DOI: 10.19650/j.cnki.cjsi.J2006162

基于深度学习的机器人最优抓取姿态检测方法*

李秀智[1,2],李家豪[1,2],张祥银[1,2],彭小彬[1,2]
(1.北京工业大学信息学部 北京 100124; 2.计算智能与智能系统北京市重点实验室 北京 100124)

摘 要:服务型机器人在抓取任务中面临的是非结构化的场景。由于物体放置方式的不固定以及其形状的不规则,难以准确计算出机器人的抓取姿态。针对此问题,提出一种双网络架构的机器人最优抓取姿态检测算法。首先,改进了 YOLO V3 目标检测模型,提升了模型的检测速度与小目标物体的识别性能;其次,利用卷积神经网络设计了多目标抓取检测网络,生成图像中目标物体的抓取区域。为了计算机器人的最优抓取姿态,建立了 IOU 区域评估算法,筛选出目标物体的最优抓取区域。实验结果表明,改进后的 YOLO V3 目标检测精度达到 91%,多目标抓取检测精度达到 86%,机器人最优抓取姿态检测精度达到 90%以上。综上所述,所提方法能够高效、精确地计算出目标物体的最优抓取区域,满足抓取任务的要求。

关键词: 深度学习;目标检测;抓取检测;机器人最优抓取
中图分类号: TP242 TH86 **文献标识码**: A **国家标准学科分类代码**: 510.40

Detection method of robot optimal grasp posture based on deep learning

Li Xiuzhi[1,2], Li Jiahao[1,2], Zhang Xiangyin[1,2], Peng Xiaobin[1,2]
(*1.Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China;*
2.Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing 100124, China)

Abstract: The service robot is faced with unstructured scene in the task of grasp. Because of the irregular placement and shape of the objects, it is difficult to accurately calculate the robot's grasp posture. Aiming at this problem, a robot optimal grasp posture detection algorithm with dual network architecture is proposed. Firstly, the YOLO V3 target detection model is improved, which improves the detection speed of the model and the recognition performance of small target objects. Secondly, convolutional neural network is used to design multi-target grasp detection network, which generates the robot grasp area in the image. In order to calculate the optimal grasp posture of the robot, the IOU area evaluation algorithm is established, which screens out the optimal grasp area of the target object. The experiment results show that the target detection accuracy of improved YOLO V3 reaches 91%, and the detection accuracy of the multi-target grasp reaches 86%, the detection accuracy of the robot optimal grasp posture reaches above 90%. In summary, the proposed method can efficiently and accurately calculate the optimal grasp area of the target object to meet the requirements of the grasp task.

Keywords: deep learning; object detection; grasp detection; robot optimal grasp

0 引 言

视觉检测与定位是智能机器人伺服控制的关键技术之一[1]。深度学习以其突出的目标检测精度和可靠性,成为近年来机械手臂伺服抓取控制领域的研究热点[2]。

随着研究的深入,如何准确确定不规则目标的抓取姿态,一度成为智能抓取的技术瓶颈。基于点云的抓取姿态估计算法(PointNet-grasp pose detection, PointNet-GPD)[3],使用 3D 神经网络 PointNet[4] 进行抓取姿态估计,这种方法依赖于物体的点云信息,实际应用中易受物体遮挡、截断等因素的影响。相比之下,基于 Faster RCNN(faster region based convolutional neural networks)网络[5]的多物体抓取模型[6]具有较好的泛化能力和检测精度;Asif 等[7]从图像的不同层级来预测抓取区域,克服了只从单个层级预测图像抓取区域的局限性,就准确度而

收稿日期:2020-03-06 Received Date:2020-03-06
*基金项目:国家自然科学基金(61703012)、北京自然科学基金(4182010)项目资助

言,该方法优于 Cornell 抓取数据集[8]上的最新方法。

除了识别可抓取区域,目标识别的类别信息同样不可或缺。陈丹等[9]通过联立目标检测框与物体的最小包围矩形[10]计算出目标物体的抓取区域,该方法只是计算出物体相对于图像的旋转位姿,并没有考虑该物体的最优抓取区域,对于不规则形状物体的抓取有一定的局限性。金欢[11]利用目标检测算法识别目标物体,将原始图像分割为多个仅包含单个物体的小块,再利用基于深度学习的抓取检测网络完成抓取任务,此方法采用级联式的网络结构,损失了机器人抓取算法的运行效率。

针对上述问题,本文提出了一种双网络架构的机器人最优抓取姿态检测算法。首先,通过改进 YOLO (you only look once) V3[12]的网络结构,提升了模型的运行效率与小目标物体识别的性能;其次,优化了文献[6]的网络结构,设计了一个轻量级多目标抓取检测网络。针对背景与抓取目标区域存在的互斥关系,提出了交并比(intersection over union,IOU)抓取区域评估算法,使机器人能够在目标物体位置上计算出最优抓取区域,提升机器人抓取姿态的计算效率。本文在实际场景下进行了实验验证,根据已标定的手眼关系进行坐标转换,实现执行末端的抓取控制,验证了上述检测算法的有效性。

1　抓取姿态检测方法总体框架

机器人抓取方法总体框架如图 1 所示。首先,使用深度相机 D435 获取场景中的深度图与彩色图,同时将深度图与彩色图进行配准;其次,采用目标检测与抓取检测双网络架构生成场景中的物体边界与抓取矩形;最后,以目标物体的边界信息为基准,计算出与其 IOU 最大值的候选抓取矩形,并利用坐标映射关系,将该抓取矩形映射成三维空间中的机器人抓取姿态。

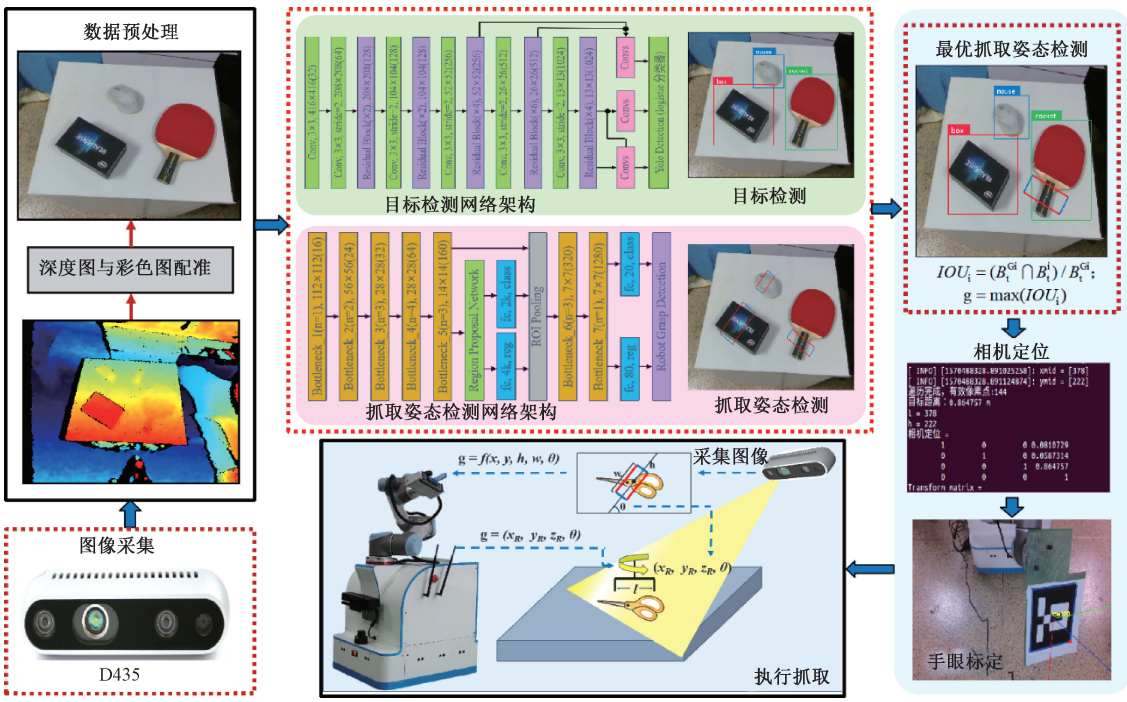


图 1　机器人抓取方法总体框架

Fig.1　General framework of robot grasping method

抓取检测问题不同于普通检测问题的回归边界任务,需要预测出目标物体的最优抓取区域[13]。为了对场景中潜在的抓取姿态进行预测,可以将抓取检测表述为在图像中找到最佳抓取区域 g,这种执行抓取的方式被定义为机器人五维抓取,g 可表示为:

$$g = f(x, y, h, w, \theta) \tag{1}$$

式中:(x, y)是图像中抓取矩形中心点坐标;(h, w)是矩形框的宽和高;θ 是抓取矩形框相对于水平轴的夹角。利用三维成像模型确定出抓取矩形对应的机器人最优抓取姿态。

2　YOLO V3 目标检测

2.1　YOLO V3 目标检测框架

YOLO 目标检测器的核心思想是将目标识别转化为回归任务来解决,绕开了候选区域的生成与评估,以端到

端的方式极大地提高了计算效率。由 Redmon 等[14]最新提出的 YOLO V3 目标检测,在特征提取网络、多尺度预测、多标签分类等方面进行了改进,成为目前最先进的目标检测器之一。YOLO V3 主体部分是由特征提取网络与多尺度检测器两部分组成,其整体框架如图 2 所示。特征提取网络由高质量图像分类的标准架构截断而成,主要用于提取图像特征;而多尺度检测器在特征提取网络的基础上进一步抽象和融合后得到 3 个尺度的预测特征,在 3 个尺度上进行检测,最终输出图像中物体类别与检测框坐标。

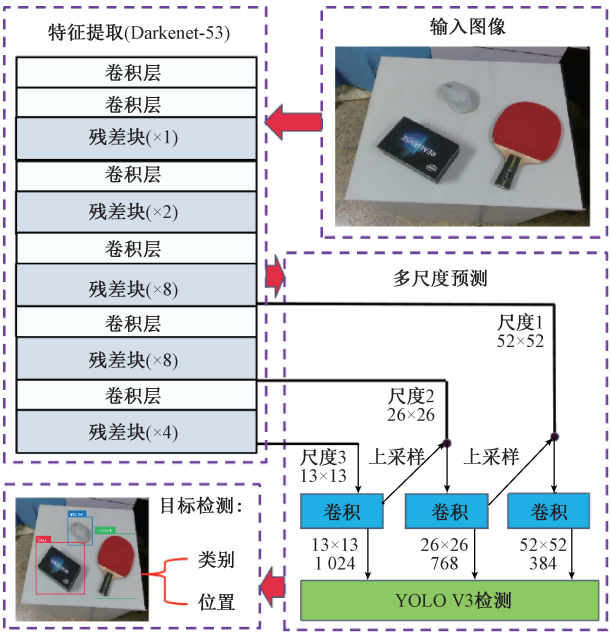


图 2 YOLO V3 网络结构

Fig.2 The network structure of YOLO V3

由图 2 可知,YOLO V3 的特征提取网络是 Darknet-53。该卷积神经网络借鉴残差结构[15],在网络中大量使用 1×1 卷积核,以及利用步长为 2 的 3×3 卷积核替代最大池化层等方式减小了网络参数量。此外,YOLO V3 所采用的框架是基于 C 语言的神经网络框架 Darknet。虽然 YOLO V3 通过网络结构的优化来提升模型的计算效率,但较为庞大的网络还是给模型运行速度带来沉重的负担。在单阶段目标检测算法中,围绕效率对模型进行调整是常用的手段。例如,提出 YOLO V3 的作者另外提出了更加轻量化的版本 YOLO V3-tiny,该网络在 YOLO V3 的基础上删去一些层,只保留两个预测分支,虽然牺牲目标检测精度,但大幅提升了网络的运算速度,YOLO V3 与其他网络在 COCO 数据集上的性能对比[12]如表 1 所示。

计算量单位为每秒浮点运算次数(floating point operations per second, FLOPS)。从表 1 可知,特征提取网络层数与计算量对网络的推断速度影响较大。虽然特征提取网络 Darknet-53 使 YOLO V3 在 COCO 等公开数据集上精度提升显著,但是其 53 层卷积层在数据集较小的检测任务中,对比 YOLO V3-tiny 的 12 层以及 YOLO V2 的 19 层特征提取网络显得较为冗余。

表 1 COCO 数据集上的性能对比

Table 1 Performance comparison on the COCO dataset

网络	层数	平均精度	计算量	帧率
YOLO V3(416)	53	55.3	65.86	35
YOLO V3-tiny(416)	12	33.1	5.56	220
YOLO V3 (608)	53	57.9	140.69	20
YOLO V2 (608)	19	48.1	62.94	40

2.2 网络结构改进

卷积神经网络在结构上存在的特性:最开始浅层网络卷积出的特征图较大,维度较少。网络越深,卷积出的特征图越小,维度越大。根据卷积层参数量计算式(2)可知,增加浅层网络层数对整个网络的参数量影响较小。

$$parameters = (K_h * k_w * C_{in}) * C_{out} + C_{out} \quad (2)$$

式中:K_h 与 K_w 是卷积核的宽高;C_{in} 与 C_{out} 分别是输入通道数与输出通道数。

在浅层网络生成的特征图中,每个像素点对应的感受野重叠区域较小,使网络能够捕获图像更多的细节。由陈丹等[9]提出的改进型 Faster RCNN 模型可知,通过对浅层网络进行扩充,使网络不容易忽略成像小的目标物体特征,提升小目标物体的检测性能。戴伟聪等[16]利用 YOLO V3 进行单目标物体检测时,对 Darknet-53 进行裁剪,另根据数据集特性对网络结构进行调整,使改进后的 YOLO V3 模型检测精度与速度都有所提升;并在文中指出,简单的卷积神经网络模型具有更好的泛化性,尤其是在数据集较小、数据复杂多变的情况下,且 YOLO V3 参数量大,在较小数据集的检测任务中容易过拟合。借鉴上述思想,本文提出了参数量与层数较少,运算复杂度较低的卷积神经网络 Darknet-43。

图 2 中,Darknet-53 的特征提取器共由 5 个部分组成,每部分残差块的数量分别是 1、2、8、8、4。本文所改进的 Darknet-43 模型也是由多个残差块经典结构组合而成,但在结构上做出了调整,具体结构如表 2 所示。首先,第 1 部分扩充为两个残差块。因为第 1 部分的输出特征图较大,在该部分卷积层输出的特征图中包含更多的特征信息,增加残差块的目的是为了使浅层卷积层的特征提取更加充分;其次,第 3 部分与第 4 部分,分别设计为 4 个残差块与 6 个残差块,相较于原模型分别减小了 4 个残差块与 2 个残差块。考虑到第 4 部分输出的特

征直通到 26×26 的预测分支上,该分支卷积模块输出的特征会参与到另外两个尺度预测分支的特征拼接,所以在该部分中保留更多的残差模块,会使整体的网络结构保持更好的特征提取性能。

表 2　Darknet-43 网络结构

Table 2　The network structure of Darknet-43

		卷积核数量	卷积核大小	输出特征
	卷积层	32	3×3	416×416,32
	卷积层	64	3×3,步长=2	208×208,64
残差块×2	卷积层	32	1×1	104×104,64
	卷积层	64	3×3	
	残差连接			
	卷积层	128	3×3,步长=2	104×104,128
残差块×2	卷积层	64	1×1	104×104,128
	卷积层	128	3×3	
	残差连接			
	卷积层	256	3×3,步长=2	52×52,256
残差块×4	卷积层	128	1×1	52×52,256
	卷积层	256	3×3	
	残差连接			
	卷积层	512	3×3,步长=2	26×26,512
残差块×6	卷积层	256	1×1	26×26,512
	卷积层	512	3×3	
	残差连接			
	卷积层	1 024	3×3,步长=2	13×13,1 024
残差块×4	卷积层	512	1×1	13×13,1 024
	卷积层	1 024	3×3	
	残差连接			

改进后的 Darknet-43 在网络的深度、参数量、计算量相较于 Darknet-53 都大幅下降。该模型层数为 43 层,比 Darknet-53 减少了 10 层。其参数量约有 3.65×10^7 个,比 Darknet-53 约少了 4.1×10^6 个。在输入图像尺寸为 416×416 的情况下,本文所改进的 Darknet-43 模型 FLOPS 为 47.9,比 Darknet-53 约少 10.96。

为了保证小物体的识别率,本文仍采用 YOLO V3 结构中浅层与高层相融合的策略,将第 19 层与 52×52 预测分支的卷积层输出的特征进行特征拼接;将第 33 层与 26×26 预测分支的卷积层输出的特征进行特征拼接,使网络同时学习深层和浅层特征,表达效果更好。

针对 YOLO V3 目标检测算法对于不同角度、尺度或新的物体泛化性较差的问题,本文在训练时,通过对输入图像进行数据增广,使得算法在物体可能的方向、比例尺寸与光照影响等问题上具有更强的鲁棒性。随着不同数据集图像中物体特征尺度的变化,训练时先验框的尺寸也需要进行调整。所以本文对训练数据采用 K-means 聚类算法,聚类出 9 种尺度的先验框。

3　多目标抓取检测网络

3.1　多目标抓取检测网络结构的设计

本文将抓取检测视为抓取角度分类与抓取位置的回归,并借鉴基于区域提取的抓取检测网络结构[6],设计了轻量级多目标抓取检测模型。首先,选用轻量级卷积神经网络 Mobilenet V2[17] 提取图像特征,并在 Mobilenet V2 中间层(bottleneck_5 之后)插入候选推荐网络(region proposal network, RPN)[5] 网络。RPN 在生成的特征图上预测候选区域位置(锚框),并将生成的特征向量(锚框参数)分别送入到两个全连接层中,即分类层与回归层。分类层判断锚框是否为抓取区域;回归层负责预测锚框的位置。经过两个全连接层的处理,便可得到每个锚框的评价得分与回归坐标。锚框的评价得分和回归坐标分别由 2 个分数和 4 个坐标值表示。其中,2 个分数用于判断锚框是否为抓取区域;4 个坐标表示锚框的物理参数,记为(x, y, w, h)。x、y 表示锚框中心点坐标,w、h 表示锚框的宽高。在预测图像中所有抓取区域时,引入损失函数:

$$L_{\text{gpn}}(\{(p_i,t_i)_{i=1}^{\text{I}}\}) = \sum_i L_{\text{gp\_cls}}(p_i,p_i^*) + \lambda\sum_i p_i^* L_{\text{gp\_reg}}(t_i,t_i^*) \tag{3}$$

式中:$L_{\text{gp_cls}}$是交叉熵损失函数,用于分类;$L_{\text{gp_reg}}$是回归损失函数,用于预测位置;λ 表示权重;I 表示所有候选区域的集合;i 表示在小批样本中候选区域的索引,$p_i^*=1$ 表示锚框 i 为正标签,即抓取区域,$p_i=0$ 时表示 i 为负标签,不是抓取区域;t_i 表示锚框的参数;t_i^* 表示正标签锚框 i 映射到图像中的坐标向量。将锚框与 Mobilenet V2 中间层特征用 ROI Align 进行特征归一化处理,继而将其送入 Mobilenet V2 剩余的网络继续提取特征。在末端接入两个全连接层,用于抓取框角度的分类与坐标位置的回归,定义抓取区域预测的损失函数如式(4)所示。

$$L_{\text{ger}}(\{(\rho_l,\beta_l)\}_{c=0}^{\text{C}}) = \sum_c L_{\text{ger\_cls}}(\rho_l) + \lambda_2\sum_c 1_{c\neq0}(c) L_{\text{ger\_reg}}(\beta_c,\beta_c^*) \tag{4}$$

式中:C 是角度类别数,共计 20 个角度间隔类别;ρ_l 是锚框为抓取矩形角度 l 的类别概率;β_l 为 ρ_l 对应的抓取边界框;$L_{\text{ger\_cls}}$ 是交叉熵损失函数,用于预测抓取角度所属类别;$L_{\text{ger\_reg}}$ 是抓取框坐标回归损失函数;λ_2 表示权重,用来均衡两个损失函数的大小;β_c^* 表示网络候选推荐框真

值。综上所述,总损失函数如下所示。

$$L_{total} = L_{gpn} + L_{ger} \tag{5}$$

多目标抓取检测网络结构如图 3 所示,借鉴区域提取二阶目标检测算法[5]的思想,首先判断 RPN 候选推荐网络生成多个抓取矩形框能否用于抓取物体;其次,预测抓取框的角度与其对应的边界参数。

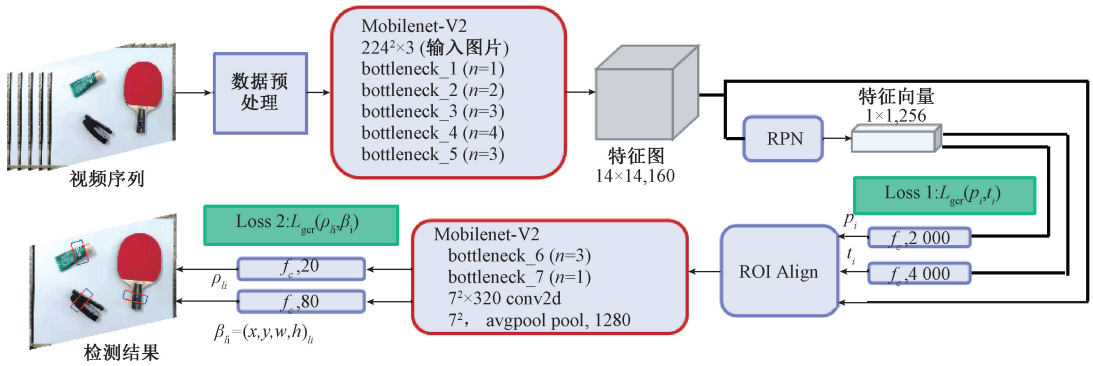


图 3 多目标抓取检测网络框架

Fig.3 Network framework of multi-target grasp detection

原模型[6]通过 ROI Pooling 层进行特征归一化处理,该层在量化阶段对数据取整操作,容易导致信息丢失。而 ROI Align 在重映射特征图时,采用双线性插值法进行计算,保留浮点数。这样的方法会保留更多的输入特征,所以选用 ROI Align 进行特征归一化处理。

3.2 RPN 锚框的设计

数据集中抓取矩形的长宽比都接近 1∶2或 2∶1的比例。因此,本文去除 RPN 预测长宽比为 1∶1的锚框,只保留 3 种尺度、长宽比为 1∶2、2∶1的锚框,如图 4 所示。

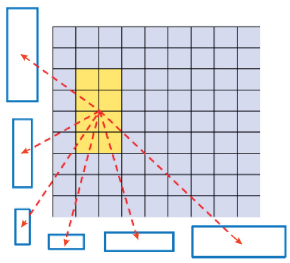


图 4 锚框机制

Fig.4 Anchor box mechanism

这种超参数的设计在保证精度的同时,可以有效地减小网络的参数量与计算量。

4 机器人最优抓取姿态检测算法

4.1 基于 IOU 区域评估的最优抓取姿态检测

由于在抓取检测任务中,背景是目标的补集,它们是互斥关系。只有以目标物体为背景计算出的抓取区域,机器人才能够以该抓取区域确定出机械夹爪的抓取姿态。综上,本文提出了一种双网络架构的机器人最优抓取姿态检测算法,该算法通过目标检测与抓取检测计算目标物体的最优抓取区域。

目标检测对图像中的目标物体进行识别和定位。抓取检测生成抓取矩形,获取图像中的候选抓取位置。以图像中待抓取物体的目标检测框为背景,计算抓取矩形与目标检测边界的 IOU,通过筛选出 IOU 最大值与在目标边界内的抓取矩形,即得出目标物体的最优抓取区域。完整的算法结构如下所示。

输入:深度图、彩色图//将深度图对齐到彩色图
DATA:t 帧通过 YOLO V3 预测的物体位置信息 B_t^i
DATA:t 帧抓取检测网络预测的抓取矩形信息 B_t^{Gi}
1: for each B_t^i in B_t do//遍历 t 帧上所有目标检测框
2: 确定 t 帧时刻图像中的抓取目标 B_t^i
3: 将 B_t^i 设为图像中抓取区域 ROI
4: if ($B_t^{Gi} \subset B_t^i$)
then g ← B_t^{Gi}
5: else $IOU_i = (B_t^{Gi} \cap B_t^i)/B_t^{Gi}$
6: if (IOU_i>0.7)
then g←max(IOU_i)
7: else go to 2:
8: end if
9: end if
10: $d_c \leftarrow f(x_c, y_c)$//g=($x_c$, y_c, w, h, θ);g 是抓取区域参数
11: if (d_c> threshold)//d_c 最大抓取距离
12: then go to 2:
13: else g ← (x_w, y_w, z_w, θ)//机器人最终抓取姿态
14: end if
15: end for

如上述伪代码所示,为了求解机器人的抓取姿态,首

先应将深度图对齐到彩色图,从而可利用坐标转换公式计算出像素点在机器人坐标系下的三维坐标值;其次,将彩色图像分别输入到目标检测模块与抓取检测模块中,分别得到目标物体的边界 B_t^i 与机械手可执行抓取区域 B_t^{Gi}。为了在抓取目标位置上选取最优的抓取区域,本文建立了 IOU 区域评估算法,如算法中的步骤 5)~9)所示。该算法以待抓取目标的边界为背景,计算抓取区域,如图 5 所示。

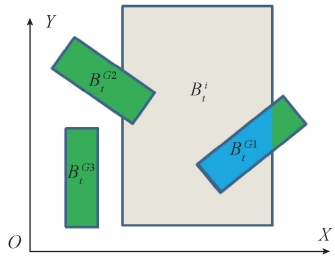
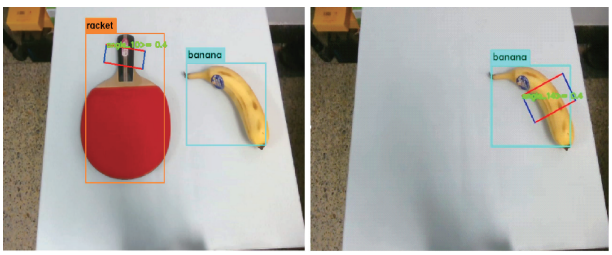


图 5　基于 IOU 的抓取区域评估

Fig.5　IOU based grasp area estimation

通过计算所有抓取区域与抓取目标边界的 IOU,当 $IOU>0.7$ 或 $B_t^{Gi} \subset B_t^i$ 时,将 B_t^{Gi} 视为 B_t^i 的可抓取区域,如图 5 中的抓取矩形 B_t^{G1} 所示。以该抓取矩形为基准,利用矩形中心像素值计算机器人末端执行器的三维抓取点,并以抓取矩形相对于图像中 X 轴的夹角作为机器人末端执行器的旋转角 θ,即可得到机器人的最优抓取姿态。最后,采用机器人抓取系统中的最大工作距离 d_c 判断生成抓取区域的可靠性。基于 IOU 区域评估算法结果如图 6 所示。



(a) 基于IOU区域评估　(b) 基于IOU区域评估

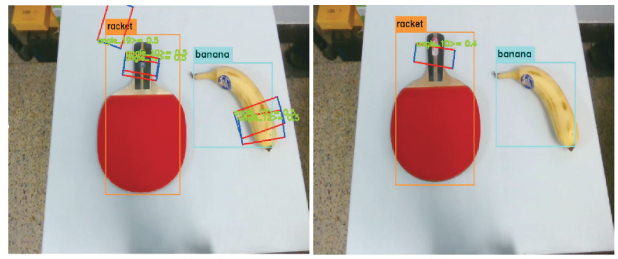
(a) Using regional evaluation of IOU　(b) Using regional evaluation of IOU

图 6　基于 IOU 区域评估算法的抓取检测结果

Fig.6　Using regional evaluation of IOU algorithm result graph

由图 6 可知,本文所提的机器人最优抓取姿态检测算法以目标物体为基准,计算该物体的最优抓取矩形。当场景中的物体被抓取后,再进行下一目标的最优抓取位姿计算,有利于机器人连续抓取。将本文算法与目标检测、抓取检测直接输出的结果作对比,如图 7 所示。

图 7(a)是无 IOU 区域评估算法的抓取检测与目标检测结果,抓取检测不能以抓取目标确认抓取位置,且抓取检测会受目标检测框的干扰。图 7(b)是基于 IOU 区域评估算法的结果。通过对比可知,本文所提的 IOU 区域评估算法有效的避免了背景对抓取检测的干扰,适用于非结构化的抓取场景。



(a)无IOU区域评估　(b) 基于IOU区域评估

(a) Without regional evaluation of IOU　(b) Using regional evaluation of IOU

图 7　不同抓取检测方法输出的对比

Fig.7　Comparison of different grasp detection method outputs

4.2　基于二维抓取矩形的机器人抓取方式

本文将图像中生成的最优抓取矩形作为机器人抓取姿态的参照。其中抓取矩形相对于 X 轴的夹角作为机器人抓取的旋转角度,抓取检测框的中心点视为机器人抓取点在视觉传感器上的二维映射。所以,在抓取任务中,还需计算出检测框中心点在机器人基座坐标系下对应空间点的三维坐标值。

首先,根据相机标定原理,计算出该像素点在相机坐标系下的三维坐标值。由于世界坐标系与相机坐标系相重合,因此在相机坐标系与世界坐标系下,同一物体具有相同的深度值。则世界坐标点 $M(X_C, Y_C, Z_C)$ 到像素点 $m(u, v)$ 的转换公式为:

$$Z_C\begin{bmatrix}u\\v\\1\end{bmatrix}=\begin{bmatrix}f/\mathrm{d}x & 0 & u_0\\0 & f/\mathrm{d}x & v_0\\0 & 0 & 1\end{bmatrix}\begin{bmatrix}X_C\\Y_C\\Z_C\end{bmatrix} \tag{6}$$

由式(6)的变换矩阵,可得到像素点 $m(u, v)$ 到相机坐标系下坐标点 $M(X_C, Y_C, Z_C)$ 的变换公式为:

$$\begin{cases}X_C = z \times (u - u_0) \times \mathrm{d}x/f\\Y_C = z \times (v - v_0) \times \mathrm{d}y/f\\Z_C = d\end{cases} \tag{7}$$

其次,联立相机与机器人的位置关系,计算出该点在机器人基座坐标系下的坐标值 (X_R, Y_R, Z_R),即机器人末端执行器抓取时的位置量,如式(8)所示。

$$\begin{bmatrix}X_R\\Y_R\\Z_R\\1\end{bmatrix}=\begin{bmatrix}\boldsymbol{R}_{RC} & \boldsymbol{T}_{RC}\\0^T & 1\end{bmatrix}\begin{bmatrix}X_C\\Y_C\\Z_C\\1\end{bmatrix} \tag{8}$$

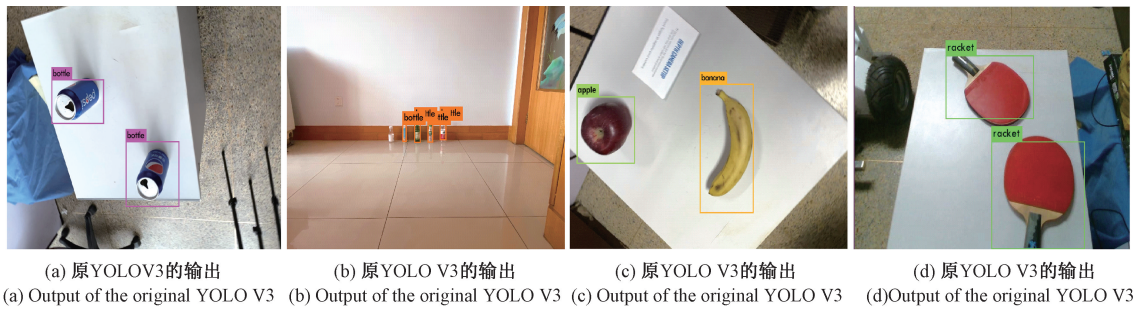
式中:$\boldsymbol{R}_{RC}$ 与 $\boldsymbol{T}_{RC}$ 是由手眼标定计算出的摄像机坐标系到机器人坐标系的旋转矩阵与平移矩阵,由视觉传感器与

机器人的相对位姿所决定。综上所述,通过式(6)~(8),便可计算出像素点对应三维空间点在机器人基座坐标系下的坐标值。

5 实验结果与分析

5.1 改进型 YOLO V3 目标检测实验

改进的 YOLO V3 网络模型训练时,设置动量为 0.9,*batch*=128,初始学习率 *lr*=0.001,学习率采用 policy 的更新方式,衰减系数为 0.1,变化步数为 6 000,总步数为 24 000 步。本文所制作的数据集取材于实验室抓取场景,拍摄了 6 种物体做数据集,包括有瓶子、乒乓球拍、苹果、订书机、鼠标、香蕉等,共计 4 000 张图片。将训练好的模型在配置有 1050ti-GPU 的笔记本上进行测试,并将改进后的 Darknet-43 与 Darknet-53 做对比,部分实验结果对比如图 8 所示。



(a) 原YOLOV3的输出 (a) Output of the original YOLO V3
(b) 原YOLO V3的输出 (b) Output of the original YOLO V3
(c) 原YOLO V3的输出 (c) Output of the original YOLO V3
(d) 原YOLO V3的输出 (d)Output of the original YOLO V3
(e) 改进后YOLOV3的输出 (e)Output of improved YOLO V3
(f) 改进后YOLOV3的输出 (f) Output of improved YOLO V3
(g) 改进后YOLOV3的输出 (g) Output of improved YOLO V3
(h) 改进后YOLO V3的输出 (h)Output of improved YOLO V3

图 8 Darknet-43 与 Darknet-53 检测结果对比

Fig.8 Comparison of detection results between Darknet-43 and Darkenet-53

图 8(a)、(b)、(c)、(d)为原 YOLO V3 模型,基于 Darknet-53 的目标检测结果。图 8(e)、(f)、(g)、(h)为改进后 YOLO V3 模型,基于 Darknet-43 的目标检测结果。两者对比可知,基于改进型的 YOLO V3 目标检测所预测的检测框更精准,且不容易漏检小目标。经过分析可知,虽然 Darknet-43 网络层数减少,但是通过对浅层网络进行扩充,可以使网络捕获更多细节,从而使网络在小物体识别与检测框的回归上预测更加准确。具体结果如表 3 所示。

表 3 目标检测网络性能对比

Table 3 Performance comparison of target detection networks

网络	*P*(精度)/%	R_{IOU}/%	速度/(f·s^{-1})
YOLO V3(Darknet-53)	90.8	0.74	11
Ours(Darknet-43)	91.0	0.75	15

由表 3 可知,本文改进后的 YOLO V3 模型,相较于原 YOLO V3 模型,速度提升 4 f/s 左右,精度 *P* 与交并比略有提升。

5.2 多目标抓取检测实验

本文在训练过程中所用的数据集是 Cornell 抓取数据集,该数据集共有 240 个不同样本的 885 个图像。每个图像都有多个抓取矩形标签,标记为抓取区域预测的正负样本,专门为机器人抓取设计。训练前,将 cornell 抓取数据集的图像随机划分,训练集、验证集、测试集比例为 5∶1∶1。

本文采用矩形抓取度量作为网络精度评估的方法,并与其他抓取检测模型作对比。矩形度量用抓取矩形进行评价,如果同时满足以下两点,则认为抓取矩形可用于抓取物体。首先,预测框的抓取角度与真值标签的角度相差小于 30°;其次,预测的 Jaccard 相似系数大于 25%。Jaccard 相似系数预测抓取区域与真值标签之间的相似性,定义为:

$$J(G_p,G_t)=\frac{(G_p\cap G_t)}{(G_p\cup G_t)} \tag{8}$$

式中:G_p 为预测抓取矩形区域;G_t 为真值的抓取矩形区

域。网络训练的硬件配置是 nvidia TitanX。训练参数如下所示:$batch\_size=128$,$lr=0.0001$,衰减系数为 0.1,变化步数为 20 000,总步数为 100 000。

将本文改进的模型与其他抓取检测模型进行比较,在 Cornell 抓取数据集中与真实物理场景中挑选 10 种不同类型的物体对模型测评,使用的硬件设备是配置有 1050ti-gpu 电脑。在 Cornell 抓取数据集与真实物理场景的测试结果如表 4 所示。

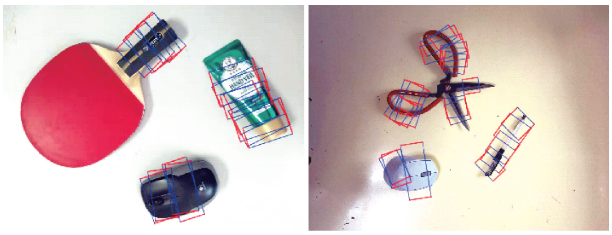
表 4　抓取检测网络对比实验

Table 4　Comparison experiment of grasp detection networks

方法	Cornell data		真实场景	
	时间/s	准确率/%	时间/s	准确率/%
文献[18]方法	0.89	74	0.90	71
文献[6]方法	0.33	91	0.35	87
本文方法	0.16	89	0.17	86

实验结果表明,本文所设计的基于区域提取的抓取检测模型在保证模型精度的同时,能大幅减少模型运算时间,满足机器人抓取的要求。

本文设计的多目标抓取检测算法输出结果如图 9 所示。可以看出在多物体场景下,模型预测出的抓取区域表现良好。



(a) 抓取检测结果
(a) Grasp detection result

(b) 抓取检测结果
(b) Grasp detection result

图 9　多目标抓取检测

Fig.9　Multi-target grasp detection

5.3　机器人最优抓取实验

机器人抓取实验平台如图 10 所示。配有凌华工控机的 UR5 机械臂、REH-64 电动夹爪、深度相机 D435、配有 1050ti-GPU 的笔记本电脑。抓取实验过程中,采集的深度图像大小为 1 238 pixel×1 238 pixel,彩色图像大小为 640 pixel×480 pixel。抓取对象包括 6 种常见的生活物体,有瓶子、乒乓球拍、苹果、订书机、鼠标、香蕉。

实验中,首先利用 D435 采集图像,将采集到的深度图配准到彩色图,逐像素获取深度值;其次,将融合后的图像利用本文所提的 IOU 区域评估算法在图像中生成最优抓取矩形。最后由式(6)~(8)计算出抓取矩形中心点对应在机器人坐标系下空间点的三维坐标值,即机器人的抓取点,并以抓取矩形相对于图像中 X 轴的夹角作为机器人末端执行器的旋转角 θ。本文将物体按照不同的摆放方式分别配置,并在抓取成功率与检测速度上进行对比,如表 5 所示。

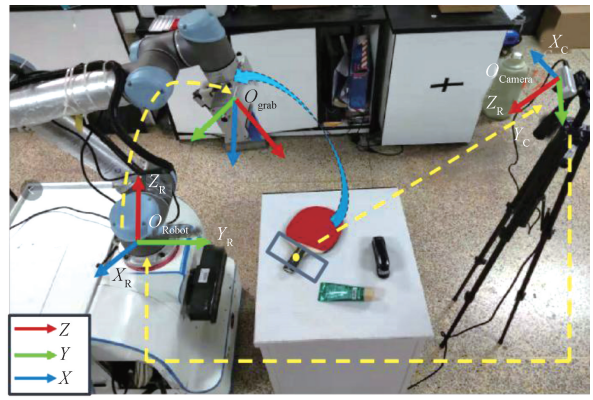


图 10　抓取实验平台

Fig.10　Grasp experiment platform

表 5　最优抓取姿态检测与抓取实验结果

Table 5　Optimal grasp posture detection and grasp experiment results

物体	目标检测次数	检测准确率/%	抓取次数	抓取成功率/%	检测时间/s
瓶子	20/20	100	19/20	95	0.24
香蕉	20/20	100	20/20	100	0.24
球拍	20/20	100	17/20	85	0.25
瓶子+香蕉	19/20	95	18/20	90	0.27
香蕉+球拍	20/20	100	17/20	85	0.26
瓶子+球拍	19/20	95	16/20	80	0.27
瓶子+球拍+香蕉	18/20	90	15/20	75	0.28

由表 5 实验结果可知,本文所提出的机器人最优抓取姿态检测算法抓取成功率较高,可满足抓取任务的要求。原因在于本文所提的 IOU 区域评估算法在计算抓取区域时,融合图像中物体位置的局部特征,缩小了寻找抓取区域的范围;在目标区域中选取 IOU 最大的抓取矩形作为抓取区域的输出,极大地降低了抓取检测出错的概率。

6　结　　论

本文提出了一种融合目标检测与抓取检测的方法,

用来识别物体及其最优抓取区域。该方法在确定机器人抓取姿态时,以待抓取目标边界信息为参考,对图像中的抓取区域进行筛选,提高了目标物体最优抓取区域的检测精度,增强了机器人在抓取任务中对非结构化场景与不规则物体的适应性。虽然双网络架构显得较为冗余,但本文分别对其进行了结构改进和优化,达到精炼网络和提升运行速度的目的。且本文在目标检测模型中,尽可能保存网络浅层的特征,提升目标检测在小目标物体识别上的性能。

综上所述,对于非结构化和带有不规则物体的场景,本文所提的算法可以较好地完成抓取任务。但是本文所提的多目标抓取检测由两个深度学习网络组成,结构上略显繁冗。未来将借鉴多任务目标检测的思想,对双网络架构进行整体优化,以实现更高效率的物体类别与抓取检测。

参考文献

[1] 洪海涛. “新工科”背景下的视觉检测与机器人技术[J]. 大学教育, 2019(7): 74-76.
HONG H T. Vision detection and robot technology in the background of “New Engineering”[J]. University Education, 2019(7): 74-76.

[2] GUO D, KONG T, SUN F, et al. Object discovery and grasp detection with a shared convolutional neural network[C]. International Conference on Robotics and Automation, 2016: 2038-2043.

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

[4] QI C R, SU H, MO K, et al. Pointnet: Deep learning on point sets for 3d classification and segmentation[C]. IEEE Conference On Computer Vision and Pattern Recognition, 2017: 652-660.

[5] 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 and Machine Intelligence, 2017, 39(6): 1137-1149.

[6] CHU F J, XU R, VELA P A. Real-world multi-object, multi-grasp detection[J]. IEEE Robotics and Automation Letters, 2018, 3(4):3355-3362.

[7] ASIF U, TANG J, HARRER S. Densely supervised grasp detector (DSGD)[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019:8085-8093.

[8] 马倩倩,李晓娟,施智平. 轻量级卷积神经网络的机器人抓取检测研究[J]. 计算机工程与应用, 2020, 56(10):141-148.
MA Q Q, LI X J, SHI ZH P. Research on robot grab detection based on lightweight convolutional neural network[J]. Computer Engineering and Applications, 2020, 56(10):141-148.

[9] 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 optimal grabbing method of 3D target based on cascade faster RCNN[J]. Chinese Journal of Sientific Instrument, 2019, 40(4): 229-237.

[10] 张雷涛,胡玉霞,曹秀鸽. 注塑件最小包围矩形算法的研究[J]. 甘肃高师学报, 2005(2):20-22.
ZHANG L T, HU Y X, CAO X G. Research on algorithm of minimum enclosing rectangles for injection molded parts[J]. Journal of Gansu Normal Colleges, 2005(2): 16-18.

[11] 金欢. 基于卷积神经网络的机器人抓取检测研究[D]. 哈尔滨:哈尔滨工业大学, 2019.
JIN H. Research on robot grab detection based on convolutional neural network[D]. Harbin: Harbin Institute of Technology, 2019.

[12] JU M, LUO H, WANG Z, et al. The application of improved YOLO V3 in multi-scale target detection[J]. Applied Sciences, 2019, 9(18): 3775.

[13] KUMRA S, KANAN C. Robotic grasp detection using deep convolutional neural networks[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems, 2017: 769-776.

[14] REDMON J, DIVVAL S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779-788.

[15] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

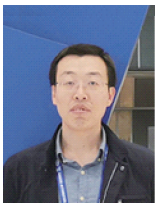
[16] 戴伟聪,金龙旭,李国宁,等. 遥感图像中飞机的改进 YOLOv3 实时检测算法[J]. 光电工程, 2018, 45(12):84-92.
DAI W C, JIN L X, LI G N, et al. Improved algorithm of real-time detection of aircraft in remote sensing

image[J]. Opto-Electronic Engineering, 2018, 45(12): 84-92.

[17] SANDLER M, HOWARD A, ZHU M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]. IEEE conference on computer vision and pattern recognition, 2018: 4510-4520.

[18] REDMON J, ANGELOVA A. Real-time grasp detection using convolutional neural networks[C]. International Conference on Robotics and Automation, 2015: 1316-1322.

作者简介



李秀智,2008 年于北京航天航空大学获得博士学位,现为北京工业大学副教授、硕士生导师,主要研究方向为智能机器人导航、计算机视觉。

E-mail: xiuzhi.lee@163.com

Li Xiuzhi received his Ph. D. from the Beihang University in 2008. Now, he is an associate professor and master student supervisor in Beijing University of Technology. His main research interest includes intelligent robot navigation and computer vision.

李家豪(通信作者), 2015 年于上海理工大学获得学士学位,现为北京工业大学硕士研究生,主要研究方向为智能机器人抓取方法与计算机视觉。

E-mail: 825334787@qq.com

Li Jiahao (Corresponding author) received his B. Sc. degree from University of Shanghai for Science and Technology in 2015. Now, he is an M. Sc. candidate in Beijing University of Technology. His main research interests include intelligent robot grasp method and computer vision.

张祥银(通信作者),2016 年于北京航空航天大学获得博士学位,现为北京工业大学讲师、硕士生导师,主要研究方向为机器人控制与规划。

E-mail: xy_zhang@bjut.edu.cn

Zhang Xiangyin (Corresponding author) received his Ph. D. from the Beihang University in 2016. Now, he is an lecture and master student supervisor in Beijing University of Technology. His main research interest is robot control and planning.