一种机械臂像素级抓取检测方法的研究

何联格[1,2,3], 聂远航[1], 李天华[1], 妥吉英[1,3]
1. 重庆理工大学 汽车零部件先进制造技术教育部重点实验室,重庆 400054
2. 北京信息科技大学 现代测控技术教育部重点实验室,北京 100192
3. 重庆青山工业有限责任公司,重庆 402761

摘 要:目的 针对在平面抓取的场景下,机械臂如何感知目标物体的位置信息并完成抓取作业,提出了一种兼顾检测速度和精度的抓取检测网络。**方法** 根据抓取检测任务中输入与输出尺寸大小相同的特点,采用语义分割思想设计了抓取检测网络;输入为随机裁剪后的深度图片,输出为同尺寸的抓取置信度、抓取角度和抓取宽度特征图;为了提高平面抓取任务的效率,在综合考虑检测网络速度和精度的情况下,对网络结构进行了改进,除了在网络结构中加入了注意力机制外,还使用 U-net 和 Deeplabv3 算法替换了网络的主体结构;通过对比实验认为加入注意力机制的检测网络在检测速度和检测精度上平衡得较好,能够实现抓取任务,将抓取位姿传输给 ROS 系统,通过一系列的坐标变换和运动规划进行了抓取作业。**结果** 添加注意力机制后,检测网络的推理时间仍为毫秒级,最大抓取置信度提高了 7.2%;采用 U-net 和 Deeplabv3 的网络检测速度较慢,U-net 网络的抓取置信度 $\widetilde{Q}_{\max}$ 提高了 18.8%,Deeplabv3 网络的准确率 $\boldsymbol{A}_{cc}$ 提高了 12.8%;由于抓取检测网络应同时考虑检测速度和精度,因此加入注意力机制的检测网络性能较优。**结论** 实验结果表明:添加注意力机制后的抓取网络在检测速度与精度上兼顾的比较理想,能够成功抓取场景中的目标物体,此方法具有一定的实际应用价值。

关键词:平面抓取;抓取检测;语义分割;机械臂抓取任务
中图分类号:TP249 **文献标识码:**A doi:10.16055/j.issn.1672-058X.2024.0002.003

Research on Pixel-level Grasping Detection Method of Robot Arm

HE Liange[1,2,3], NIE Yuanhang[1], LI Tianhua[1], TUO Jiying[1,3]
1. Key Laboratory of Advanced Manufacture Technology for Automobile Parts, Ministry of Education, Chongqing University of Technology, Chongqing 400054, China
2. Key Laboratory of Modern Measurement & Control Technology, Ministry of Education, Beijing Information Science and Technology University, Beijing 100192, China
3. Chongqing Tsingshan Industrial Co., Ltd., Chongqing 402761, China

Abstract: Objective A grasping detection network that balances detection speed and detection accuracy was proposed for how a robotic arm can sense the position information of a target object and complete a grasping operation in a planar grasping scenario. **Methods** Based on the feature that the input size and the output size are the same in the grasp detection task, a grasp detection network was designed using semantic segmentation ideas. The input was a randomly cropped depth image, and the output was images of the same size containing features like grasp confidence, grasp angle

收稿日期:2023-01-04 **修回日期:**2023-03-01 **文章编号:**1672-058X(2024)02-0018-08
基金项目:重庆市基础科学与前沿技术研究专项课题(CSTC2020JCYJ-MSXMX0331);重庆市教育委员会科学技术研究计划青年项目(KJQN202101144).
作者简介:何联格(1984—),男,陕西省周至县人,博士,讲师,从事整车热分析和深度学习研究.
通讯作者:妥吉英(1988—),男,甘肃省东乡县人,博士,讲师,从事计算机视觉和振动控制研究. Email:tjy@cqut.edu.cn.
引用格式:何联格, 聂远航, 李天华,等.一种机械臂像素级抓取检测方法的研究[J].重庆工商大学学报(自然科学版),2024,41(2):18—25.
HE Liange, NIE Yuanhang, LI Tianhua, et al. Research on pixel-level grasping detection method of robot arm[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2024, 41(2): 18—25.

and grasp width. To improve the efficiency of the planar grasping task, the network structure was improved by replacing the backbone structure of the network using the U-net and Deeplabv3 algorithms, in addition to adding an attention mechanism to the network structure, taking into account the speed and accuracy of the detection network. Comparative experiments concluded that the detection network with the attention mechanism has a better balance of detection speed and detection accuracy, and can achieve the grasping task, transmitting the grasping poses to the ROS system and performing the grasping operation through a series of coordinate transformations and motion planning. **Results** After adding the attention mechanism, the inference time of the detection network was still in the millisecond range, and the maximum grasping confidence was improved by 7.2%. The detection speed of the networks using U-net and Deeplabv3 was slower, the grasping confidence $\widetilde{Q}_{\max}$ of the U-net network was improved by 18.8%, and the accuracy $\boldsymbol{A}_{cc}$ of the Deeplabv3 network was improved by 12.8%. Since the grasp detection network should consider both detection speed and accuracy, the detection network with the addition of the attention mechanism has better performance. **Conclusion** The experimental results show that the grasping network with the attention mechanism has a better balance of detection speed and accuracy and can successfully grasp the target objects in the scene, and the method has some practical application value.

Keywords: planar grasping; grasping detection; semantic segmentation; robotic arm grasping tasks

1　引　言

机械臂的自主抓取是机器人获取、移动、运输目标物体等动作的前提,具有广泛的应用场景。近年来深度学习以其突出的检测精度和可靠性,成为机械臂抓取控制领域的研究热点[1-3]。机械臂抓取算法指的是在包含待抓取物体和机械手模型的场景中,以视觉信息为输入,通过算法计算出一个优良的抓取位姿,使机械臂在该位姿下能够稳定方便的抓取场景中的物体。

根据机械手模型的抓取位姿进行分类,抓取算法可以分为平面抓取和 6 自由度抓取两类。其中平面抓取的方法限制机械臂只能垂直于桌面进行抓取,方法需要预测的参数较少,实现难度较低,适用于工业抓取。目标检测技术快速发展的同时也带动了抓取检测技术的发展,不论是基于锚框的单阶段、双阶段检测方法,还是无锚框的检测方法,通过在矩形抓取框的基础上添加抓取角度 φ 便可以实现抓取检测[4-5]。经过多年的发展,目标检测技术在机械臂视觉抓取检测任务中已经取得了长足的进步,但仍然存在着问题和挑战:作为离散预测方式的矩形框难以描述球形物体抓取情况;多数目标检测网络含有大量的参数,网络结构比较复杂,很难平衡检测精度与速度。

为了改善抓取检测的适用性和平衡检测过程中的精度和速度。Mahler 等[6]采集了场景中的深度信息并生成候选抓取位置,每个抓取位置对应一个抓取质量,选择质量最高的抓取位置进行抓取,方法需要生成大量的候选位置且需要点云信息,检测速度较慢。Morrison 等[7]将场景中深度信息图片作为 GG-CNN 网络的输入,输出抓取置信度,抓取宽度以及角度,依旧使用了矩形框描述的抓取表示构建了数据集。Wang Dexin 等[8]提出了一种像素级数据集的标定方法,使用抓取点、抓取宽度以及角度等参数来描述单侧,双侧以及环形抓取标签,文中设计的网络以彩色图片作为输入,进行抓取任务时需另外读取深度信息,使抓取任务效率下降。

为了提高平面抓取任务的效率,综合考虑抓取检测模型的速度和精度,在文献[7]、文献[8]的基础上,采用语义分割思想,提出了一种以深度图片作为输入、平面抓取位姿作为输出的抓取检测网络并对其结构进行了改进。具体来说,除了为原有的网络添加注意力机制外,还使用了在语义分割任务中表现良好的 U-net 和 Deeplabv3 网络,并对这些方案进行了比较。结合像素级标定方法在真实场景和 Cornell 数据集上进行标注,从而实现机械臂的快速抓取检测。在 Cornell 和实际采集数据集上的测试表明,抓取检测网络在添加了注意力模块后,检测速度和检测精度的平衡较好,最后使用安诺机械臂视觉平台实现了真实环境中目标物体的抓取操作。

2　二指夹爪的抓取表示方法

随着技术的不断发展,更多机器人系统都配有各式各样的传感器来获得彩色图片、点云等环境信息。机械臂首先要对目标物体的方位、方向和抓取位置进行感知,从而进行路径规划和抓取检测计算,实现目标抓取。空间六自由度抓取方法可以实现多角度灵活抓取任务,方法需预测出描述对象位置、旋转和抓取宽度等 7 个参数,通过实例分割场景点云信息以保证点云精确并保障信噪比,且计算量大。平面抓取方式适合应用在实际的工业生产中,方法限制机械臂只进行垂直桌面的抓取任务,只需要预测[x,y,z,θ,w]5 个参数,

无须预测绕 X 轴、Y 轴旋转参数。对于机械臂平面抓取方式,抓取表示方法又可分为方向矩形框和像素点 2 种。

2.1 基于方向矩形框的抓取表示

目标检测任务常采用矩形框来表示物体的类别以及位置信息,矩形抓取框在此基础上删除了物体类别 class 并添加了旋转角度 φ 以表示抓取检测。使用 $\tilde{\boldsymbol{g}}$ 表示像素坐标系下的抓取模型,如式(1)所示,其示意图如图 1 所示。

$$\tilde{\boldsymbol{g}}=[u,v,\varphi,w,m] \tag{1}$$

式(1)中:(u,v) 表示像素坐标系下的坐标位置信息,φ 表示夹爪与图片水平轴的夹角,w 表示夹爪的张开宽度,m 表示机械臂夹爪的物理尺寸在图片坐标系下的映射。

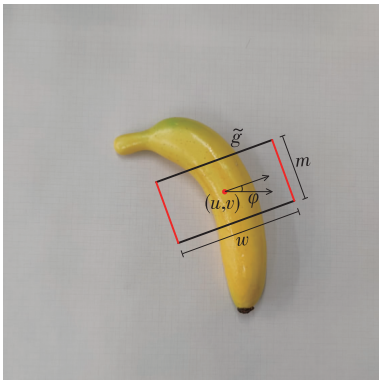

图 1 方向矩形框抓取表示

Fig. 1 Oriented rectangle-based grasping representation

世界坐标系下的抓取位姿 $\boldsymbol{g}$ 可由 $\tilde{\boldsymbol{g}}$ 经过坐标变换得到。变换关系如式(2)所示:

$$\boldsymbol{g}=\boldsymbol{t}_{WC}(\boldsymbol{t}_{CI}(\tilde{\boldsymbol{g}})) \tag{2}$$

式(2)中:$\boldsymbol{t}_{WC}$ 为世界坐标系(机械臂坐标系)与相机坐标系之间的变换矩阵,$\boldsymbol{t}_{CI}$ 为二维图片坐标系与三维相机坐标系之间的映射矩阵。

Chu Fujen 等[9]受 Faster RCNN 的启发,在 ResNet-50 网络的基础上实现了端到端的抓取预测。将抓取预测拆分为两部分来解决场景中多目标多抓取问题,一是预测目标的位置信息(x,y,h,w)的回归预测,二是对抓取方向(φ)进行分类预测。

在矩形框表示的抓取检测模型中,大量的先验框均为负样本,易导致样本不均衡的问题,还会消耗大量的计算时间。此外矩形框这一抓取表示方法有一定的局限:CNN 网络预测的夹爪尺寸 m 为一个变量,与实际情况不符;矩形框这种离散的表示方式无法全面地描述场景中球形物体的所有抓取情况;参数 m 不需要专门进行预测,在数据集标注时可通过碰撞检测来隐性表达。

2.2 基于像素点的抓取表示

像素点抓取方法是指对图片信息中每个像素点都生成抓取位姿,由抓取点、抓取角度和抓取宽度三部分组成,表示方法由式(3)所示,图 2 为其示意图。

$$\tilde{\boldsymbol{g}}=[u,v,\tilde{\varphi},\tilde{W}] \tag{3}$$

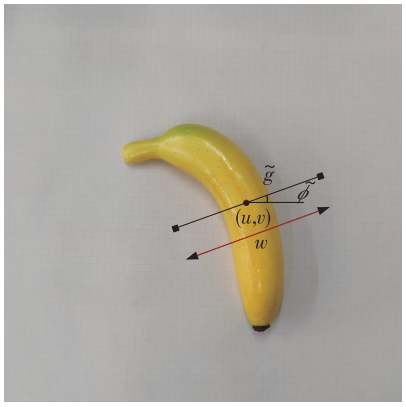

图 2 像素点抓取表示

Fig. 2 Pixel-based grasping representation

通过式(2)的坐标转换,世界坐标系下的像素点抓取表示如式(4)所示:

$$\boldsymbol{g}=[x,y,z,\varphi,w] \tag{4}$$

式(4)中:(x,y,z) 表示目标物体的位置信息;φ 表示夹爪绕 Z 轴的旋转角度;w 表示夹爪的张开宽度。

基于像素点抓取方法,文献[7]提出一种全卷积网络实现深度图到抓取位姿的映射,在检测精度和推理速度达到较好的平衡。平面抓取任务采用像素级抓取表示更为合理[10],基于图像分割任务在 RGB 图像上预测出三指夹爪合适的抓取区域以及对应的抓取角度和宽度。文献[11]使用了残差-挤压-激励网络(RSEN)提取场景中深度信息,实现了像素级学习和抓取位姿估计。

3 基于语义分割任务的抓取检测网络

3.1 抓取表示

平面抓取检测需要在给定场景信息(如彩色图像、深度信息和点云信息)情况下,获得目标物体的抓取位姿。在式(3)的基础上添加表示抓取点抓取成功率的参数 q,使用已知内参的相机采集的深度图片 $I=R^{H\times W}$ 中,抓取位姿如式(5)所示:

$$\tilde{\boldsymbol{g}}=(p,q,\tilde{\varphi},\tilde{W}) \tag{5}$$

式(5)中:p 表示抓取点的坐标(u,v),单位为像素值;q 表示抓取点(u,v)的抓取成功率,取值为$[0,1]$。

机械臂视觉系统标定完成后,世界坐标系下抓取表示 $\boldsymbol{g}$ 如式(6)所示:

$$\boldsymbol{g}=(P,q,\varphi,w) \tag{6}$$

式(6)中:P 表示世界坐标系的抓取点的位置(x,y,z)。

实际场景中物体的抓取位姿并不是唯一的,采用参数 $\tilde{G}$ 表示像素坐标系下的抓取集合,如式(7)所示:

$$\tilde{G}=(\tilde{Q},\tilde{\Phi},\tilde{W}) \tag{7}$$

式(7)中:$\tilde{Q},\tilde{\Phi},\tilde{W}$ 分别使用 $q,\tilde{\varphi},\tilde{W}$ 来填充,其尺寸与输入的深度图片一致,即$(\tilde{Q},\tilde{\Phi},\tilde{W})=R^{3\times H\times W}$。

抓取置信度 $\tilde{Q}$ 中的最大值 $q_{\max}$ 为像素坐标系下的最优抓取点 $p^*=(u^*,v^*)$,在抓取角度 $\tilde{\Phi}$ 和抓取宽度 $\tilde{W}$ 上相应位置 p^* 可以得到相应的抓取角度 $\tilde{\varphi}^*$ 和抓取宽度 $\tilde{W}^*$,那么像素坐标系下的最优抓取位姿为 $\tilde{G}^*=\max_{\tilde{Q}}\tilde{G}$。

3.2　抓取检测网络的结构

根据 2.1 节中的描述,抓取检测模型 M 应实现从深度图片 $I\in R^{H\times W}$ 到抓取集合 $\tilde{G}\in R^{3\times H\times W}$ 的映射,即 M:$I\rightarrow(\tilde{Q},\tilde{\Phi},\tilde{W})$。模型 M 输入为深度图片 I,输出为抓取集合 $\tilde{G}$,旨在使深度图片中每个像素点与抓取位姿一一对应;语义分割任务的目标是对输入图片中每个像素赋予标签,也就是像素级分类任务,二者目标一致。因此模型 M 首先通过特征提取网络对输入图片进行下采样操作,然后通过特征融合网络对特征图执行上采样和融合操作来完成抓取检测任务。

全卷积网络已经被证明在抓取检测任务中表现良好[12],Morrison 等[7]设计了一种轻量级的抓取检测网络 GG-CNN2。现在 GG-CNN2 的基础上开展了改进研究,抓取检测网络 M 的整体结构如图 3 所示。Wang 等[13]提出的空洞卷积层能在参数不变的情况下增加模型的感受野。GG-CNN2 含有两层空洞卷积且膨胀率分别为 2 和 4,而空洞卷积组内的膨胀率 r 的公约数不应大于 1,为了使网络能够捕捉到图像中更多的全局信息,增加一层空洞卷积并将膨胀率修改为 $r=(1,2,3)$[13]。模型 M 为全卷积神经网络,采用 ReLU 激活函数,其主要架构如下:

(1) 左边网络为特征提取网络。输入为单通道 360×360 的深度图片,利用卷积层和 padding 操作提取特征信息但不改变特征图的尺寸;2 次最大池化层使特征图的尺寸减小为 90×90×16;3 次空洞卷积后特征图尺寸变为 90×90×32。

(2) 中间网络为特征融合网络。采用双线性上采样,将特征图尺寸还原为 360×360×16。

(3) 右侧网络为预测网络。采用卷积核大小为 1×1 的卷积得到 4 张特征图:

抓取置信度 $\tilde{Q}$,抓取宽度 $\tilde{W}$,抓取角度 $\sin(2\tilde{\varphi})$ 和 $\cos(2\tilde{\varphi})$,经过 $\tilde{\Phi}=\arctan\left(\frac{\sin(2\tilde{\Phi})}{\cos(2\tilde{\Phi})}\right)/2$ 得到抓取角度 $\tilde{\Phi}$。对 3 个特征图进行高斯滤波和 sigmoid 激活函数处理,并与真实标签进行 loss 计算之后再反向传播。

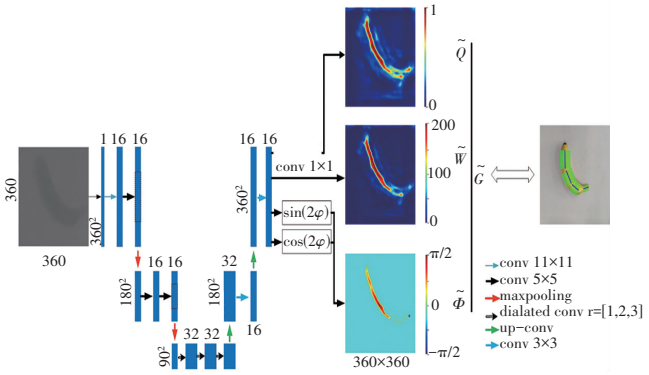

图 3　抓取检测网络 M 的整体结构

Fig. 3　The overall architecture of the grasping detection network M

3.3　注意力机制

模型的输入包含 4 个维度$[B,C,H,W]$,但模型自身并不能判别输入训练样本的“好坏”,因此注意力机制添加在$[C,H,W]$ 3 个维度上,分为通道注意力机制(SENet),空间注意力机制和两者组合(Coordinate Attention)。注意力模块类似于一个计算单元,可应用在网络的任一中间张量。当模块的输入为张量 $\boldsymbol{X}=[X_1,X_2,\cdots,X_C]\in\mathbb{R}^{C\times H\times W}$,输出为同尺寸的张量 $\boldsymbol{Y}=[Y_1,Y_2,\cdots,Y_C]$,那么第 c 个通道的全局平均池化可以表示为

$$y_c=\frac{1}{H\times W}\sum_{i}^{H}\sum_{j}^{W}x_c(i,j)$$

全局平均池化将全局空间信息压缩到通道维度中,但很难体现空间信息。文献[14]提出了一种将空间维度信息嵌入到通道维度的移动网络注意力机制 Coordinate Attention(CA)。CA 模块分别沿着高度 W 和宽度 H 对输入进行池化操作,将全局池化分解为一对特征编码,使用池化核尺寸为$[W,1]$对垂直坐标 H 进行编码,高度 h 处的第 c 个通道的池化可以表示为

$$y_c^h=\frac{1}{W}\sum_{0\leqslant j<W}x_c(h,j)$$

同理,宽度 w 处的第 c 个通道可以表示为

$$y_c^w=\frac{1}{H}\sum_{0\leqslant i<H}x_c(i,w)$$

高度和宽度两个维度的特征聚合使得注意力模型可以获得空间信息中的远程依赖关系,并保留精确的

位置信息。当输入尺寸为 $C\times H\times W$ 张量时,CA 模块的网络结构如图 4 所示。

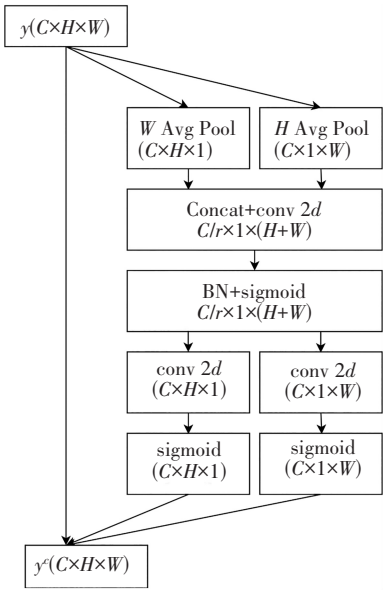

图 4 Coordinate Attention 网络结构

Fig. 4 Network structure of Coordinate Attention

经过 sigmoid 函数处理,得到尺寸为 $C\times H\times 1$ 和 $C\times 1\times W$ 两个特征图并将其分别命名为 y^h 和 y^w,两个特征图的值为[0,1]。CA 模块的最终输出如式(8)所示:

$$y^c = y\times y^h \times y^w \tag{8}$$

Coordinate Attention 模块在图像分类,目标检测以及图像分割任务中都有较好的表现。将 CA 模块添加在特征提取网络(空洞卷积层除外)和特征融合网络中。其对模型的改进效果将在实验部分详细阐述。

3.4 损失函数

对于抓取检测网络 M 而言,其数据集包括:输入图片 $I=[I_1,I_2,\cdots,I_n]$ 及相应标签文件 $L=[L_1,L_2,\cdots,L_n]$。其中 I_i 表示通过已知内参的深度相机采集的深度图片;L_i 表示编码后的抓取姿态,包含位置 p,抓取宽度 w 以及抓取角度[$\sin(2\varphi)$,$\cos(2\varphi)$]。

将输入深度图片和预测值的映射关系视为分类问题,即对输入图片中每个像素都进行分类处理,参数 y 表示对标签文件编码后得到的抓取姿态参数,参数 $\hat{y}$ 与之对应,表示模型的预测值。给定一个参数 p,当像素点为正样本时,p 等于模型的预测值 $\hat{y}$

$$p=\begin{cases}\hat{y} & \text{if } y=1\\ 1-\hat{y} & \text{otherwise}\end{cases}$$

使用 Focal loss 函数作为抓取网络模型 M 的损失函数,定义损失函数如式(9)所示:

$$Loss(p)=-\alpha_t(1-p)^r\log(p) \tag{9}$$

式(9)中:α_t 用于解决正负样本不均衡的问题;$(1-p)^r$ 用于解决难样本的问题,增大难样本在损失函数中的比重,$\log(p)$ 为交叉熵损失函数。α_t 和 r 均为超参数,设置 $\alpha_t=0.9$,$r=2$。

为了提高抓取检测网络的性能,在网络中增加注意力机制,以提高网络的检测精度。除了预测网络部分,模型中的其他部分分别使用 U-net 和 Deeplabv3 网络进行了替换,期望抓取置信度和检测精度能得到进一步提升。通过将推理时间与检测准确率相结合的方法来评估各网络模型的检测结果,最终得到了速度和精度兼顾的模型。

4 仿真实验及结果分析

4.1 数据集的标定及编码

康奈尔数据集是通过深度相机在现实世界中采集的,含有 240 个不同对象的 885 张照片,每张图片分辨率为 640×480,包含 RGB 图像及其相应点云信息。抓取检测网络 M 的数据集由深度图片 $I\in R^{H\times W}$ 和抓取集合 $\tilde{g}\in R^{3\times H\times W}$ 组成,使用 Realsense d435i 相机采集场景中物体的 RGB 图片和深度图片,深度图片作为网络 M 的输入,将对应的 RGB 图片使用文献[8]中的方法进行标注,如图 5 所示。

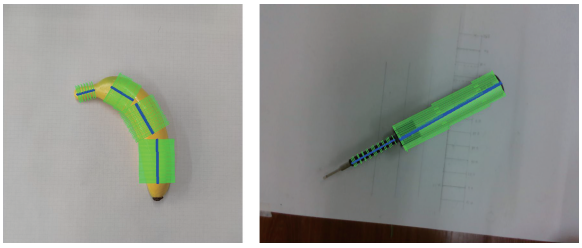

图 5 数据集标注示意图

Fig. 5 View of dataset annotation

为方便模型训练并提高其泛化能力,须对数据集进行规范化处理和数据增强。数据集输入到模型 M 之前,使用中心裁剪将深度图片及标签尺寸裁剪为 360×360 pixels,然后执行随机缩放、裁剪和旋转等数据增强操作并在数据增强的过程中调整深度图片的深度值。

标注后的抓取集合 $\tilde{\boldsymbol{g}}\in R^{3\times H\times W}$ 包含 3 个通道,分别执行如下的编码操作:

(1) 抓取置信度。将抓取置信度标签 $\tilde{Q}^{\mathrm{T}}$ 作为二分类问题来处理,抓取点的值设为 1,其他点的值设为 0。模型 M 预测出的抓取置信度在特征图中数值越趋近 1 表明抓取成功率越高。

(2) 抓取宽度 $\tilde{W}^{\mathrm{T}}$。数据集中 $\tilde{W}^{\mathrm{T}}$ 的最大值为 120 pixels,也就是 $\tilde{W}$ 的取值范围为[0,120]。为避免数据增强过程时抓取宽度的值超过 120,设置参数 $\omega=$

200,训练时将 $\tilde{W}^{\mathrm{T}}$ 的值按照 $1/\omega$ 的比例缩放到区间[0,1)。抓取置信度 $\tilde{Q}^{\mathrm{T}}=1$ 处设置对应的抓取宽度 $\tilde{W}^{\mathrm{T}}$,未标记处的抓取宽度设为 0。

(3) 抓取角度 $\tilde{\Phi}^{\mathrm{T}}$。抓取角度 $\tilde{\Phi}$ 具有对称性,模型对抓取角度的检测也应具有连续性,因此将抓取角度的范围设为 $[-\pi/2,\pi/2]$,并将其编码为 $\sin(2\tilde{\Phi}^{\mathrm{T}})$ 和 $\cos(2\tilde{\Phi}^{\mathrm{T}})$ 两个参数。当 $\tilde{\Phi}\in[-\pi/2,\pi/2]$ 时,$\sin(2\tilde{\Phi}^{\mathrm{T}})$ 和 $\cos(2\tilde{\Phi}^{\mathrm{T}})$ 的值与 $\tilde{\Phi}$ 对应且对称分布,利于网络训练。后处理时,通过三角函数得到抓取宽度 $\tilde{\Phi}$。与抓取宽度类似,在抓取置信度 $\tilde{Q}^{\mathrm{T}}=1$ 处设置对应的 $\sin(2\tilde{\Phi}^{\mathrm{T}})$ 和 $\cos(2\tilde{\Phi}^{\mathrm{T}})$,其他点处值为 0。

4.2　网络的输出和评估方法

以单通道的深度图片作为抓取检测网络 M 的输入,检测模型在检测精度和速度下取得较好的平衡。图 6 显示了部分场景下网络的输出特征图,前 3 行分别为抓取置信度、抓取宽度和抓取角度,最后一行为最优抓取位姿。

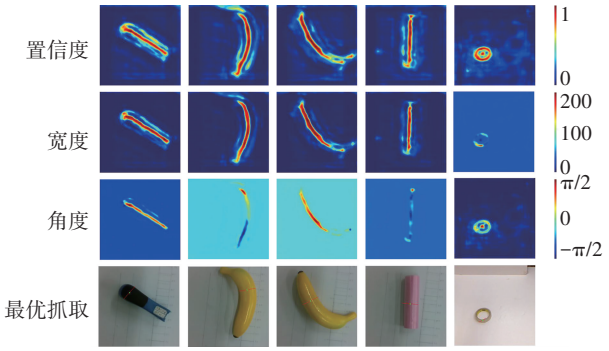

图 6　抓取检测网络的输出

Fig. 6　The output of the grasping detection network

训练时将数据集按 8∶1∶1 的比例划分为训练集、验证集和测试集。GG-CNN2 原论文中采用 IoU 度量评价方法,定义两个参数 $\tilde{Q}_{\max}$ 和 A_{cc} 用于评估抓取检测网络 M 的预测结果,其中 $\tilde{Q}_{\max}$ 为预测网络的特征图 $\tilde{Q}$ 中的最大值,$\tilde{Q}\in[0,1]$,值越趋近 1 表明抓取成功率越大;参数 A_{cc} 使用了网络全部输出参数进行评估,当输出同时满足以下两个条件时,认为预测正确:

(1) 抓取点的预测值与标签的距离小于 5 像素;

(2) 抓取宽度的预测值与标签之比小于 0.8;

(3) 抓取角度的预测值与标签的误差小于 $\pi/6$。

4.3　对比试验

为进一步改进抓取检测方法的性能,在模型 M 中添加注意力机制 CA,并将 M 的主体网络分别替换为 U-net 和 Deeplabv3 语义分割网络,从最大置信度 $\tilde{Q}_{\max}$ 和抓取成功率 A_{cc} 两个指标对 4 种算法进行比较分析。

U-net 是一个经典的语义分割模型,前半部分进行特征提取,后半部分进行上采样,网络的结构左右对称,并将深层与浅层的特征图通过通道数进行拼接,实现浅层的位置信息与深层的语义信息相融合。

Deeplabv3 也是一个语义分割模型[15],被认为是语义分割的新高峰。方法优点有:在特征提取网络中使用了多层的空洞卷积,在参数不变的情况下增加模型的感受野,使得每个卷积层的输出包含较大范围的信息,并将推理时间和检测精度平衡的很好;采用 ASPP 结构,将多个不同膨胀率的空洞卷积层并联,使得模型在多尺度物体上的表现更好。

采用 pytorch 框架实现抓取检测网络 M 的搭建,CA 注意力机制的添加,以及主干网络的替换,使用 4.1 节中的抓取数据集对 4 种算法进行了训练了 1 000 epoch,batch-size 的设置为 8,选用 Adam 优化器,初始学习率 η 为 0.001 并使用 MultiStepLR 的衰减方法。使用 $\tilde{Q}_{max}$ 和 A_{cc} 两个参数对算法进行评估,训练结果如图 7 和图 8 所示。

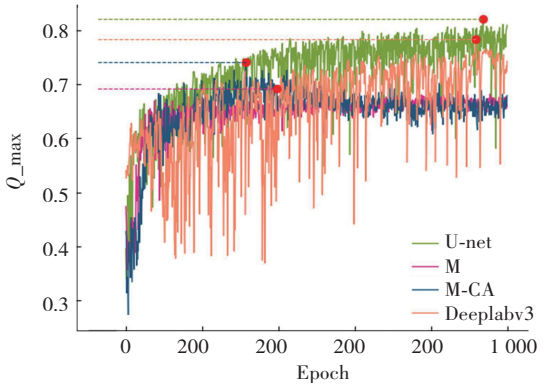

图 7　最大抓取置信度 $\tilde{Q}_{\max}$

Fig. 7　The maximum grasping confidence $\tilde{Q}_{\max}$

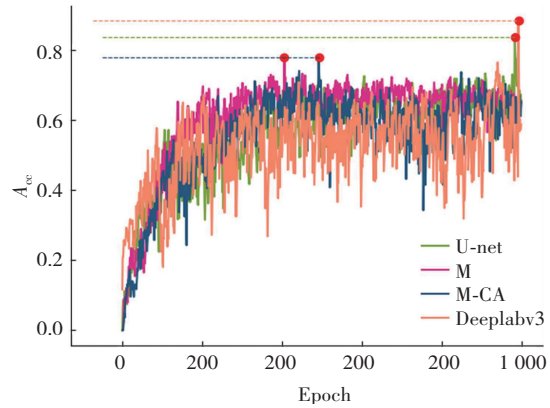

图 8　预测成功率 A_{cc}

Fig. 8　The prediction success rate A_{cc}

在装有 RTX 3060 的电脑上进行了训练和模型推理,具体的结果如表 1 所示。

表 1 抓取检测模型的性能对比

Table 1 The performance comparison of grasping detection networks

算 法	推理时间/ms	$\tilde{Q}_{\max}$	A_{cc}/%
模型 M	2.7	0.69	78
模型 M_{CA}	5.2	0.74	78
U-net	37.5	0.82	84
Deeplabv3	41.7	0.78	88

结果表明:就推理时间而言,加入注意力机制前后的模型 M 均为毫秒级,U-net 和 Deeplabv3 网络的检测速度较慢;在抓取置信度 $\tilde{Q}_{\max}$ 这一指标中,U-net 网络的表现最好,相比于模型 M 提升了 18.8%;在抓取检测任务的准确率 A_{cc} 中,Deeplabv3 网络的表现最好,相比于模型 M 提升了 12.8%。机械臂抓取任务由抓取检测,抓取运动规划和抓取控制系统组成,因此抓取检测网络需兼顾模型检测的速度和准确率,根据表 1 中数据认为添加注意力机制后的模型 M 效果最好,将这一方法(M_{CA})移植到机械臂抓取平台进行抓取实验。

4.4 机械臂视觉抓取平台

机械臂视觉抓取平台如图 9 所示。配有控制柜的安诺 J601A 机械臂、电动夹爪、Realsense d435i 相机。抓取实验中深度图片和 RGB 图片的大小均为 640×480 pixels,抓取对象有瓶子、香蕉、苹果、订书机等。

图 9 安诺机械臂抓取平台

Fig. 9 Robot experiment setup with an Anno robot arm

抓取检测模型 M 得到的结果 $\tilde{\boldsymbol{g}}$ 位于像素坐标系下,$\tilde{\boldsymbol{g}}$ 由 4 个参数组成:

$$\tilde{\boldsymbol{g}}=[u,v,\tilde{\varphi},\tilde{W}]$$

开展抓取实验前,需分别进行相机内参标定与手眼标定实验,经过一系列坐标转换得到机械臂坐标系下的抓取位姿 g。内参标定得到的内参矩阵$\boldsymbol{t}_{CI}$,深度信息 Z_C 可根据深度图得到,将像素点坐标(u,v)对应到相机坐标系下的三维坐标$[X_C,Y_C,Z_C]$,如式(10)所示:

$$Z_C\begin{bmatrix}u\\v\\1\end{bmatrix}=\begin{bmatrix}f_x&0&u_0&0\\0&f_y&v_0&0\\0&0&1&0\end{bmatrix}\begin{bmatrix}X_C\\Y_C\\Z_C\\1\end{bmatrix} \tag{10}$$

根据张正友棋盘格标定法得到 d435i 深度相机的内参矩阵如下:

$$\boldsymbol{t}_{CI}=\begin{bmatrix}606.825&0&321.103\\0&608.296&245.419\\0&0&1\end{bmatrix}$$

通过手眼标定得到相机与机械臂之间的坐标关系$\boldsymbol{t}_{WC}$,将相机坐标系下坐标转换至世界坐标系下$[X_W,Y_W,Z_W]$,如式(11)所示。机械臂抓取平台采用“眼在手上”的方式,通过相应的功能包进行标定实验,标定结果为相机坐标系与机械臂终端之间的位置关系。

$$\begin{bmatrix}X_W\\Y_W\\Z_W\\1\end{bmatrix}=\begin{bmatrix}R_{3\times3}&T_{1\times3}\\\vec{0}&1\end{bmatrix}\begin{bmatrix}X_C\\Y_C\\Z_C\\1\end{bmatrix} \tag{11}$$

式(11)中:R 与 T 分别为相机坐标系与机械臂坐标系之间的旋转矩阵和平移矩阵,使用四元数来描述旋转矩阵 $R_{3\times3}$,手眼标定实验结果如下:

$$R=[0.707,-0.707,0,0]$$

$$T=[0.064,0.033,-0.011]$$

通过式(10)和式(11),抓取点(u,v)已转换至机械臂坐标系$[X_C,Y_C,Z_C]$;世界坐标系下绕 Z 轴旋转角度 φ 与 $\tilde{\varphi}$ 相同;抓取宽度 $\tilde{W}$ 单位为像素,由式(10)可求得实际的抓取宽度 w。从而场景中目标物体的平面抓取位姿 $\boldsymbol{g}$ 如式(12)所示:

$$\boldsymbol{g}=[X_C,Y_C,Z_C,\varphi,w] \tag{12}$$

在抓取检测过程中,首先使用 d435i 相机获取场景信息并对 RGB 图和深度图进行对齐;再将深度图片输入模型 M 得到目标物体的最优抓取位姿$\tilde{\boldsymbol{g}}^*$;经过一系列坐标转换得到机械臂底座坐标系下的抓取位姿$\boldsymbol{g}^*$;使用工作空间规划即可完成抓取任务。

5 结 论

基于像素点抓取表示对平面抓取任务进行了研究,提出了一种全卷积抓取检测方法,并在此基础上进行了改进研究,最后实现了机械臂快速抓取检测。结

论如下:

(1) 相比于方向矩形框表示方式,像素点抓取表示无需先验框和预测夹爪物理尺寸 m,更贴合机械臂的抓取任务。

(2) 抓取检测网络需兼顾检测速度和检测精度,相比于抓取网络 M,U-net 网络的抓取置信度 $\tilde{Q}_{\max}$ 提高了 18.8%,Deeplabv3 网络的准确率 A_{cc} 提高了 12.8%,但模型的检测速度较慢。根据结果对比分析,加入注意力机制的模型 M_{CA} 更适合平面抓取任务。

(3) 将抓取检测方法 M_{CA} 移植到安诺机械臂视觉抓取平台上,可以实现对若干典型物体的实时抓取。

参考文献(References):

[1] 刘亚欣, 王斯瑶, 姚玉峰, 等. 机器人抓取检测技术的研究现状[J]. 控制与决策, 2020, 35(12): 2817—2828.
LIU Ya-xin, WANG Si-yao, YAO Yu-feng, et al. Current status of research on robot grasp detection technology [J]. Control and Decision, 2020, 35(12): 2817—2828.

[2] 李世裴, 韩家哺, 路博凡, 等. 视觉引导下协作机器人抓取技术研究[J]. 重庆工商大学学报(自然科学版), 2022, 39(1): 42—48.
LI Shi-pei, HAN Jia-fu, LU Bo-fan, et al. Research on vision-guided collaborative robot grasping technology [J]. Journal of Chongqing University of Technology and Business (Natural Science Edition), 2022, 39(1): 42—48.

[3] GUO D, KONG T, SUN F C, et al. Object discovery and grasp detection with a shared convolutional neural network [C]//proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2016: 2038—2043.

[4] 王勇, 陈荟西, 冯雨齐. 基于改进 CenterNet 的机械臂抓取检测[J]. 中南大学学报(自然科学版), 2021, 52(9): 3242—3250.
WANG Yong, CHEN Hui-xi, FENG Yu-qi. Improved CenterNet-based robotic arm grasping detection[J]. Journal of Central South University (Natural Science Edition), 2021, 52(9): 3242—3250.

[5] YANG J Y, CHEN U K, CHANG K C, et al. A novel robotic grasp detection technique by integrating YOLO and grasp detection deep neural networks [C]//2020 International Conference on Advanced Robotics and Intelligent Systems (ARIS). IEEE, 2020: 1—4.

[6] MAHLER J, MATL M, LIU X, et al. Dex-net 3.0: Computing robust vacuum suction grasp targets in point clouds using a new analytic model and deep learning[C]//2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018: 5620—5627.

[7] MORRISON D, CORKE P. Learning robust, real-time, reactive robotic grasping [J]. The International Journal of Robotics Research, 2020, 39(2): 183—201.

[8] WANG D X, LIU C S, CHANG F L, et al. High-performance pixel-level grasp detection based on adaptive grasping and grasp-aware network[C]//IEEE Transactions on Industrial Electronics. IEEE, 2022: 11611—11621.

[9] CHU F J, XU R N. Real-world multiobject, multigrasp detection[J]. IEEE Robotics and Automation Letters, 2018, 3(4): 3355—3362.

[10] SHI C Q, MIAO C X, Zhong X G, et al. Pixel-level grasp detection for unknown objects with encoder-decoder-inception deep network[C]//2022 IEEE 5th International Conference on Electronics Technology (ICET). IEEE, 2022: 1153—1157.

[11] CAO H, CHEN G, LI Z J, et al. Residual squeeze-and-excitation network with multi-scale spatial pyramid module for fast robotic grasping detection[C]// Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021: 13445—13451.

[12] ZENG A, SONG S R, YU K T, et al. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching[J]. The International Journal of Robotics Research, 2022, 41(7): 690—705.

[13] WANG P Q, CHEN P F, YUAN Y, et al. Understanding convolution for semantic segmentation[C]//Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018: 1451—1460.

[14] HOU Q B, ZHOU D Q, FENG J S. Coordinate attention for efficient mobile network design [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 13713—13722.

[15] CHEN L C, Zhu Y, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 801—818.

责任编辑:田　静