DOI:10.19344/j.cnki.issn1671-5276.2020.02.052

基于图像语义分割的物体位姿估计

王宪伦,张海洲,安立雄
(青岛科技大学 机电工程学院,山东 青岛 266042)

摘　要:为提高机器人在复杂场景中对物体的辨识与定位能力,提出一种基于图像语义分割技术的物体位姿估计方法。将RGBD传感器拍摄的RGB图像放入语义分割网络中,完成对图像的分割与物体分类;将分割出来的目标物体与深度图配准,得到目标物体点云图;将点云图与模型库中的模型运用ICP算法完成对物体的位姿估计。研究结果表明,该方法分割准确率可达82.26%,完成一次位姿估计时间1.35 s。

关键词:机器人;图像识别;图像语义分割;物体识别;位姿估计
中图分类号:TP24　　**文献标志码**:B　　**文章编号**:1671-5276(2020)02-0216-05

Object Pose Estimation Based on Image Semantic Segmentation

WANG Xianlun, ZHANG Haizhou, AN Lixiong
(College of Electromechanical Engineering, Qingdao University of Science and Technology, Qingdao 266042, China)

Abstract: To improve the ability of object recognition and pose estimation in complex scene, this paper presents a method for the object pose estimation based on image semantic segmentation, and puts the RGB image generated by RGBD sensor into semantic segmentation network for finishing the image segmentation and object classification. The point cloud of target object is gotten by registering the object image with depth map. The ICP algorithm is used to estimate the pose of the object with the point cloud and model library. The results indicate that the segmentation accuracy is 82.26% and the time of finishing the estimation at one time is 1.35 s.

Keywords: robot; image recognition; image semantic segmentation; object recognition; pose estimation

0　引言

随着机器人技术、计算机视觉与人工智能技术的发展,复杂场景下的物体识别与姿态估计问题逐渐成为机器视觉领域研究的重点。对于三维物体的识别最早出现在20世纪50年代。开始的研究者主要探索单一背景的物体识别,并将此问题叫做"积木世界"问题。ROBERTS提出了一套针对此问题的识别方案,将多面体分为多个部件进行识别[1]。但是此模型对于真实世界做了过分的简化,因此对于曲面较多的物体难以适用。BOLLES开发出一种面向三维工件的识别方案,根据工件的CAD模型,在工件的顶点、边、面、体之间建立拓扑关系,通过图匹配的方法实现工件的识别[2]。CHEN开发了一种以不同视角的图像建立物体立体模型的方法,在图像匹配的过程中使用方向、轮廓等信息,进行识别定位[3]。近年来,针对此问题的解决方案大多是以点云数据为基础。首先是离线采集点云数据,通过分割算法进行点云聚类[4],计算点云的特征描述子,获得物体的最小包围盒[5],实现对目标物体的识别与定位。国内外许多科研机构在这方面都取得了较好的成果,例如:Willow Garage公司PR2机器人、中国科技大学的"可佳"机器人等。但是随着深度学习的普及,尤其是卷积神经网络[6]在图像分类取得突破之后,深度学习也开始应用于图像语义分割,所以对比之前的解决方案使用的3D点云数据,现在只需要通过对2D图像进行语义分割便可以实现对图像分割与分类,大大加速了物体的识别过程。

本研究设计的基于图像语义分割技术的物体姿态估计算法,它的最终目标是实现复杂场景下目标物体的识别与定位,依靠Kinect视觉传感器拍摄场景RGB图像与深度图,利用全卷神经网络[7]实现RGB图像语义分割,将分割后的目标图像与深度图配准得到点云图,最后利用迭代最近点[8]算法将目标点云图与模型库中已有的目标模型进行配准,最终实现物体的识别与位姿估计。

1　物体识别与位姿估计算法

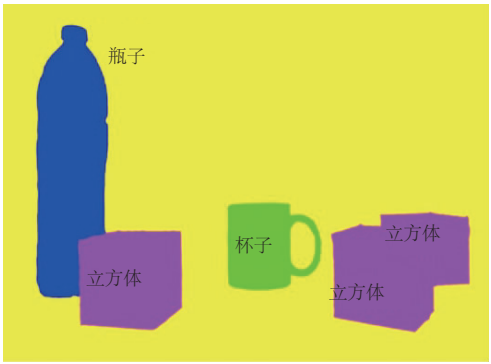
物体识别与位姿估计算法主要包括基于图像语义分割技术的物体识别与位姿估计两部分。

1.1　物体识别

图像语义分割属于图像理解的范畴。语义分割要求将图中的每一个像素点分割并标注为某个物体类别(图1)。

第一作者简介:王宪伦(1978—),男,山东济宁人,副教授,博士,研究方向为机器人技术。

(a) 原图



(b) 分割图

图 1　场景原图与分割图

本文采用的语义分割网络是 FCN(全卷积神经网络),通过改进主流的深度卷积神经网络在数据集上训练,以 RGB 图像为输入,然后输出图像的像素语义分割图,实现像素级别端对端的分割。在语义分割网络的设计中输出的语义分割图像需要与输入的图像有相同的尺寸,但是在做图像分类的网络模型中使用了多个卷积层与池化层,在经过多次的卷积与池化操作之后会让原始输入的图像尺寸越来越小,所以为了得到与输入图像尺寸一致的输出,必须进行反卷积[9]操作。

整个语义分割网络的设计架构以分类效果较好的 ResNet[10](深度残差网络)为基础。ResNet 的产生是为了解决随着网络深度不断增加而出现梯度消失或者梯度爆炸的问题。形式上,输入的数据表示为 x,将期望的底层映射表示为 $H(x)$,将堆叠的非线性层拟合后的映射为 $F(x)$。传统的卷积网络直接以 $F(x)$ 作为下一层的输入,但是随着网络深度不断增加会使梯度变得越来越小,丢失更多的原始信息。所以 ResNet 以 $H(x)=F(x)+x$ 作为下一层的输入,使网络深度可以达到较深的层数,从而使训练模型对于输出与输入的微小波动更加敏感,最终的分类效果更好。

语义分割网络为了获得更好的分割效果,选择以 ResNet 为基础,去掉原网络的全局池化层,因为全局池化层会丢失图像的空间信息,然后将全连接层替换为核尺寸为 1 的卷积层,最后一层接入一个卷积转置层,使网络的输出等于输入尺寸。输出的图像中可以将不同类别的物体以不同颜色表示出来,并且将实现不同物体之间的分割。整体网络设计架构示意图见图 2。

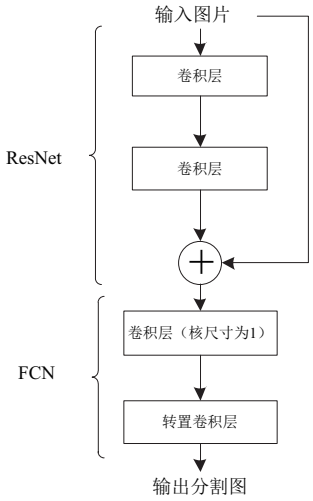


图 2　语义分割网络设计图

1.2　位姿估计

位姿估计部分主要分为两部分,第一部分是目标点云的获得,第二部分是将目标点云模型库中目标的点云配准,求取物体位姿。

目标点云的获得主要依赖于 RGBD 传感器的成像模型。在不考虑镜头畸变的情况下,摄像机的成像模型为小孔成像,如图 3 所示。空间中的点坐标为 $P=[x,y,z]$ 与其在图像成像平面上坐标 $p=[u,v]$ 满足如下关系:$u=(x\cdot f)/z$,$v=(y\cdot f)/z$,其中 f 为 Kinect 视觉传感器的焦距。深度图中的每一个像素的值 $d(u,v)$ 保存了场景中的点到镜头中心的距离。通过标定[11],可以将彩色图与深度图的各个像素一一对应起来,实现彩色图像中像素点到空间中三维坐标的映射:

$$\begin{aligned} z&=d(u,v)\\ x&=\frac{(u-c_x)z}{f}\\ y&=\frac{(v-c_y)z}{f} \end{aligned} \tag{1}$$

其中 c_x、c_y 为摄像机光心坐标。将分割出来的目标物体的彩色图与深度图对齐之后便可以得到其点云图。

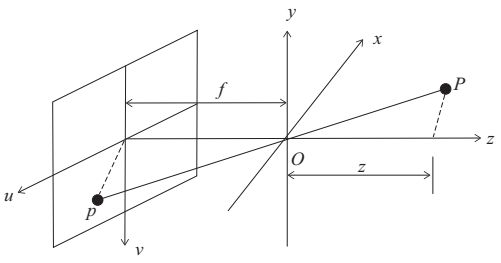


图 3　摄像机成像模型

点云配准过程一般分为两个阶段:第一阶段为粗配准,使得目标点云与模型库中模型之间位姿差距最大可能地减小,达到大致的重合状态;第二阶段为精配准,其目的是通过精确的配准算法使得目标点云与模型库中点云达

到最佳的重合状态。

设 $P=\{P_i\}_{i=1}^{N}$,$Q=\{Q_i\}_{i=1}^{N}$为两个点云集,P 是上文中获得的目标点云,Q 是模型库中对应的模型点云。粗对准算法步骤如下:

首先计算点云 P 和 Q 的协方差矩阵$\boldsymbol{C}_p$和$\boldsymbol{C}_q$:

$$\boldsymbol{C}_p = \frac{1}{N}\sum_{i=0}^{N}(P_i-\bar{P})(P_i-\bar{P})^{\mathrm{T}} \tag{2}$$

$$\boldsymbol{C}_q = \frac{1}{N}\sum_{i=0}^{N}(Q_i-\bar{Q})(Q_i-\bar{Q})^{\mathrm{T}} \tag{3}$$

其中:$\bar{P}$是点云 P 的中心,$\bar{Q}$是模型点云 Q 的中心,$\bar{P}=\frac{1}{N}\sum_{i=1}^{N}P_i$,$\bar{Q}=\frac{1}{N}\sum_{i=1}^{N}Q_i$,然后通过对协方差矩阵进行 SVD 分解可以得到如下形式:

$$\boldsymbol{C}_p=\boldsymbol{U}_p D_p V_p^{\mathrm{T}} \tag{4}$$

$$\boldsymbol{C}_q=\boldsymbol{U}_q D_q V_q^{\mathrm{T}} \tag{5}$$

其中:$\boldsymbol{U}_p$是点云 P 的特征向量,$\boldsymbol{U}_q$是点云 Q 的特征向量。

则点云 P 和 Q 之间的旋转矩阵 $\boldsymbol{R}$ 和平移向量 $\boldsymbol{T}$ 为:

$$\boldsymbol{R}=\boldsymbol{U}_p\boldsymbol{U}_q^{-1} \tag{6}$$

$$\boldsymbol{T}=\bar{Q}-\boldsymbol{R}\bar{P} \tag{7}$$

$\boldsymbol{R}$ 和 $\boldsymbol{T}$ 便是点云粗匹配的结果。

在精确匹配时,普遍运用 ICP(iterative closest point)算法,该方法通过逐步迭代的方法寻找两个点云集中的匹配对应点,并计算两个点云集之间的刚体变换参数,直到达到最大迭代次数或者达到给定的收敛精度,最终求得两个点云集之间的刚体变换参数,即目标物体的位姿。设待精配准的两个点云集为 $P=\{p_i\}_{i=1}^{N_p}$,$X=\{x_i\}_{i=1}^{N_x}$为待配准的两片点云。ICP 算法首先对点云集 P 中的每个点p_i,寻找其在点云集 X 中的距离最近点q_i作为对应点。设点云集 P 中每个点寻找到的最近点集合为 $Q=\{q_i\}_{i=1}^{N_x}$。

算法建立如下损失函数:

$$E(q) = \frac{1}{N_x}\sum_{i=0}^{N_x}\| q_i - \boldsymbol{R}(\boldsymbol{q})\,p_i - \boldsymbol{T}(\boldsymbol{q})\|^2 \tag{8}$$

其中:$\boldsymbol{R}(\boldsymbol{q})$表示旋转矩阵;$\boldsymbol{T}(\boldsymbol{q})$为平移向量。求解式(8),使其值最小。

初始迭代时,令目标点云的初始位置为P_0,刚体变换向量为 $\boldsymbol{m}$,迭代次数 $k=0$。迭代执行步骤如下:

1) 寻找对应点集:计算点云集 P 的最近点集合为 Q,$Q=C(P,X)$,其中 C 为搜索最近点操纵;

2) 计算配准参数:按照粗对准算法计算两个点云集之间的 $\boldsymbol{R}$、$\boldsymbol{T}$;

3) 将配准参数作用到P_0得到新的位置:$P_{k+1}=q_k(P_0)$;

4) 若相邻两次迭代求得的误差d_k小于给定的阈值 t,即$d_k-d_{k+1}<t$,则迭代终止;否则,$k=k+1$,转步骤 1)。

最终通过目标点云与模型库中模型点云的精配准得到目标点云的旋转矩阵 $\boldsymbol{R}$ 与平移向量 $\boldsymbol{T}$,即目标物体的位姿。

为了加速精确配准中两块点云的对应点对匹配,本文采用 Kd-tree[12] 近邻搜索算法进行加速。Kd-tree 是对于高维数据中的快速最近临查找算法,其本质是一种二分查找树,被 ZHANG 等[13] 第一次运用到 ICP 算法中。对于点云数据来说,Kd-tree 中存储的是三维数据,Kd-tree 的建立过程就是对三维空间的一个划分过程,实际运用中,建立过程如下:

1) 计算三个维度中各个维度上的数据方差,将最大方差的维度定义为划分轴。

2) 在划分轴上计算中值作为临界值,将全部的三维空间分为两份。同时创建一个结点,用于存储划分的维度与划分值。

3) 将三维数据在划分轴的维度上与临界值进行对比,小于临界值的数据归为左子树,大于临界值的数据归为右子树。

4) 对左右两棵子树循环进行步骤 1)到步骤 3),直到全部的子集合都不能再划分,并将该数据保存为叶子节点。

Kd-tree 为点云数据建立了一个快速查找的拓扑结构,加速了点云的匹配点查找过程。

2　实验结果与分析

整个实验主要分为两步,第一步是基于图像语义分割技术的目标物体分割,第二步是物体位姿的估计。因为第一步的图像分割质量对位姿估计结果影响较大,所以首先对第一步的分割质量做了明确的评价。

实验环境采用 MXNet[14] 平台搭建网络,MXNet 是 Amazon 旗下的深度学习框架。实验硬件配置如下:CPU 为 Intel(R)i5-8600K,GPU 为 GeForce GTX1070Ti,操作系统为 Window10。

本次实验采用 PASCAL challenge[15] 目标检测评价体系中提出的评价标准 IoU(intersection over union)。此标准是计算预测值与真值的交集和预测值与真值并集的比值。交并比的定义如下:

$$IoU=\frac{DR\cap GT}{DR\cup GT} \tag{9}$$

其中:GT 是真实的目标物体分割图的像素区域;DR 是预测的目标物体分割图的像素区域,如图 4、图 5 所示。

图 4　*DR* 区

本文采用的数据集是自建数据集,其格式按照 PASCAL VOC2012 格式建立,其中涉及 6 种类别,分别为“motor”、“ruler”、“box”、“mouse”、“cola”和“ball”。原始

图 5　*GT* 区

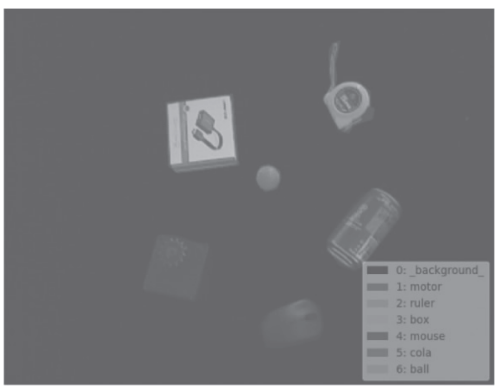
数据集包括 1 000 张,其中 400 张用于训练,600 张用于模型测试。

在对目标网络进行训练时,模型训练的超参数设置如下:基础学习率为 0.001,学习率的变化率为 0.1,最大迭代次数为 20 000,步长为 3 000。

在 Windows 下目标检测速度为 0.978 秒/张。算法结果在数据集上的 *IoU* 值最终为 82.26%。目标检测实验结果如图 6 所示。

从图 6 可以看出,原来场景中 5 种物品全部被分割出来,并且以不同的颜色代表不同的物体,从而实现了物体识别的功能。把可乐作为目标物,将可乐提取出来,与深度图配准,得到可乐的点云图(图 7)。

(a)实际场景图



(b)场景标签图

(c)场景分割图

图 6　目标检测实验结果

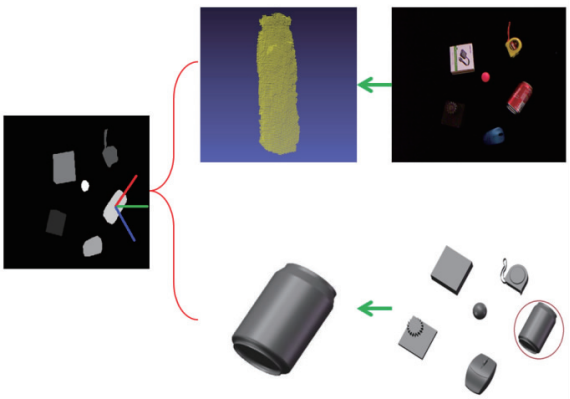


图 7　目标点云图

将获得的目标点云图与模型库中可乐模型运用基于 ICP 算法的点云匹配方法进行位姿计算,可以得到目标物体准确位姿。

为了评价最终位姿估计的精度,引入机械臂作为结果量化工具。具体操作方法是:通过算法计算出一组位姿数据,并且转换到机器人坐标系下,同时人工操作机械臂到达实际物体的位姿位置,记录下一组位姿数据(表 1)。数据格式为 x、y、z、α、β、γ,距离单位为 mm,角度单位为(°)。

实验结果分析,算法平均误差为 3.96 mm、1.59 mm、2.64 mm、2.35°、1.7°、4.15°。通过以上数据可以看出位姿估计基本可以满足机器人对于可乐、水杯类似物体的抓取,具有一定的应用性。

表 1　8 组算法与实际测量的误差对比

x 误差	y 误差	z 误差	α 误差	β 误差	γ 误差
−4.2	1.2	3.2	2.3	1.2	3.5
5.3	2.0	2.7	1.9	2.0	4.3
3.2	1.3	1.9	−2.1	1.1	−4.5
4.1	−1.4	−2.9	2.4	2.1	−3.9
−3.9	1.7	2.3	−2.7	−1.9	4.2
2.7	2.0	−3.2	2.4	1.4	3.7
3.2	1.2	2.5	−1.9	−1.2	4.2
5.1	1.9	−2.4	3.1	2.7	−4.9

3　结语

本研究提出的基于语义分割的物体位姿估计算法得到了实验验证,实验结果表明此方法可以实现复杂场景下目标物体的识别以及位姿估计,可以达到实际应用的要求,并为以后的研究提供了重要参考依据。

在下一阶段,本研究将会把此算法应用到机械臂抓取的实验中,将此算法与机械臂联合使用完成目标物体识别抓取的工作。由于此算法并未达到100%的物体识别率,所以在高精度的应用场景中无法使用,因此在今后的研究中,可能需要进一步改进基于语义分割的物体识别方法,减少实验误差,以实现高精度作业。

参考文献:

[1] ROBERTS L G. Machine perception of three-dimensional solids [D]. Cambridge: Massachusetts Institute of Technology, 1963.

[2] BOLLES R C, HORAUD P. 3DPO: A three-dimensional part orientation system [M]. Springer, Boston: Three-Dimensional Machine Vision, 1987: 399-450.

[3] CHEN C H, KAK A C. A robot vision system for recognizing 3d objects in low-order polynomial time[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1989, 19(6): 1535-1563.

[4] SHI B,LIANG J,LIU Q. Adaptive simplification of point cloud using K-means clustering[J]. Computer-Aided Design,2011,43(8):910-922.

[5] R. Girshick, J. Donahue, T. Darrell, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2014: 580-587.

[6] LECUN Y, BOTTOUN L, BENGGIO Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11):2278-2324.

[7] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation [C]. USA: Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, IEEE, 2015: 3431-3440.

[8] BESL P J, MCKAY N D. Method for registration of 3-D shapes [C]. Sensor fusion IV: control paradigms and data structures. International Society for Optics and Photonics, 1992, 1611: 586-606.

[9] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[M]. Computer vision-ECCV 2014. Springer International Publishing, 2014:818-833.

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

[11] 郭连朋,陈向宁,刘彬.Kinect传感器的彩色和深度相机标定[J]. 中国图像图形学报, 2014,19(11):1584-1590.

[12] BENTIEY J. L. Multidimensional binary search trees used for associative searching[C].Com. of ACM. 1975, l8:509-517.

[13] ZHANG Z Y. Iterative point matching for registration of free-from curves and surfaces [J]. International Journal of Computer Vision, 1994,13(2):119-152.

[14] CHEN T, LI M, LI Y, et al. MXNet: a flexible and efficient machine learning library for heterogeneous distributed systems [EB/OL]. https://arxiv.org/abs/1512.01274.

[15] EVERINGHAM M, VAN Gool L, WILLIAMS C K I, et al. The pascal visual object classes (voc) challenge[J]. International Journal of Computer Vision, 2010, 88(2):303-338.

收稿日期:2019-01-21