Vis ECCV*, 2016: 21–37
- 20 Ren S, He K, Girshick R, et al. Object detection networks on convolutional feature maps. *IEEE Trans Pattern Anal Mach Intell*, 2017, 39: 1476–1481
- 21 Lu C W, Krishna R, Bernstein M, et al. Visual relationship detection with language priors. *Comput Vis ECCV*, 2016: 852–869
- 22 Krishna R, Zhu Y, Groth O, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *Int J Comput Vis*, 2017, 123: 32–73
- 23 Peyre J, Laptev I, Schmid C, et al. Weakly-supervised learning of visual relations. In: 2017 IEEE International Conference on Computer Vision. IEEE, 2017. 5179–5188

Robotic grasping in multi-object stacking scenes based on visual reasoning

ZHANG HanBo, LAN XuGuang, ZHOU XinWen, TIAN ZhiQiang, ZHANG Yang & ZHENG NanNing

Institute of Artificial Intelligence and Robotics, National Engineering Laboratory of Visual Information Processing and Applications, School of Telecommunications, Xi'an Jiaotong University, Xi'an 710049, China

Vision is one of the most important ways to solve robotic grasping problem. We propose a framework that can help the robot grasp target object in multi-object scenes based on visual reasoning, which includes two stages: Perception and execution. During perception stage, our proposed framework includes two parts: Visual manipulation relationship reasoning and robotic grasp detection. In visual manipulation relationship reasoning, Visual Manipulation Relationship Network (VMRN) is proposed to simultaneously detect objects and gets manipulation relationships between each pair of objects. We design Object Pairing Pooling Layer to implement end-to-end training of object detection and visual manipulation relationship reasoning in VMRN, which makes the algorithm faster and more robust. In robotic grasp detection, fully convolutional grasp detection network based on oriented anchor box is proposed to implement real-time grasp detection for any object, which is the state-of-the-art robotic grasp detection algorithm on the standard Cornell Grasp Dataset. During execution stage, by combining depth information and perception results, grasp point and grasp vector in camera coordinate are computed first. Then, they are transformed into robot coordinate for robot to execute grasping motion. Experimental results show that our framework can help robot grasp target in multi-object scene in the right order.

robotic grasping, visual manipulation relationship reasoning, grasp detection, visual relationship manipulation dataset

doi: [10.1360/N092018-00169](https://doi.org/10.1360/N092018-00169)