DOI: 10.19650/j.cnki.cjsi.J1905814

基于深度卷积网络的多目标动态三维抓取位姿检测方法*

杨傲雷[1],曹 裕[1],徐昱琳[1],费敏锐[1],陈 灵[2]
(1. 上海大学机电工程与自动化学院 上海 200444; 2. 湖南师范大学工程与设计学院 长沙 410081)

摘 要:在非结构化环境机器人抓取任务中,获取稳定可靠目标物体抓取位姿至关重要。本文提出了一种基于深度卷积网络的多目标动态三维抓取位姿检测方法。首先采用 Faster R-CNN 进行多目标动态检测,并提出稳定检测滤波器,抑制噪声与实时检测时的抖动;然后在提出深度目标适配器的基础上采用 GG-CNN 模型估算二维抓取位姿;进而融合目标检测结果、二维抓取位姿以及物体深度信息,重建目标物体点云,并计算三维抓取位姿;最后搭建机器人抓取平台,实验统计抓取成功率达到 95.6%,验证了所提方法的可行性及有效性,克服了二维抓取位姿固定且单一的缺陷。

关键词: 深度卷积网络;抓取位姿;目标检测;稳定检测滤波器
中图分类号: TP391 TH86 **文献标识码**: A **国家标准学科分类代码**: 510.4050

Dynamic multi-target 3D grasp posture detection approach based on deep convolutional network

Yang Aolei[1], Cao Yu[1], Xu Yulin[1], Fei Minrui[1], Chen Ling[2]
(*1.School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China*;
2.College of Engineering and Design, Hunan Normal University, Changsha 410081, China)

Abstract: In the robot grasping task in unstructured environment, it is important to acquire stable and reliable grasp pose of the object. In this paper, a dynamic multi-target 3D grasp pose detection approach based on deep convolutional network is proposed. Firstly, the Faster R-CNN is utilized to conduct dynamic multi-target detection, and a stabilization detection filter is proposed to reject the noise and jitter in real time detection. Then, based on proposing depth target adapter, the GG-CNN model is used to estimate the 2D grasp pose. Furthermore, the target detection result, 2D grasp pose and object depth information are fused to reconstruct the point cloud of the object, and calculate the 3D grasp pose. Finally, a robot grasping platform was established. The experiment results show that the statistical grasping success rate reaches 95.6%, which not only verifies the feasibility and effectiveness of the proposed approach, but also overcomes the defect of fixed and single result for 2D grasp pose.

Keywords: deep convolutional network; grasp posture; object detection; stabilization detection filter

0 引 言

机器人在人类生活诸多方面发挥着重要作用,如协同控制[1]、自主避障[2]、辅助抓取[3]等方面。而机器人抓取作为机器人与人类协同工作的基本能力,是完成后续各类复杂自主任务的根本保证。经过多年的研究,机器人抓取在结构化环境中已可取得很好的抓取精度与成功率,例如结构化生产车间中工业机器人程序化抓取单一同类物体,成功率能达到 99% 以上。然而,机器人工作环境一旦切换至非结构化复杂动态场景时,原有的一系列方法都很难适用。因此,有必要针对非结构化复杂环境,进一步研究机器人动态、多物体的抓取问题,计算识别出多物体的三维可抓取位姿。

目前,研究人员已提出许多方法来检测物体可抓取位姿。Pas 等[4-5]先后提出了两种基于几何条件约束的检测方法,其主要思想是在物体上定义若干几何条件以支撑后续计算最佳抓取位姿。Peng 等[6]利用机械手与

收稿日期:2019-11-15 Received Date:2019-11-15
*基金项目:国家自然科学基金(61873158, 61703262)、上海市自然科学基金(18ZR1415100)项目资助

抓取对象进行几何形状层面的拟合,在体素化后的三维物体上进行抓取位姿的研究。此外,从另一角度,若目标物体本身的位姿已知,那么计算可抓取位姿的难度也就能大大降低。Wu 等[7]引入了一个完全基于合成位姿数据的位姿解释神经网络。近年来,例如区域神经网络(region convolution neural network,RCNN)等基于卷积网络的深度学习方法[8-10]也间接推动了机器人抓取问题的研究。Wu 等[11]利用混合层级特征设计了一种抓取位姿估计方法。Kumra 等[12]使用 RGB-D 图像训练卷积网络来预测图像平面中的最佳抓取位姿。Guo 等[13]设计了一个端到端网络来预测目标物体可能的抓取点,并使用参考矩形表示图像中的抓取位置。另外,级联深度学习模型在机器人抓取领域也有着进一步的应用[14-15]。另外,Morrison[16]提出了生成抓取卷积神经网络(generative grasping convolution neural network, GG-CNN),该网络可以在像素层面上得出抓取位姿,但该方法只考虑图像平面的二维抓取位姿,也无法得知所抓物体的类别。

纵观以上方法,大多数机器人抓取方法的研究是针对静态、二维、单个目标物体的可抓取位姿识别与检测问题,将其用于真实三维空间抓取任务存在很大局限性。本文提出了一种基于深度卷积网络的机器人多物体动态三维抓取位姿的检测方法。首先,采用更快区域卷积神经网络(faster region-based convolutional neural networks, Faster-RCNN)进行动态多目标检测,并设计了稳定检测滤波器以消除干扰所致的不稳定性;其次,采用 GG-CNN 网络在非结构化复杂场景中检测多个物体的二维原始抓取位姿;然后,融合目标检测结果、二维抓取位姿以及物体深度信息,重建目标物体点云,并在多个目标物体表面分别计算出对应的三维可抓取位姿;最后,通过大量的实物抓取实验,验证所提方法的可行性与有效性。

本文的主要提出了一种基于二维位姿与深度图像的三维抓取位姿生成方法。该方法获取的物体三维抓取位姿重合于真实空间中,而非传统方法中的二维图像平面中。同时,针对环境干扰等因素带来的多目标物体检测不稳定问题,提出了一种多目标稳定检测滤波器算法,提高了所提方法的适用性。

1 问题描述及方法架构

1.1 坐标系及三维抓取位姿定义

机器人抓取任务中主要的 5 个坐标系如图 1 所示。世界坐标系$\{\boldsymbol{W}\}$是一个通用架构,机器人、目标点及相机位置皆可以相对于$\{\boldsymbol{W}\}$进行定义。机器人基坐标系$\{\boldsymbol{B}\}$位于机器人底座,与$\{\boldsymbol{W}\}$的关系为${}^{W}_{B}\boldsymbol{T}$。相机坐标系$\{\boldsymbol{C}\}$通常与相机的拍摄主轴方向一致,与$\{\boldsymbol{W}\}$的关系为${}^{W}_{C}\boldsymbol{T}$。机器人夹持器$\{\boldsymbol{F}\}$位于末端,其原点通常位于机器人手指之间,$\{\boldsymbol{F}\}$相对于$\{\boldsymbol{B}\}$的定义为${}^{B}_{F}\boldsymbol{T}$。目标坐标系$\{\boldsymbol{G}\}$用于表示夹持器$\{\boldsymbol{F}\}$需要到达的位置,$\{\boldsymbol{G}\}$相对于$\{\boldsymbol{W}\}$的定义为${}^{W}_{G}\boldsymbol{T}$。如上所述,为不失一般性,文中矩阵${}^{M}_{N}\boldsymbol{T}$被定义为坐标系$\{\boldsymbol{N}\}$相对于坐标系$\{\boldsymbol{M}\}$的变换矩阵${}^{M}_{N}\boldsymbol{T}$,该变换矩阵也可被认为是“位姿”描述。

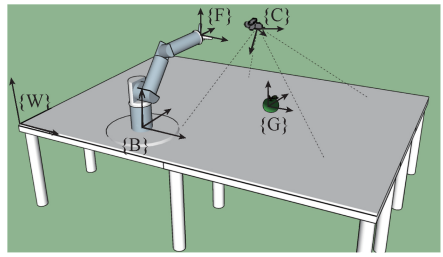


图 1 机器人抓取坐标系定义

Fig.1 Definition of the robot grasping coordinate system

目前抓取位姿多为二维抓取位姿$\boldsymbol{G}_{2D}=\{w,l,x,y,\theta\}$,其参数分别表示夹持器宽度、夹持器张开长度、抓取位姿中心点以及抓取位姿夹角。为克服二维抓取位姿在三维空间中的局限性,式(1)定义了三维物体的抓取位姿${}^{W}\boldsymbol{G}_{3D}$。

$$ {}^{W}\boldsymbol{G}_{3D}=\{x,y,z,\vec{x},\vec{y},\vec{z}\} \tag{1} $$

假设上述定义的目标坐标系$\{\boldsymbol{G}\}$与本文所提方法最终获取的三维抓取位姿一致,则:

$$ {}^{W}_{C}\boldsymbol{T}={}^{W}\boldsymbol{G}_{3D}=\begin{bmatrix} & {}^{W}_{C}\boldsymbol{R}_{\vec{x},\vec{y},\vec{z}} & & x\\ & & & y\\ & & & z\\ 0 & 0 & 0 & 1 \end{bmatrix}_{4\times4} \tag{2} $$

式中:单位向量$\{\vec{\boldsymbol{x}},\vec{\boldsymbol{y}},\vec{\boldsymbol{z}}\}$表示目标坐标系$\{\boldsymbol{G}\}$在世界坐标系$\{\boldsymbol{W}\}$的姿态,${}^{W}_{C}\boldsymbol{R}_{\vec{x},\vec{y},\vec{z}}$为旋转矩阵,$(x,y,z)$为平移向量。机器人所有抓取任务都可以采用上述定义的坐标系统及三维抓取位姿${}^{W}\boldsymbol{G}_{3D}$进行表示与描述。机器人抓取任务的实施实际上是将$\{\boldsymbol{F}\}$移动到$\{\boldsymbol{G}\}$,使$\{\boldsymbol{F}\}$与$\{\boldsymbol{G}\}$重合的过程。

1.2 整体方法架构

为实现机器人抓取物体的动作,这就意味着需要使得${}^{B}_{F}\boldsymbol{T}$与目标物体三维抓取位姿${}^{W}_{G}\boldsymbol{T}$重合。由于上面两个变换矩阵定义在不同坐标系中,需要将两个变换统一到相同坐标系中,最终目标是获取${}^{B}_{G}\boldsymbol{T}$,以更方便驱动机器人进行抓取动作的实施${}^{B}_{F}\boldsymbol{T}\rightarrow closing\rightarrow{}^{B}_{G}\boldsymbol{T}$。然而,${}^{B}_{G}\boldsymbol{T}$并不能直接获得,但是如果${}^{C}_{G}\boldsymbol{T}$能够通过本文所提出的方法计算得出,则目标矩阵${}^{B}_{G}\boldsymbol{T}={}^{B}_{C}\boldsymbol{T}{}^{C}_{G}\boldsymbol{T}$,其中${}^{B}_{C}\boldsymbol{T}$可通过相机标定得出,${}^{C}\boldsymbol{G}_{3D}={}^{C}_{G}\boldsymbol{T}$表示目标在相机坐标系中的位姿。换句话说,机器人抓取问题最终转化成求取${}^{C}\boldsymbol{G}_{3D}$,即求取${}^{C}_{G}\boldsymbol{T}$的问题。为简洁,后文中采用$\boldsymbol{G}_{3D}$表示${}^{C}\boldsymbol{G}_{3D}$。这样,三维抓取位姿${}^{W}\boldsymbol{G}_{3D}$就可以通过${}^{W}\boldsymbol{G}_{3D}={}^{W}_{G}\boldsymbol{T}={}^{W}_{B}\boldsymbol{T}{}^{B}_{C}\boldsymbol{T}$得到。

本文提出方法体系架构如图 2 所示，主要包含两部分：多目标动态检测和抓取位姿生成。多目标动态检测的作用是对多目标物体进行稳定目标识别，包括视频流控制器、多目标检测算法、多目标稳定检测滤波器。抓取位姿生成模块则是融合前者的检测结果及传感器提供的深度数据，估计二维进而计算三维抓取位姿。包括 GG-CNN 检测网络、深度目标适配器、三维点云重建、三维抓取位姿生成算法。另外，相机参数标定模块一方面给出了相机外参 ${}^{B}_{C}\boldsymbol{T}$，从而获得相对于 $\{\boldsymbol{B}\}$ 的最终三维抓取位姿，即 ${}^{B}_{G}\boldsymbol{T}$，另一方面，它提供相机内参 $\boldsymbol{P}_{\text{intrinsic}}$ 来重建 3D 点云。

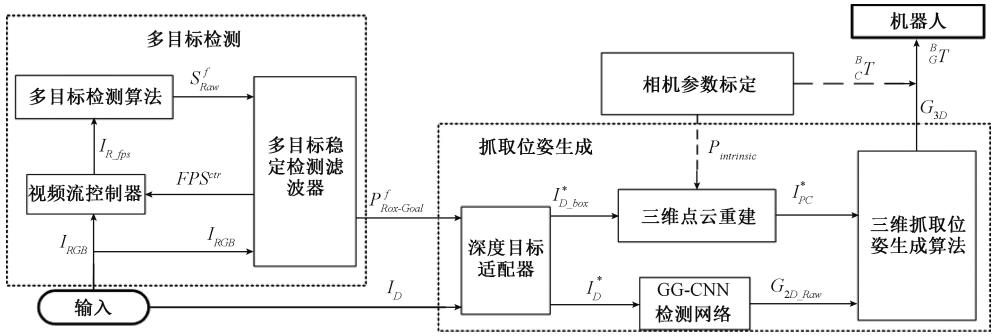


图 2　整体方法架构

Fig.2　Overall architecture of the proposed method

2　多目标动态检测及稳定检测滤波器

本文采用 Faster R-CNN，作为多目标检测算法，其高准确率保证了物体的低误识别率。为了减少资源消耗，抓取场景设计成两种检测模式：静态模式和动态模式。静态模式是指抓取场景在一定的时间间隔内几乎没有波动的情况，而动态模式是指抓取场景不稳定地连续变化的情况。帧差参考高低阈值 V_{ref_U} 及 V_{ref_L} 定义两种模式相互转换的区间。由视频流控制器根据控制参数 FPS_{ctr} 进行动态模式与静态模式的切换。当抓取场景处于动态模式时，检测算法持续运行，保证目标检测的实时性。一旦场景转换为静态模式，由于物体不再移动，检测算法挂起，减少计算资源消耗。然而，由于抓取场景中诸如不良光照、采样频率等随机噪声的干扰，检测算法的原始检测结果通常呈现不稳定状态，导致当前帧检测到的物体类别、数量与前一帧检测到的类别、数量并不完全对应，即无法保证每一帧都能检测到稳定的目标物体。因此，本文设计了多目标稳定检测滤波器，用于抑制此类不稳定性问题。

算法 1：多目标稳定检测滤波器算法

输入：

I_{RGB}：彩色图像流。

$FPS^{ctr}_{\max}$：最大控制帧率；N：滤波器窗口大小。

V_{ref_U}, V_{ref_L}：给定帧差的上限与下限。

$\boldsymbol{S}_N = [S^f_{Raw}, S^{f-1}_{Raw}, S^{f-2}_{Raw}, \cdots, S^{f-N+1}_{Raw}]$：多目标检测算法输出的最新 N 帧检测结果集合。$\boldsymbol{S}^f_{Raw} = [T^f_1, T^f_2, \cdots, T^f_i]$ 表示第 f 帧检测结果，其中：

i 表示该帧中检测到的物体个数；$T^f_i = \{C^f_{IN_i}, B^f_{IN_i}\}$ 表示第 f 帧中第 i 个检测结果，其中：$C^f_{IN_i}$ 表示检测目标 T^f_i 的类别，例如，$C^f_{IN_1}$ = 苹果。$B^f_{IN_i} = \{P^f_{B\_C_i}, l^f_{B_i}, w^f_{B_i}\}$ 表示 T^f_i 的检测框信息，其中，$P^f_{B\_C_i}$ 为检测框中心点坐标，$l^f_{B_i}$ 为检测框的长度，$w^f_{B_i}$ 为检测框的宽度。

输出：

FPS^{ctr}：每秒进行检测的帧数（控制参数）。

$P^f_{Box\_Goal} = \{O_1, O_2, \cdots, O_j\}$：稳定的检测结果。

1）初始化 P_{Box_Goal} 为空集

如果　图像帧差平均值 $V_{fd} \geqslant V_{ref\_U}$ 则

　　场景处于动态模式，FPS^{ctr} 调节至最大控制帧率 $FPS^{ctr} = FPS^{ctr}_{\max}$

否则如果　图像帧差平均值 $V_{fd} \leqslant V_{ref\_L}$ 则

　　场景处于静态模式，FPS^{ctr} 调节至 $FPS^{ctr} = 0$

否则

2）采用聚类排序操作对 S^f_{Raw} 按物体对象类别、个数进行排序，得到 S^f_{Rank}

3）依次保存每帧的排序结果 S^f_{Rank} 至大小为 N 的循环链表 S_{Rank}（N 也就是滤波器窗口大小）

4）如果 Size（S_{Rank}）$< N$ 则

5）控制帧率 $FPS^{ctr} = FPS^{ctr}_{\max}$

6）否则

$$\text{控制帧率 } FPS^{ctr} = FPS^{ctr}_{\max}\frac{V_{fd} - V_{ref\_L}}{V_{ref\_U} - V_{ref\_L}}$$

7）利用均值滤波操作从保存的大小为 N 的循环链表 S_{Rank} 中计算出最终结果 $P^f_{Box\_Goal}$

8）返回 $FPS^{ctr}, P^f_{Box\_Goal}$

算法 1 描述了多目标稳定检测滤波器算法工作原

理。该算法运行的前提条件是需要连续采集并计算至少 N 帧图像的原始检测结果。也就是说，该检测滤波器算法的窗口大小是 N，当检测识别的帧数大于等于 N 时，进行聚类排序操作和均值滤波操作，当帧数小于 N 时，只进行聚类排序操作。

聚类排序操作遵循原则：1）对每一帧原始检测结果按类别优先、同类聚集的原则进行排序，并保存最近的 N 帧排序结果 S^f_{Rank}；2）针对多目标检测算法对第 f 帧的检测结果，取出每一个 T^f_i，与 S_{Rank} 的物体依次进行类别判断与交并比（intersection-over-union，IOU）计算，如果类别相同且检测框 IOU 大于设定值 V（根据不同的目标大小选取），则 T^f_i 对应的物体与 S_{Rank} 中相应物体为同一物体，并归入同一列中，以 T^f_{ij} 表示，代表归入第 j 列中（同一列表示不同帧识别跟踪到的同一物体）。

均值滤波操作遵循原则：1）对聚类排序后的结果以列为单位进行按“同一物体”的滤波计算；2）在同一列中，均值滤波操作的窗口大小为 $N+1$，即对这 $N+1$ 个位置信息进行滤波操作。“$N+1$”由两部分组成，其中第一部分的“N”个对象是最新 N 帧的识别与排序后的结果，另一部分的“1”是前一次滤波结果（即上一次的算法输出）；3）滤波计算过程，分别计算每列 $N+1$ 个 T^f_{ij} 的有效数量 k，如果 $k < \gamma_{sf}(N+1)$，则 $C^f_{\mathrm{OUT}_j} = \mathrm{NULL}, B^f_{\mathrm{OUT}_j} = \mathrm{NULL}$，如果 $k \geqslant \gamma_{sf}(N+1)$，则 $C^f_{\mathrm{OUT}_j}$ 存在，且 $B^f_{\mathrm{OUT}_j} = mean(P^{f-1}_{Box\_Goal}, B^f_{\mathrm{IN}_{ij}}, B^{f-1}_{\mathrm{IN}_{ij}}, \cdots, B^{f-(k-2)}_{\mathrm{IN}_{ij}})$。另外，$\gamma_{sf}$ 表示滤波器算法的敏感因子，用于调节算法对动态目标物体检测与滤波的响应速度。下面以场景中包含 3 类 4 个物体（2 个苹果、1 个瓶子、1 个易拉罐）的情况进行实例说明，此例中选取滤波器窗口 $N=4$，滤波器算法的敏感因子 $\gamma_{sf}=50\%$。如图 3 所示，左边是聚类排序之前的输入。可以看到，虽然已知场景中包含 3 类 4 个物体，但受到外界环境干扰，前端多目标检测算法在每帧图像上检测到的实际结果（S^f_{Raw}）并不稳定，第 6 帧 $f=6$ 时，S^6_{Raw} 中没有检测到图像中的“易拉罐”，第 7 帧 $f=7$ 时，S^7_{Raw} 检测到所有物体，第 8 帧 $f=8$ 时，S^8_{Raw} 没有检测到“易拉罐”和另外一个“苹果”，第 9 帧 $f=9$ 时，S^9_{Raw} 没有检测到“瓶子”和“易拉罐”。经聚类排序操作后的结果如图 3 右半部分所示，其中采用不同颜色字体表示不同物体类别，不同列表示不同的物体（同一列表示同一物体在不同帧中的信息）。另外，图中标注的 $P^8_{Box\_Goal}$ 表示在 $f=8$ 保存的滤波器输出结果，$P^9_{Box\_Goal}$ 表示当前 $f=9$ 时计算所得的结果。需要指出的是，针对最右列的易拉罐，当前计算结果为“空”，是因为此时此列的有效数量 $k=2<\gamma_{sf}\cdot(N+1)=50\%\cdot(4+1)$。其他 3 列都是有效值，且 $B^9_{\mathrm{OUT}_j} = mean(B^8_{\mathrm{OUT}_j}, B^9_{\mathrm{IN}_{ij}}, B^8_{\mathrm{IN}_{ij}}, B^7_{\mathrm{IN}_{ij}}, B^6_{\mathrm{IN}_{ij}})$。

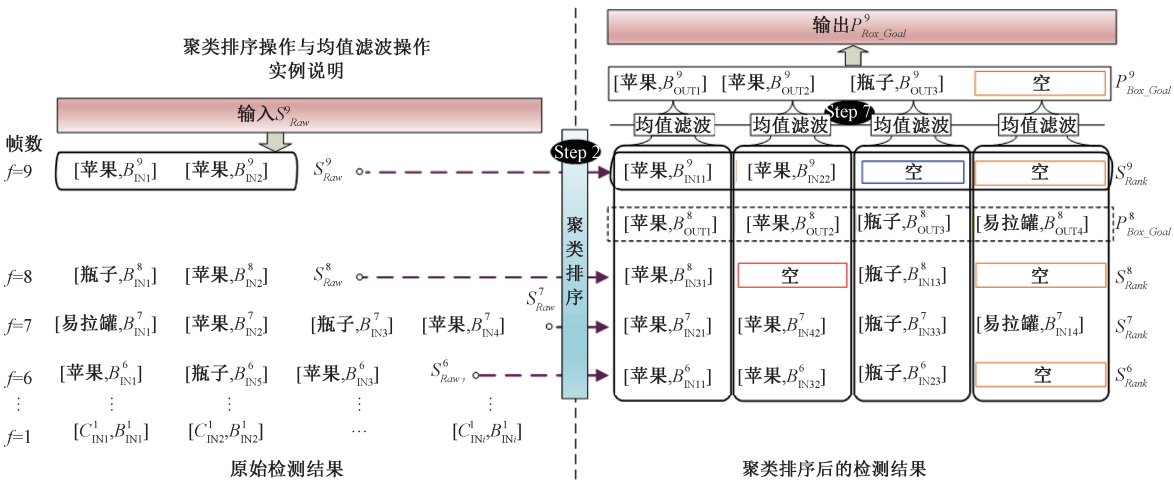


图 3 算法 1 中步骤 2 与 7 的实例演示

Fig.3 Example demonstration of the step 2 and 7 in algorithm 1

3 三维抓取位姿生成方法

机器人三维抓取位姿相较于二维抓取具有明显优势。本文在获取的二维抓取位姿基础上，融合视觉相机获取的深度空间信息，提出了三维抓取位姿生成算法。

3.1 二维抓取位姿获取及深度目标适配器

本文采用 GG-CNN 轻量神经网络模型，采用大量深度图数据集进行训练，进而应用所得模型对目标物体二维抓取位姿进行估计，其网络输入为 300×300×1 的单通道深度图，输出为 300×300×1 的三通道特征图。每个像素点代表一个抓取位姿，共 90 000 个抓取位姿。

由于 RGB-D 相机获取的深度图像，极易受到外界环境干扰，故有必要实施预处理操作，如高斯平滑滤波等。另外，由于 GG-CNN 网络的输入大小为 300×300 深度图，多数情况下与所选用相机获取的深度图像大小不一致，

因此提出了一种深度目标适配器：以物体检测框中心点为中心，裁剪出 300×300 的深度图像矩形区域，若裁剪区域超出深度相机的图像采集范围，如图 4 所示，则以边缘像素值进行对称映射填充，其中内框为物体检测框，外框为裁剪区域，分别为图 2 中标注的 $I^*_{D\_box}$ 与 I^*_D。

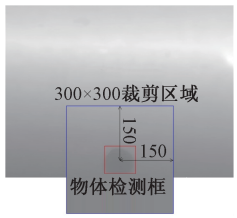


图 4　深度目标适配器效果

Fig.4　The effect of depth target adapter

3.2　点云重建及三维抓取位姿生成算法

为生成三维抓取位姿，点云重建必不可少。在计算机视觉中，通常将相机模型简化成小孔成像模型，定义图像坐标系 $\{\boldsymbol{I}\}$ 和像素坐标系 $\{\boldsymbol{P}\}$。图像坐标系是为了描述成像过程中物体从相机坐标系 $\{\boldsymbol{C}\}$ 到图像坐标系 $\{\boldsymbol{I}\}$ 的透射投影关系而引入，像素坐标系是为了描述物体成像后的像点在数字图像上的坐标而引入。$\{\boldsymbol{I}\}$ 用物理单位（如 m）表示像素在图像中的位置，而 $\{\boldsymbol{P}\}$ 的表示单位为个（像素数目）。因此，根据小孔成像模型的内参标定方法，可以推导出深度图像的像素坐标到相机坐标三维点云数据的计算公式，如式（3）所示。

$$\boldsymbol{I}^*_{PC}=d_{x,y}\begin{bmatrix}\dfrac{f}{s_u} & 0 & c_u\\ 0 & \dfrac{f}{s_v} & c_v\\ 0 & 0 & 1\end{bmatrix}^{-1}\begin{bmatrix}x\\ y\\ 1\end{bmatrix} \tag{3}$$

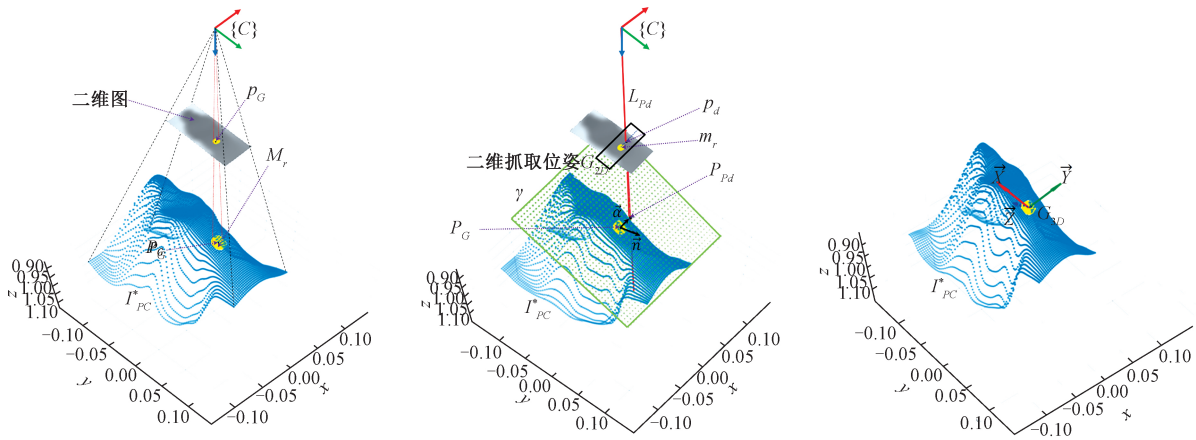
式中：$I^*_{PC}=(X_C,Y_C,Z_C)^T\subseteq\mathscr{R}^3$ 为相机坐标系 $\{\boldsymbol{C}\}$ 下的点云数据坐标。物体检测框 $I^*_{D\_box}$ 中 $d_{x,y}$ 为像素点 (x,y) 对应空间点到相机镜头平面的距离。$P_{intrinsic}=(f,s_u,s_v,c_u,c_v)$ 表示相机内参，其中 f 为相机主距，s_u 和 s_v 为每个像素在图像平面 x 和 y 方向上的物理尺寸，(c_u,c_v) 为图像坐标系原点在像素坐标系中的坐标。

三维抓取位姿 G_{3D} 的计算主要由两个部分组成：三维抓取点 $P_G=(x_G,y_G,z_G)$；抓取点姿态单位向量 $\vec{x}$，$\vec{y}$，$\vec{z}$。P_G 确定三维抓取位姿空间位置，$(\vec{x},\vec{y},\vec{z})$ 确定机器人终端夹持器最终抓取姿态，$\vec{z}$ 确定夹持器抓取时向物体靠近的方向，而 $\vec{x}$ 和 $\vec{y}$ 决定夹持器旋转角度。

如图 5 所示。以瓶子为例详细展示三维抓取位姿生成算法。令 $p_G=(x,y)$ 为二维图中二维抓取点，采用式（3）可求出三维抓取点 $P_G=(x_G,y_G,z_G)$。假设二维图中以 p_G 为圆心、半径为 r（例如 $r=5$）个像素圆为 m_r，在相机坐标系 $\{\boldsymbol{C}\}$ 中，两维区域 m_r 在三维空间所投影出的点云为 M_r，其为整个物体点云 $\boldsymbol{I}^*_{PC}$ 中点 P_G 的邻域，即 $M_r\subseteq I^*_{PC}$。采用最小二乘法，在 M_r 上拟合出最优切平面 $\gamma\subseteq\mathscr{R}^3$，进而计算三维抓取点 P_G 处于平面 γ 垂直的法线 $\vec{n}$。

另外，在二维图中，以 p_G 为起点，在二维抓取位姿 G_{2D} 的长边方向任取一点 p_d。由相机坐标系 $\{\boldsymbol{C}\}$ 的原点（光心）指向 p_d 的射线定义为 L_{p_d}，其与 γ 的交点定义为 P_{p_d}，计算过程可表示为 $P_{p_d}=\gamma\cap L_{p_d}$。进而，由 P_G 指向 P_{p_d} 的向量定义为 $\vec{\alpha}$，并利用式（4）即可计算三维抓取位姿的单位向量 $\vec{x}$，$\vec{y}$，$\vec{z}$。

$$\vec{\boldsymbol{x}}=\frac{\vec{\boldsymbol{\alpha}}\times(-\vec{\boldsymbol{n}})}{\|\vec{\boldsymbol{\alpha}}\times(-\vec{\boldsymbol{n}})\|_2},\vec{\boldsymbol{y}}=\frac{\vec{\boldsymbol{\alpha}}}{\|\vec{\boldsymbol{\alpha}}\|_2},\vec{\boldsymbol{z}}=\frac{-\vec{\boldsymbol{n}}}{\|-\vec{\boldsymbol{n}}\|_2} \tag{4}$$



(a) 瓶子的点云 I^*_{PC}
(a) The point cloud I^*_{PC} of the bottle

(b) 向量 $\vec{\alpha}$ 以及表面法线 $\vec{n}$
(b) Vector $\vec{\alpha}$ and surface normal $\vec{n}$

(c) 最终三维抓取位姿 G_{3D}
(c) Final 3D grasp pose G_{3D}

图 5　三维抓取位姿生成算法（以瓶子为例）

Fig.5　3D grasp pose generation algorithm (taking a bottle as exmple)

综上所述,结合三维抓取点 $P_G=(x_G,y_G,z_G)$ 以及单位向量 $\overrightarrow{x}$, $\overrightarrow{y}$, $\overrightarrow{z}$, 便可得到三维抓取位姿 $G_{3D}=\{x_G,y_G,z_G,\overrightarrow{x},\overrightarrow{y},\overrightarrow{z}\}$, 如图 5(c)所示。

4 机器人抓取平台验证及分析

4.1 平台搭建

实物验证所用机器人为 Kinova Jaco2 机械臂,相机选用 Realsense D435 RGB-D 深度相机,抓取物体选取生活中常见的 5 类物体:瓶子,香蕉,苹果,橙子,易拉罐。对于多目标检测算法,本文对 Open Image V4 数据集以及自制数据集进行旋转,模糊等数据增强处理,得到共 10 500 张训练样本进行训练。GG-CNN 网络训练所用数据集为康奈尔抓取数据集,其中包含 1 035 张标有正例与反例的二维抓取位姿的深度图像。实验所用计算平台的显卡为 NVIDIA GeForce GTX 1080,CPU 为 3.4 GHz Intel Core i7-6800k。图 6 所示为实物验证平台。

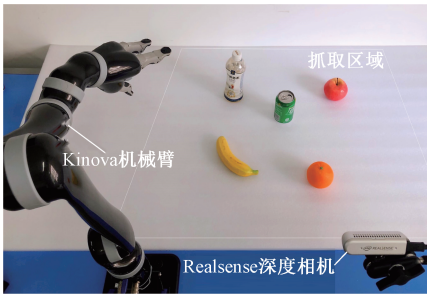


图 6 抓取实验平台

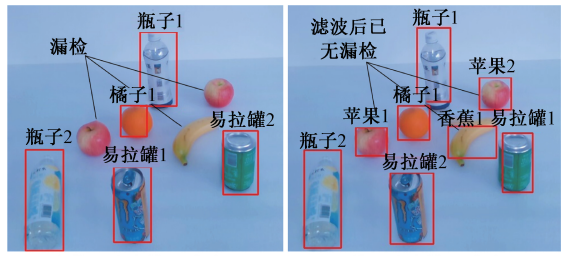
Fig.6 Grasping experiment platform

4.2 多目标稳定检测滤波器验证

为验证多目标稳定检测滤波器的效果,实验中对其进行了多次测验,图 7 所示展示了在两个不同的场景中,未使用该滤波器和使用后的两组效果图,对比图 7(a)与图 7(b)可以看出,多目标稳定检测滤波器对于检测结果不稳定的问题有明显的抑制作用。

4.3 物体抓取位姿生成与抓取验证

采用 GG-CNN 预测二维抓取位姿,其效果如图 8 所示,实验显示 GG-CNN 对 5 类物体的二维抓取位姿估计与显示效果。图 9 所示展示了采用本文提出的三维抓取位姿生成方法,所得到的两个场景中 5 类多个物体的三维抓取位姿估计与显示效果。



(a) 未使用稳定检测滤波器
(a) Results without the stabilization detection filter

(b) 使用稳定检测滤波器
(b) Results with the stabilization detection filter

图 7 多目标稳定检测滤波器的效果

Fig.7 The effect of multi-object stabilization detection filter

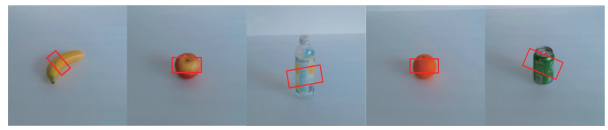


图 8 GG-CNN 二维抓取位姿效果

Fig.8 2D grasps pose effect diagram generated with GG-CNN

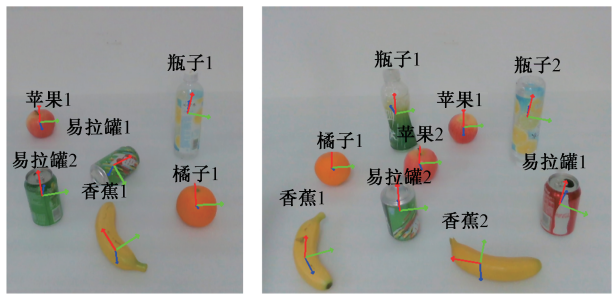


图 9 三维多物体抓取位姿

Fig.9 3D multi-object grasp pose

图 10 所示演示了机器人对“瓶子”抓取过程,整个过程包括起始、预抓取、抓取和后抓取等连贯过程。根据所提方法计算出三维物体当前可抓取位姿,驱动机器人,使得末端夹持器坐标系与计算所得三维抓取位姿重合,实现抓取任务。

(a) 起始 (a) Start

(b) 预抓取 (b) Pre-grasping

(c) 抓取 (c) Grasping

(d) 后抓取 (d) Post-grasping

(e) 起始 (e) Start

图 10 机器人抓取过程

Fig.10 Robot grasping process

4.4　机器人抓取结果分析

为验证三维抓取位姿生成算法的可行性与有效性，在图 6 所示的抓取平台上对选取的 5 类物体进行了 1 000 次抓取实验，每种类别进行 200 次抓取验证。抓取实验的统计数据如表 1 所示。

表 1　5 类物体抓取实验统计数据

Table 1　Grasping experiment statistic data for 5 categories of objects

类别	抓取次数	成功次数	成功率/%	识别用时/(s/次)	抓取用时/(s/次)
苹果	200	196	98	0.29	9.65
橘子	200	199	99.5	0.27	9.97
易拉罐	200	193	96.5	0.32	9.63
香蕉	200	181	90.5	0.3	10.02
瓶子	200	187	93.5	0.29	9.45
平均指标			95.6	0.29	9.74

可以看出，除香蕉和瓶子外，其他类别的物体都具有很高的抓取成功率。主要原因是：1）夹持器指尖设计不利于小物体的抓取，香蕉相较于机器人手指末端较小，抓取难度较大；2）瓶子的材质是透明的，采用 Realsense 深度相机无法在透明部分获得精确的深度信息，进而导致计算所得的三维抓取位姿存在一定偏差。虽然上述原因在一定程度上影响了抓取效果，但从整体数据可以看出，机器人能以 95.6% 的平均成功率稳定抓取到目标物体，也充分验证了本文提出的多物体动态三维抓取位姿检测方法的可行性及有效性。另外，从表 1 中可看出，本方法对于单个物体识别用时达 0.29 s，可满足平时大多数变化不太激烈的抓取场景。机器人抓取用时受机器人运动速度限制，但也可控制在单个物体在 10 s 内成功抓取。

5　结　　论

本文提出了一种动态多目标三维抓取位姿估计方法，克服了二维抓取位姿在实际抓取过程中的局限性。结合基于 Faster R-CNN 的目标检测与抓取位姿估计方法，实现了物体类别识别与三维抓取位姿计算两大任务。针对目标检测中光线等干扰导致的不稳定性，设计了多目标稳定检测滤波器，为真实机器人稳定抓取提供了有力支撑。在 1 000 次的实际抓取实验中，本文提出的三维抓取位姿检测方法平均抓取成功率达 95.6%，验证了此方法的可行性与有效性。

参考文献

[1] YANG A, NAEEM W, IRWIN G W, et al. Stability analysis and implementation of a decentralized formation control strategy for unmanned vehicles [J]. IEEE Transactions on Control Systems Technology, 2014, 22(2):706-720.

[2] YANG A, NAEEM W, FEI M, et al. A co-operative formation-based collision avoidance approach for a group of autonomous vehicles [J]. International Journal of Adaptive Control & Signal Processing, 2016, 31(4): 489-506.

[3] 惠文珊，李会军，陈萌，等. 基于 CNN-LSTM 的机器人触觉识别与自适应抓取控制[J].仪器仪表学报，2019, 40(1):211-218.
HUI W SH, LI H J, CHEN M, et al. Robotic tactile recognition and adaptive grasping control based on CNN-LSTM [J]. Chinese Journal of Scientific Instrument, 2019, 40(1):211-218.

[4] PAS A T, GUALTIERI M, SAENKO K, et al. Grasp pose detection in point clouds [J]. The International Journal of Robotics Research, 2017, 36 (13-14): 1455-1473.

[5] PAS A T, PLATT R. Using geometry to detect grasps poses in 3D point clouds[C]. Proceedings of Advanced Robotics, 2018:307-324.

[6] PENG S, FU ZH Q, LIU L G. Grasp planning via hand-object geometric fitting[J]. The Visual Computer, 2016, 34:257-270.

[7] WU J, ZHOU B, RUSSELL R, et al. Real-time object pose estimation with pose interpreter networks [C]. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018:6798-6805.

[8] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]. IEEE International Conference on Computer Vision (CVPR), 2014: 580-587.

[9] GIRSHICK R. Fast R-CNN [C]. IEEE International Conference on Computer Vision (CVPR), 2015: 1440-1448.

[10] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 39(6):1137-1149.

[11] WU G, CHEN W, CHENG H, ZUO W, ZHANG D. Multi-object grasping detection with hierarchical feature fusion[J]. IEEE Access, 2019, 7:43884-43894.

[12] KUMRA S, KANAN C. Robotic grasp detection using deep convolutional neural networks [C]. IEEE/RSJ International Conference on Intelligent Robots & Systems,

2017:769-776.

[13] GUO D, SUN F, KONG T AND LIU H. Deep vision networks for real-time robotic grasp detection [J]. International Journal of Advanced Robotic Systems, 2017, 14(1):1-8.

[14] 喻群超, 尚伟伟, 张驰. 基于三级卷积神经网络的物体抓取检测[J]. 机器人, 2018, 40(5):762-768.
YU Q CH, SHANG W W, ZHANG CH. Object grasp detecting based on three-level convolutional Neural Network[J]. Robot, 2018, 40(5):762-768.

[15] 陈丹, 林清泉. 基于级联式 Faster RCNN 的三维目标最优抓取方法研究[J]. 仪器仪表学报, 2019, 40(4):229-237.
CHEN D, LIN Q Q. Research on 3D object optimal grasping method based on cascaded faster RCNN [J]. Chinese Journal of Scientific Instrument, 2019, 40(4):229-237.

[16] MORRISON D, CORKE P, LEITNER J. Closing the loop for robotic grasping: a real-time, generative grasp synthesis approach[C]. Robotics: Science and Systems (RSS), 2018:1-10.

作者简介

杨傲雷,2004 年于湖北工业大学获得学士学位,2009 年于上海大学获得硕士学位,2012 年于英国贝尔法斯特女王大学获得博士学位,现为上海大学副教授,主要研究方向为多智能体协同控制、智能机器人与视觉学习系统、机器学习与数据科学等。

E-mail: aolei@ shu.edu.cn

Yang Aolei received his B. Sc. degree from Hubei University of Technology in 2004, M. Sc. degree from Shanghai University in 2009, and Ph. D. degree from Queen's University Belfast, UK in 2012. He is currently an associate professor at Shanghai University. His main research interests include multi-intelligence cooperative control, intelligent robot and visual learning system, UAV group cooperative formation and control.

徐昱琳(通信作者),1986 年于东华大学获取学士学位,2003 年于法国斯特拉斯堡大学获得博士学位,现为上海大学副教授,主要研究方向多变量工业系统的建模与控制、仿人灵巧手结构设计、建模与控制、服务机器人智能控制等。

E-mail: xuyulin@ shu.edu.cn

Xu Yulin (Corresponding author) received her B. Sc. degree from Donghua University in 1986, and Ph. D. degree from University of Strasbourg, France in 2003. She is currently an associate professor at Shanghai University. Her main research interests include modeling and control of multi-variable industrial systems, bionic manipulator structure design, modeling and control and service robot intelligent control.

陈灵,分别于 2008 年和 2010 年于中南大学获得学士学位和硕士学位,2014 年于埃塞克斯大学获得博士学位,现为湖南师范大学助理研究员,主要研究方向为计算机视觉,机器人定位与导航。

E-mail: lcheno@ hunnu.edu.cn

Chen Ling received his B. Sc. degree and M. Sc. degree both from Central South University in 2008 and 2010, respectively, and received his Ph. D. degree from University of Essex, UK in 2014. He is currently an associate research fellow at Hunan Normal University. His main research interests include computer vison and localization and navigation of robots.