第50卷第1期 中国科学技术大学学报 Vol. 50, No. 1
2020年1月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Jan. 2020

文章编号:0253-2778(2020)01-0001-10

基于RGB图像的二阶段机器人抓取位置检测方法

熊军林,赵 铎

(中国科学技术大学自动化系,安徽合肥 230027)

摘要:随着机械臂在越来越多的场合扮演着重要的角色,准确的抓取位置检测是整个机械臂系统顺利完成任务的关键,为此提出一种以整个图片为输入直接输出结果的端到端实时检测方案.物体的抓取点位置会影响到该物体的抓取角度,基于此给出了一种两阶段预测方案将这两个要素分开预测.首先,建立一个卷积神经网络预测物体的抓取点位置;然后,以抓取点位置为中心采集原图像中的一个方形区域.针对这一区域利用Canny算法以及Hough变换进行边缘提取和直线检测,并提出一种主方向提取算法,分析得到直线,进而确定物体的角度和抓取时平行夹持器张开的间距.抓取位置检测算法给出了基于RGB图像预测的较好准确率,神经网络与传统方法的结合使用也为以后的研究提供了参考.

关键词:抓取位置检测;卷积神经网络;边缘提取;直线提取;主方向提取

中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2020.01.001

引用格式:熊军林,赵铎. 基于RGB图像的二阶段机器人抓取位置检测方法[J]. 中国科学技术大学学报,2020,50(1):1-10.

XIONG Junlin, ZHAO Duo. Two-stage grasping detection for robots based on RGB images[J]. Journal of University of Science and Technology of China, 2020,50(1):1-10.

Two-stage grasping detection for robots based on RGB images

XIONG Junlin, ZHAO Duo

(*Department of Automation, University of Science and Technology of China, Hefei* 230027, *China*)

Abstract: Recently, robots have played big roles in more and more cases. An accurate grasp detection is a key component of a robot working process. An end-to-end method for robotic grasp detection in an RGB image containing objects is proposed in such a case, which takes the whole picture as input and gives the prediction result directly without using traditional sliding windows or region extraction. Obviously, different grasp points lead to different grasp orientations. The grasp detection method takes two steps. First, a convolutional neural network is trained to predict the positions of grasp points. Next, a square area with the preceding grasp point as the center is taken from the image, where the edges are extracted using the Canny edge detection and the lines are detected using Hough Transform. A principal-directiondetection algorithm is proposed to analyze these lines and detect grasp orientations and the distance between two parallel fingers. The method gives a better grasp detection and has an influence on computer vision using both deep learning and traditional algorithms.

Key words: robot grasping detection; convolutional neural network; edge detection; line detection; principal-direction detection

收稿日期:2018-11-15;**修回日期**:2019-05-27

基金项目:国家自然科学基金(61773357)资助.

作者简介:熊军林(通讯作者),男,1977年出生,博士/教授,研究方向:自动控制理论. E-mail:xiong77@ustc.edu.cn

0 引言

伴随着信息技术与控制技术的发展，机械臂在工业、医疗以及家庭场景中有了越来越广泛的应用. 机械臂的工作方式通常由机械臂末端操作装置的种类决定. 焊接头、钻头等可以用于加工、组装零件；吸盘、平行夹持器用于移动各种外形的物体；特制刀具、微型摄像头等辅助医疗手术. 抓取操作作为机械臂末端操作的一种，是很多对象操作类任务的基础. 家庭场景中常见物体的移动操作、工业中特定零件的抓取及运输操作，都需要精确且鲁棒性强的抓取操作. 此外，抓取操作的灵活程度对机械臂工作内容的多样性也有直接影响，抓取操作的泛化性越强，机械臂的工作能力越灵活，

抓取操作是否成功的关键在于是否得到一个准取的抓取位置. 物体的抓取位置由待抓取物体的几何以及物理特性决定，这就需要我们引入视觉信息并关注抓取位置检测算法的泛化性能. 对于物体抓取位置的检测主要关注 3 个方面的信息，即物体的抓取位置、抓取角度以及抓取时平行夹持器的张开距离. 以人类经验而言，物体的抓取位置一般位于物体质心附近，抓取角度垂直于物体较长的平行边，而理解这一点对于机器人来说并不容易.

物体抓取位置检测的研究始于 20 世纪 80 年代，早期的研究大都是针对抓取点的检测，这并不能提供一个确定的抓取方式描述. 直到 2011 年，Jiang 等[1]给出了一种抓取位置的矩形框描述，如图 1 所示，这个矩形需要一个 5 维向量 (x,y,h,w,θ) 来表示. 其中，(x,y) 表示抓取点在图像空间中的坐标；h 表示平行夹持器张开的距离；w 表示夹持器的宽度，这个量在实际机械臂系统中是一个常量；θ 表示物体抓取角度. 以此为依据，将物体抓取位置检测问题转化为在图像空间寻找这样一个矩形的问题.

2002 年，Paiter[2]使用 k-means 算法针对简单的物体进行抓取点位置检测. Morelas 等[3]使用启发式算法对某些特定形状的物体做抓取检测. 这些方法只能针对具有形状特点的物体才能有很好的效果，泛化性能很差.

深度学习的发展为泛化性问题的解决提供了新的途径. 第一代卷积神经网络 LeNet[4]在 1998 年诞生，这个网络包含了 5 个卷积层并且引入了池化层的概念. 随着计算能力的提升，2012 年，具有划时代

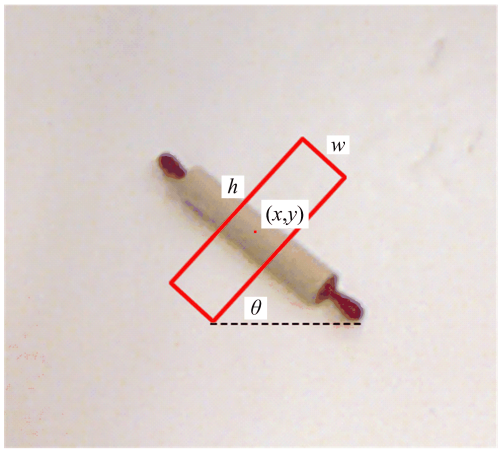


图 1 抓取位置矩形框描述法示意图

Fig. 1 Rectangle for grasping

意义的网络 AlexNet[5]诞生，AlexNet 是一个加深版的 LeNet，包含了 5 个卷积层和 3 个全连接层；引入了 ReLU 激活函数以及采用 dropout 方法处理过拟合. 之后的几年之间 VGGNet[6]，ResNet[7]，GoogleNet[8-10]等各有特点的神经网络模型相继出现，卷积神经网络也在诸如图像分类、目标检测等领域发挥了重要作用.

2014 年，Lenz 等[11]使用一个卷积神经网络进行抓取位置预测. 这种检测方式采用滑动窗口的方式进行矩形框的筛选，这样的筛选过程使得实时性较差，采用的网络结构也比较复杂. 2015 年，Redmon 等[12]利用 AlexNet 网络结构进行了抓取位置的预测，以一整张图片作为输入，输出抓取位置的矩形描述向量 $(x,y,h,w,\sin2\theta,\cos2\theta)$，这个检测方法解决了实时性较差的问题，同时降低了训练时的开销，但将抓取点位置与角度同时预测，忽略了两者之间的关联. 2016 年，Wang 等[13]利用了一个改进的级联网络提高了抓取位置预测的准确性，但同样存在网络结构较为复杂的问题.

考虑到抓取点位置与角度的关联并关注到实时性，本文主要做了如下的工作：

(Ⅰ)设计了一种基于 AlexNet 的神经网络结构，用于抓取点位置的检测. 本文重新设计了全连接层，并将批量归一化算法[14]应用到网络中，提高了网络的拟合能力和训练速度，抑制了网络过拟合.

(Ⅱ)通过抓取点位置与抓取角度的关联，提出了一种两阶段抓取位置检测算法. 以抓取点为中心，在原图像上裁取方形区. 分析此区域进行抓取角度的提取，提高了抓取角度预测的准确性.

(Ⅲ)利用 Canny[15]边缘提取和 Hough[16]变换

提取物体的边缘直线信息. 以这些直线信息为基础，本文设计了物体主方向提取算法预测抓取角度 θ 以及夹持器张开距离 h . 相较于利用神经网络的检测方法，抓取角度预测的成功率提高了 30%.

1　物体抓取点位置检测

物体的抓取点受物体的几何特性以及物理特性的影响，很难使用传统的视觉算法进行检测. 从一张图片中得到一个点的坐标，这本质上是一个非线性回归问题，而卷积神经网络恰好是一个强大的非线性拟合工具.

本文采用的网络结构是以 AlexNet 为基础，重新设计了网络的全连接层，并将批量归一化算法应用到每一层卷积层的输入，提高了网络的训练速度以及抗干扰性能.

1.1 基本网络结构

AlexNet 包含主要 5 个卷积层，3 个全连接层，卷积层之间采用 max-pooling 层连接. 卷积层的主要功能是特征提取，每层的卷积核通过训练不断地进行调整，训练结束之后的卷积核就是特征提取器. 卷积层之后的池化层是对图像特征的一种采样降维. 池化操作是基于图像特征的区域相关性，在不影响特征包含信息的情况下，降低特征的维数，以此降低整个网络系统的计算量. 卷积操作是一种图像空间中的线性操作，网络非线性回归能力由激活函数提供. AlexNet 卷积层采用的激活函数是 ReLU 函数，形式为

$$r(x)=\begin{cases}x, x>0\\0, x\leqslant 0\end{cases} \tag{1}$$

ReLU 函数求导简单，大于零部分的导数值恒为 1，避免了神经网络中梯度消失的问题. 全连接层激活函数采用的是 sigmoid 函数，形式为

$$s(x)=\frac{1}{1+e^{-x}} \tag{2}$$

$s(x)$ 函数处处可导，$\dot{s}(x)=s(x)(1-s(x))$，导数的形式方便计算机的计算，为网络提供了更强的非线性拟合能力.

本文重新设计了网络的全连接层，如图 2 所示. 预测网络包含了 3 个全连接层，连接节点数分别为 1024，256，64. 最后的输出层只包含两个节点，即 (x,y) . 由于本文的输出维数远小于卷积层给出的特征维数，需要全连接层逐渐缩小特征的尺度，以此进行特征分析和整合. 全连接层的节点数选为 2 的整数幂，这样可以加快网络的运算速度.

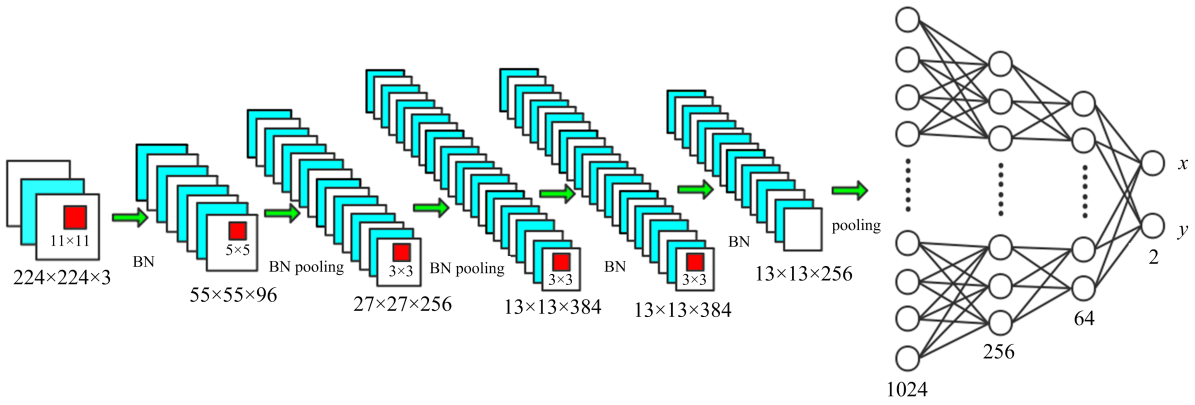


图 2　抓取点位置预测网络示意图

Fig. 2　Network for grasping points detection

由于本文的网络以抓取点位置 (x,y) 为输出，网络的预测问题本质是一个回归问题，因此我们选用均方误差作为网络的输出误差，单个样本的误差表示为

$$E=(x-\tilde{x})^2+(y-\tilde{y})^2 \tag{3}$$

式中，(x,y) 为样本的标签，$(\tilde{x},\tilde{y})$ 表示网络的输出.

1.2　数据集多标签的处理

本文训练网络的数据集源于康奈尔大学的抓取数据集. 生活中每个物体不只有一种抓取方式，抓取数据集为每一个抓取物体提供了多个正确的抓取位置. 单个物体，对应多个不同的抓取点位置 (x,y)，这是一个多标签回归问题，无法利用网络拟合这样的关系. 若随机选择其中的一个作为标签，又会造成数据集包含信息的丢失，使得最终训练的网络预测

能力不足.

本文给出了一种数据集增广方法解决这个问题.对于数据集中每一个包含多标签的样本,根据它的标签序号做图片平移操作,得到新的图片,相应地,对该标签做同样的平移处理,得到一组新的样本标签对.假如一个样本包含 N 个标签,最终将这个样本转化为 N 个标签样本对.这样既保留了数据集中的标签信息,又增广了数据集,增强了网络的拟合能力,抑制了过拟合问题.

1.3 批量归一化

本文将批量归一化(batch normalization)[14]算法应用在AlexNet网络中,提升网络训练速度的同时,也可以增强网络的拟合能力并抑制过拟合.

假设一个参与训练的批次包含 Q 个样本,表示为 $M=\{m_1,m_2,\cdots,m_Q\}$,批量归一化的核心思想就是将这些输入归一化为一组分布符合均值为0、方差为1的数据,具体操作如下:

$$\mu_M=\frac{1}{n}\sum_{i=1}^{n}m_i \tag{4}$$

$$\sigma_M^2=\frac{1}{n}\sum_{i=1}^{n}(m_i-\mu_M)^2 \tag{5}$$

$$\hat{m}_i=\frac{m_i-\mu_M}{\sqrt{\sigma_M^2+\varepsilon}} \tag{6}$$

式(4)中,μ_M 代表 M 中输入的均值;式(5)中,σ_M^2 代表 M 中输入的方差;式(6)进行了归一化处理,得到了 $\hat{M}=\{\hat{m}_1,\hat{m}_2,\cdots,\hat{m}_Q\}$,$\hat{M}$ 服从均值为0,方差为1的分布.ε 表示一个大于零的很小的量,用以避免出现分母为0的情况.

取参与训练的两个连续的 M_i 和 M_j,如果 M_i 和 M_j 有很大的不同,训练会导致整个网络的参数进行大范围调整,延长训练时间.如果 M_i 和 M_j 完全相同,则不需要进行训练.为了减少训练时间,我们应该使得 M_i 和 M_j 尽可能相同,通过批量归一化得到均值和方差相同的 $\hat{M}_i$ 和 $\hat{M}_j$,加快了网络的训练速度.这样的归一化操作的不足是,将会使得卷积层提取的特征失去原本的分布,从而导致特征信息的丢失,因此在进行了批量归一化之后,再进行如下的线性变换:

$$\hat{o}_i=\gamma\hat{m}_i+\beta \tag{7}$$

式中,γ 和 β 是需要训练的参数,用以恢复特征原本的分布,$\hat{O}=\{\hat{o}_1,\hat{o}_2,\cdots,\hat{o}_Q\}$ 即为卷积层的输出.

本文采用上述网络结构做抓取点位置的检测,采用整张图片作为输入,确保了系统的实时性;提出了一种处理多标签的方式,扩展了数据集,充分利用了数据集信息;并将批量归一化方法引入了AlexNet,提升了网络的训练速度,抑制了过拟合.

2 物体抓取角度及夹持器距离的检测

抓取角度的选取与抓取点的位置是相关的,物体的抓取角度受抓取点附近物体几何特性的影响.同样的一个杯子,抓取点选取杯子的把手或杯身,相应的抓取角度会有很大的不同.于是本文提供了一个基于已选取抓取点的抓取角度提取方案,以提高抓取角度检测准确性.

以改进的AlexNet预测出的抓取点位置结果 (x,y) 为中心,提取一个正方形区域.这个区域包含了抓取点附近的图像信息,以此区域为抓取角度检测算法的输入.抓取角度的预测可以看作分类问题,也可以作为回归问题来求解.若看作回归问题,即直接预测抓取角度;若看作分类问题,需要将[0°,180°)分割为若干个子区间,每个子区间看作一个子类,以子区间的中心角度为该类的代表角度,将图片正确分类即可进行角度预测.

卷积神经网络对处理分类和回归问题都比较擅长,但对抓取角度的检测表现不佳.我们尝试了以AlexNet为主体的网络分别以回归和分类的方式解决角度预测问题.我们将[0°,180°)区间分为18个子类,得到分类的准确率低于70%,角度回归的误差大于20°.这显然无法满足抓取角度检测的需求.由于卷积神经网络的卷积核是平行移动的,这样的特征提取器对角度特征的提取能力有限,传统的视觉方法表现得更好.本文利用传统的视觉方法分析带抓取物体的边缘特征,从而分析抓取角度.

2.1 Canny边缘检测

本文利用Canny边缘检测算法提取物体的边缘信息.首先需要利用高斯平滑滤波处理图像,防止噪点对于边缘提取的影响.利用公式

$$\left.\begin{aligned}H_{ij}&=\frac{1}{2\pi\sigma^2}\exp(K_{ij})\\K_{ij}&=-\frac{(i-(g+1))^2+(j-(g+1))^2}{2\sigma^2}\end{aligned}\right\} \tag{8}$$

生成一个大小为 $(2g+1)\times(2g+1)$ 的高斯核,其中 H_{ij} 即为卷积核第 i 行第 j 列的元素.用这个卷积核执行卷积操作进行高斯滤波,滤去原图片中的噪声.利用Sobel算子来检测去噪后图像的水平、

垂直边缘，得到图像在水平与垂直方向上的一阶导数 G_x 和 G_y，确定像素的梯度如下：

$$\left.\begin{aligned} G &= \sqrt{G_x^2 + G_y^2} \\ \varphi &= \arctan(G_y / G_x) \end{aligned}\right\} \tag{9}$$

式中，G 代表梯度的模，φ 代表梯度的方向. 对整张梯度图进行非极大值抑制，取像素点正负梯度方向上的两个点，比较梯度值. 如果这个点的梯度不是最大的，那么这个点就会被抑制. 非极大值抑制有助于图像的边缘更加准确.

上述操作只会提取局部极大的梯度值，会使得梯度值很小的像素点得以保留. 这些像素点依旧可能是伪边缘点，影响边缘提取的结果. 针对这些伪边缘点，首先，采用非极大值抑制的方式，考虑每一个边缘候选点梯度正负方向的两个候选点，如果候选的边缘点梯度的模大于正负方向的两个候选点梯度的模，这个候选点的梯度将会被保留；相反，这个候选点的梯度将会被抑制. 其次，由于上述操作依旧无法去除梯度值较小的伪梯度点，因此要引入双阈值检测的方式对梯度较小的伪边缘点进行抑制. 设定两个阈值，高阈值与底阈值，当点的梯度值大于高阈值的时候，直接标记该点为强边缘点；当点的梯度值在高阈值与低阈值之间时，该点标记为弱边缘点，即待定状态；当点的梯度小于低阈值的时候，则该点将被抑制. 最后，对待定的弱边缘点，采用孤立边缘点抑制的策略来处理弱边缘点，处理的策略是在已知的弱边缘点的周围 8 个位置寻找强边缘点. 如果存在强边缘点，则将这个边缘点标记为强边缘点；如果不存在，这个弱边缘点就被抑制. 经过这样的操作，可以有效地去除伪边缘点，避免其对后续的算法进行干扰.

2.2　Hough 变换

Hough 变换是一种直线提取方法，对于 Canny 算法得到的图像边缘，利用 Hough 变换我们就能得到关于原图像边缘的解析描述，进而分析物体的抓取角度.

在一个平面直角坐标系中，通过点 (x_0, y_0) 的直线可以用描述为

$$y_0 = kx_0 + b \tag{10}$$

式中，k 表示的直线的斜率，b 表示的直线的截距. 对式(10)稍作变换可以得到

$$b = -kx_0 + y_0 \tag{11}$$

式(11)可以看作一条 k-b 空间的斜率为 $-x_0$，截距为 y_0 的直线. 至此，通过式(11)，我们将原空间中的一个点映射为新空间的一条直线. 如果我们另外取一个新的原空间中的点 (x_1, y_1)，同样可以将其映射为 k-b 空间的一条直线，即

$$b = -kx_1 + y_1 \tag{12}$$

结合式(11)与(12)可以得到一个 k-b 空间的交点 (k_0, b_0)，k_0 即为原空间点 (x_0, y_0) 与点 (x_1, y_1) 组成直线的斜率，b_0 为截距. 这样就将原空间中的一条直线转化为 k-b 空间中的一个点. 对于 k-b 空间中的点计数保存在 Hough 矩阵中，计数的值代表原图像中共线的点的个数，计数值越大，代表原图像中该直线存在可能性越高，矩阵的索引值对应直线的斜率与截距. 由于 Hough 变换得到的直线段不一定是决定物体主方向的关键直线段，算法中对长度大于 15 的直线段都进行保留，并提出了主方向判别算法，对这些直线段的数量、长度进行分析，判断物体的主体方向.

2.3　主方向判别及抓取角度的检测

以 Hough 得到的直线组 $L = \{(k_1, b_1), (k_2, b_2), \cdots, (k_v, b_v)\}$ 为基础，本文设计了一个主方向提取算法来检测抓取角度. 首先，统计 Hough 变换提取到的直线倾斜角. 直线的倾斜角为

$$\psi_i = \arctan(k_i) \tag{13}$$

式中，ψ_i 表示第 i 条直线的倾斜角.

我们将[0°，180°)这个角度区间划分为 20 个子区间：[0°，10°)，[5°，15°)，[10°，20°)，…，[160°，170°)，[165°，175°)，[170°，180°)，将[175°，180°)和[0°，5°)合并为一个区间，它们都在 0°的左右. 将每一个 ψ_i 按照它的大小进行分类，36 个子区间对应着 36 个类别，并对每一个类别进行计数得到 $N = \{n_1, n_2, n_3, \cdots, n_{36}\}$，$n_s$ 表示第 s 个类包含直线数的计数. 子区间重叠设计是为了避免处在边界角度周围的直线被分散计数. 求取每个区间中所有直线倾斜角的平均值，将这个平均值作为这个区间的关键值，代表该区间的直线的整体朝向.

我们定义与抓取角度垂直的方向为一个物体的主方向. 对于一个球形物体，主方向可选定任意一个方向. 显然一个物体的主体方向单纯的由提取直线的数量多的子区间的关键值决定并不合适. 提取到的直线段的长度以及这些直线段到抓取点的距离也都会影响到抓取角度，也就是主方向的选择.

基于此，本文给出了一个主方向的判别方法. 依据 Hough 矩阵中的端点信息，计算出每个类中的长直线段长度的最大值 $l^s_{\max}$ 以及抓取点到各长直线段

距离的平均值 d^s_{ave}. $l^s_{\max}$ 与 d^s_{ave} 计算公式为

$$l^s_{\max}=\max_{T_s}(\sqrt{(x_{js}-x_{hs})^2+(y_{js}-y_{hs})^2}) \tag{14}$$

$$d^s_{\text{ave}}=\max\left(\frac{1}{n_s}\sum_{P_s}\frac{|k_{fs}x_c-y_c+b_{fs}|}{\sqrt{1+k^2_{fs}}},\zeta\right) \tag{15}$$

式中，T_s 代表划分到第 s 个子区间的直线段的端点集，$\{(x_{js},y_{js}),(x_{hs},y_{hs})\}\in T_s$. P_s 代表的是划分到第 s 个区间的直线的斜率与截距组成的元组的集合，$(k_{fs},b_{fs})\in P_s$，ζ 代表 d^s_{ave} 的下界. (x_c,y_c) 表示抓取点坐标，由于图像是以抓取点为中心取出，所以该向量为定值.

提取到的线段数量较多，最长直线段长度较长且到抓取点的最短距离较短的分组应被选为主方向对应分组，这里给出公式如下：

$$Sc=\max_s\left(\frac{n_s\times l^s_{\max}}{d^s_{\text{ave}}}\right) \tag{16}$$

Sc 最高对应的分组的关键值即为物体的主方向，抓取角度即为主方向的垂直方向. 算法 2.1 给出了抓取角度检测的总体算法流程.

算法 2.1　抓取角度提取算法

1 将[0°,180°)区间划分为 36 个子区间类别
2 原图片以抓取点为中心剪切出 1 个子图像
3 对子图片利用 Canny 算法进行边缘提取
4 利用 Hough 变换对边缘进行直线提取
5　for 全部提取到的直线：
6　　利用式(13)计算直线的倾斜角
7　　将直线按照倾斜角分入步骤 1 的区间类别
8　　对应区间计数 n_s 加 1
9　end for
10　for 所有的角度区间 T：
11　　for 所有的直线 in 区间 T_s：
12　　利用式(14)计算直线段的长度，维护最大长度值 $l^s_{\max}$
13　　利用式(15)计算到抓取点的距离，维护距离的平均值 d^s_{ave}
14　　end for
15　　利用式(16)计算角度区间的得分，维护其最小值及其对应的子区间
16　end for
17　最小得分区间直线平均倾斜角主方向角度
18　主方向角度正交方向角度为抓取角度.

2.4　夹持器距离的检测

抓取过程中的平行夹持器张开距离也是需要检测的量之一. 这个量与物体主方向上直线段的最大间距有关. 这个最大间距对应物体在抓取方向上的最大宽度. 只要平行夹持器之间的距离大于这个最大间隔，就可以完成抓取过程.

考虑到主方向分组中的直线段不都是平行的，需要给出直线段距离的衡量方式，本文计算分组中的点到分组中直线的距离，取这些距离中的最大值. 由于同一分组中的任意两条直线段的夹角不会超过 10°，故可以用这个点到直线段的最长距离代替直线段之间的最长距离.

最长距离 $D_{\max}$ 公式表示为

$$D_{\max}=\max_{T_r,P_r}\left(\frac{|k_{gr}x_{fr}-y_{fr}+b_{gr}|}{\sqrt{1+k^2_{gr}}}\right) \tag{17}$$

式中，T_r 表示主方向区间的端点集，$(x_{fr},y_{fr})\in T_r$. P_r 表示主方向区间的直线的斜率和截距组成的元组集合，$(k_{gr},b_{gr},)\in P_r$. 最终的平行夹持器张开距离 h 为

$$h=2D_{\max} \tag{18}$$

3　实验

3.1　抓取点位置检测实验

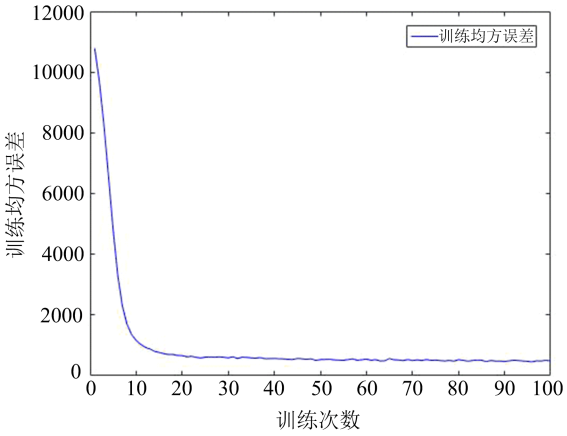
本文训练了一个基于 AlexNet 的卷积神经网络，用于检测物体的抓取点位置. 用于训练和测试的数据集采用康奈尔大学的抓取数据集，该数据集可以在 http://pr.cs.edu.cn/deepgrasping/下载. 这个数据集包含 280 个不同物体的 1 035 张 RGB-D 图片，包含物体图片、抓取框描述文件、点云文件和背景文件 4 个部分，本文的只使用其中的 RGB 三个通道. 数据集中的图片大小为 640×480×3. 由于 AlexNet 的输入图片大小为 224×224×3，本文对数据集中的图片做剪切操作，若采取缩放操作，缩放操作会导致原图片中物体变形，影响网络的判断. 剪切操作不仅可以保留原物体的特征，还可以使得物体像素在图片中的占比增加，方便特征提取. 剪切的过程中按照标签数量进行平移，处理多标签回归问题.

本文的网络以抓取点位置在图像空间坐标系下的坐标 (x,y) 为输出. 我们用一块 NVIDIA GTX1080Ti GPU 进行训练，在整个数据集上训练 20000 次. 每次训练都将数据集分为诸多小的批次，以每个批次的图片为单位进行误差反传以及参数调整，降低网络的计算量. 每一个批次包含 64 张图片，选择 2 的整数次幂作为一个批次的图片数量，便于使用 GPU 进行计算.

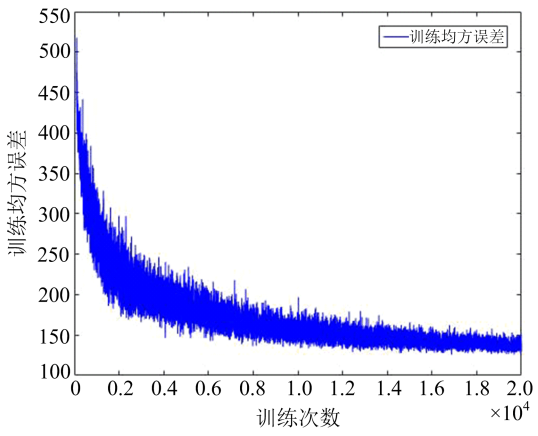
对于网络参数的调整，本文采用 Adam 参数调整方法[17]进行训练，权值 w 对应调整参数 β_1 取 0.9，偏量对应调整参数 β_2 取 0.999，学习率 η

取 10^{-6}. β_1 与 β_2 依照文献[17]作者建议进行取值，学习率取值是综合了训练速度与准确率的结果. 进行归一化时，为了避免标准差为 0，为标准差增加一个微小的大于零的量 σ，取 10^{-8}. 网络的误差描述选用均方误差，累计一个批次中所有样本的误差再进行误差反传. 为了抑制过拟合，本文引入了 $L2$ 正则化方法，为误差函数增加了一个正则的惩罚项，最终网络误差的表达形式为

$$E_i = \sum_j ((x_{ji} - \tilde{x}_{ji})^2 + (y_{ji} - \tilde{y}_{ji})^2) + \frac{\lambda}{2N_i}\sum_w w^2 \tag{19}$$



(a)第1至100次训练



(b)第100至第20000次训练

图 3　训练的误差变化

Fig. 3　Error for the training process

3.2　物体抓取角度及指间距离的检测实验

利用预测的抓取点，我们以这个抓取点为中心在图像中提取一个正方形区域. 分析这个区域，以此提取物体抓取角度. 本文在原 224×224×3 的图像中以抓取位置为中心提取一个 64×64 的区域，并将这个区域做一个尺度变换成 299×299 的图片. 这个尺度变换的目的是使得物体的边缘更加容易提取，使得抓取角度的提取更加准取.

针对上述图片，进行 Canny 边缘提取和 Hough 变换. Canny 变换中选取大小为 5×5，σ 为 1.3 的高斯核进行误差滤波，之后采用 Sobel 算子来检测图像的中的垂直水平边缘，算子形式为

$$\begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{pmatrix} \tag{20}$$

左式为垂直边缘提取算子，右式为水平边缘提取算子. Canny 算法滞后阈值检测部分低阈值取式中，E_i 表示第 i 个批次的总误差，第一部分即均方误差，第二部分为 $L2$ 正则化部分，λ 为正则化参数，本文取 0.05，N_i 表示第 i 个批次的样本总数. 同样，为了抑制过拟合现象，在全连接层中引入 dropout 机制，即在训练过程中使得一部分神经元失活来增加网络的泛化性能，本文的 dropout 参数取 0.5.

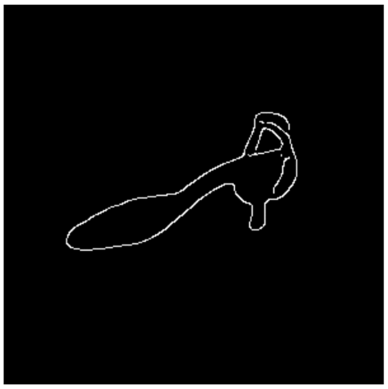
图 3 显示了网络训练过程中训练集和测试集中的误差变化. 如果我们将图片分割成 800 个大小相同的正方形，网络给出的预测点与标签中的抓取点在同一个正方形中.

10，高阈值取 25，一般高低阈之间的比值应在 2 到 3 之间.

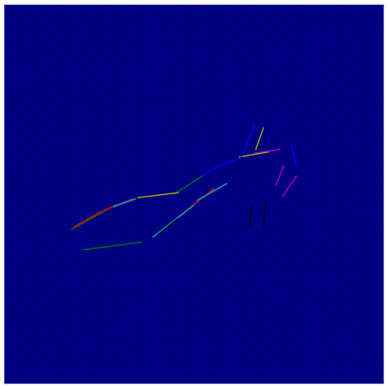
此后，对于 Canny 算法得到的图像进行 Hough 变换，限制提取直线段的长度，长度小于 15 的直线段将被舍弃，这样可以避免噪点的影响并减少后续预测角度部分算法的计算量. 利用 Canny 算法与 Hough 变换对图像处理结果如图 4 所示. 对 Hough 变换提取到的直线按照角度进行分类，按照主方向提取算法预测物体的抓取角度，进而给出夹持器距离的预测. 利用图 4 中提取到的直线进行主方向提取的实验结果如表 1 所示. 表 1 给出了检测到的直线段的相关信息，“区间”表示直线段切斜角所属的区间，n 表示倾斜角属于该区间的直线段的数量，$l_{\max}$ 表示该区间直线段长度的最大值，d_{ave} 表示该区间直线段到抓取点距离的平均值. 从图中的结果来看，Sc 最高的[30°，40°)区间被选为主方向区间，区间中所有直线倾斜角平均值为 33.67°，该角

度即为区间的代表值，也是物体的主方向. 从表 1 可以看出，单一的关注 n、l_{max} 或者 d_{ave} 中的一个量，都会造成对于物体主方向的错误判断. 利用 Sc 来进行判断是一个合理的解决方案.

(a)原图像



(b)Canny边缘提取的结果



(c)Hough变换的结果

图 4　网络训练过程中训练集和测试集中的误差变化

Fig. 4　Error changes of training and test sets in the process of network training

表 1　图 4 图像主方向提取算法结果

Tab. 1　Main orientation detection result for Fig. 4

区间	n	$l_{max}(px)$	$d_{ave}(px)$	Sc	区间	n	$l_{max}(px)$	$d_{ave}(px)$	Sc
[0°,10°)	4	**46.39**	20.73	8.95	[55°,65°)	1	19.41	58.12	0.33
[5°,15°)	4	**46.39**	20.73	8.95	[60°,70°)	2	25.63	41.04	1.24
[10°,20°)	1	18.03	12.18	1.48	[65.,75°)	3	25.63	38.67	1.99
[15°,25°)	3	24.70	13.35	5.55	[70°,80°)	1	18.03	33.92	0.53
[20°,30°)	**5**	38.95	16.17	12.05	[75°,85°)	2	19.10	51.47	0.74
[25°,35°)	**5**	38.95	14.08	13.83	[80°,90°)	2	19.10	51.47	0.74
[30°,40°)	**5**	40.61	7.33	**27.68**	[95.,105.)	1	17.46	83.51	0.21
[35°,45.)	3	40.61	**6.41**	19.01	[100°,110°)	1	17.46	83.51	0.21
[50°,60°)	1	19.41	58.12	0.33					

我们采用卷积神经网络从分类和回归的角度对抓取角度进行了检测. 卷积神经网络对于角度的回归误差为±21.4.，若将抓取角度按每 10°一类分为 18 类，卷积神经网络的分类正确率仅为 66.7%. 本文的角度提取算法本质上也是一个分类问题，抓取角度分类的准取率为 95.4%，正确分类的前提下角度预测误差小于 5°，可以满足抓取角度预测的需求. 各种算法应用于抓取角度检测的结果如表 2 和 3 所示. 由表中的数据可知，基于主方向提取的抓取角度检测算法有着更好的表现.

表 2　基于分类的抓取角度预测算法比较

Tab. 2　Accuracy for grasping angle detection based on classification

算法	分类准确率
3 层全连接	63.6%
3 层卷积+2 层全连接	51.7%
改进 AlexNet	66.7%
本文主方向提取	95.4%

表 3　基于回归的抓取角度预测算法比较

Tab. 3　Accuracy for grasping angle detection based on regression

算法	拟合误差
3 层全连接	31.7°
3 层卷积+2 层全连接	29.5°
改进 AlexNet	21.4°

注：表中的改进 AlexNet 表示一个卷积层与 AlexNet 相同，包含 4 层全连接层网路.

3.3　抓取位置检测结果及讨论

综合对抓取点的位置，抓取角度以及夹持器张开距离的预测，我们最终可以给出物体抓取位置的预测. 图 5 给出了部分物体的抓取位置判别示意图. 利用测试集中的图片进行判别，测试集中包含 392 幅待抓物体的图片，19 个物体因为抓取位置点的预测偏差导致抓取失败，18 个物体由于抓取角度的偏差导致抓取失败，13 个物体由于夹持器距离预测不佳而导致失败，总体物体抓取位置预测成功率为 87.2%.

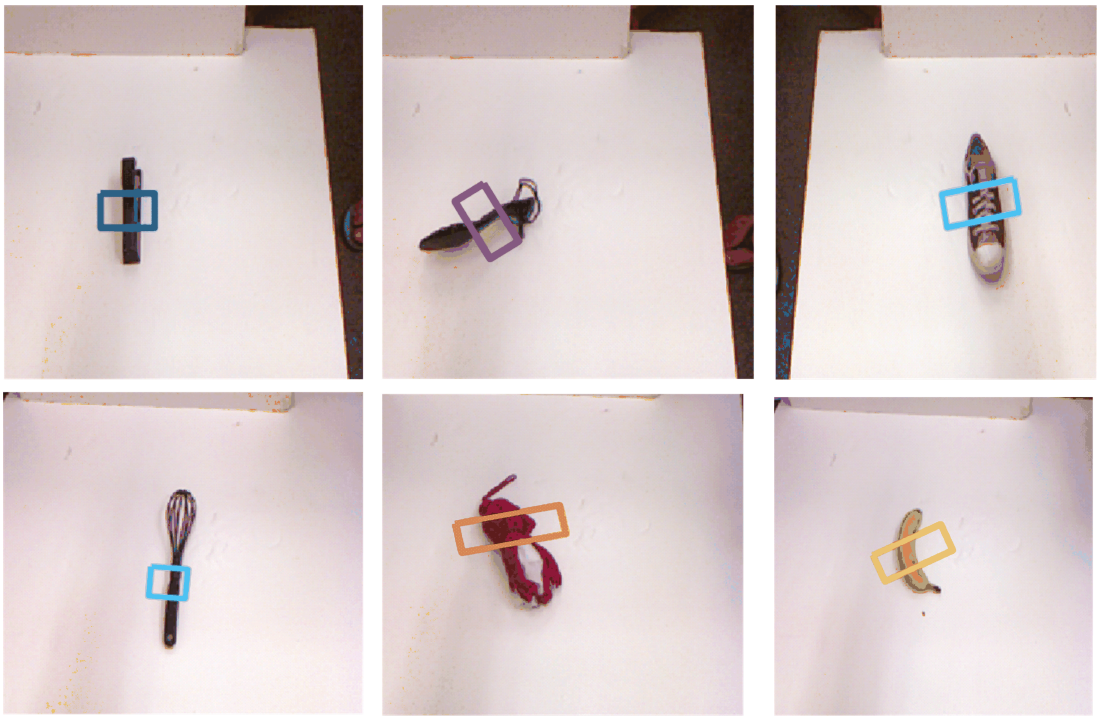


图 5　部分预测正确的抓取位置展示

Fig. 5　Grasping detection result presentation

表 4 总结了目前抓取位置预测方法正确抓取率对比. 本文的算法较文献[1]以及文献[11]的算法抓取位置预测的准确性有明显的提升，与文献[12]以及文献[13]的检测算法准确率接近. 文献[12]的方法将多标签集中于一张图片，使得图片标签复杂同时训练街图片数量也相应的减少. 这样会增加网络对标签拟合的难度，训练集的减少也使得训练难度增加，需要借助大数据集进行训练，还要利用抓取数据集对网络参数进行整定，提高了网络的训练难度. 本文方法标签简洁，数据集更大，使得神经网络训练更加简单. 本文算法可以直接进行网络的训练，不需要其他数据集辅助. 同时对抓取角度的解释性也更好. 利用传统视觉方法预测相较于网络更容易作参数的调节，算法的鲁棒性更好. 相较于 Wang[13]，本文的网络更容易训练，参数更易调整. 另外我们应用的网络结构更加简单，检测算法的实时性更好，最终也给出了更高的抓取准确率.

表 4　抓取位置检测算法准确率对比

Tab. 4　Accuracy for grasping detection

作者	算法	准确率
	随机	6.7%
Y. Jiang[1]	快速搜索	60.5%
Ian Lenz[11]	两阶段网络	73.9%
Wang[13]	两阶段＋封闭环	85.3%
Redmon[12]	多标签检测	88.0%
本文	AlexNet＋主方向提取	87.2%

本文的主要特点在于利用了抓取点位置和抓取角度之间的关系，采用传统视觉的方案代替卷积神经网络处理抓取角度的预测问题. 实际应用中，由于应用场景的背景容易获取，可以通过视觉算法去除背景，免去背景对于抓取位置预测的影响. 卷积神经网络对于图像中很小的物体的预测抓取位置偏差较大，主要是由于过小的物体在图像中提供的信息不足，导致网络判别难度增加. 如果物体本身具有非常

复杂的纹理,也可能会影响物体主方向的判断,之后的研究中考虑拟引入深度信息进行处理.

4 结论

本文给出了一种基于RGB图像的抓取位置提取方案,主要做了3个方面的工作:

(Ⅰ)改进了AlexNet网络,以更好的预测物体的抓取点位置;将批量归一化方法引入AlexNet加快训练的同时也抑制了过拟合.

(Ⅱ)考虑到抓取点位置与抓取角度的关联,以抓取点为中心提取图像区域进行抓取位置检测.在提取的过程中,每一个抓取点对应提取一个区域,增广了数据集,抑制了过拟合,同时也避免了多标签拟合问题,充分利用了数据集信息.

(Ⅲ)规避了卷积神经网络在角度上预测的弊端,采用传统视觉方式进行抓取角度的预测;并且设计了物体主方向的判别算法,以此为依据进行抓取角度的提取.

实验在测试集上表现出了较好的效果,对于物体抓取位置研究有一定的推动作用.本文的提取方法对本身纹理复杂、拍摄偏角过大的物体预测结果较差,今后的研究中可以引入深度信息对于表面纹理和拍摄角度带来的偏差进行修正.

参考文献(References)

[1] JIANG Y, MOSESON S, SAXENA A. Efficient grasping from RGBD images: Learning using a new rectangle representation [C]//Proceedings of IEEE International Conference on Robotics and Automation. Shanghai, China: IEEE, 2011: 3304-3311

[2] PIATER J H. Learning visual features to predict hand orientations [C]//International Conference on Machine Learning. Beijing, China: Computer Science Department Faculty Publication Series, 2002: 148-153.

[3] MORALES A, SANZ P J, DEL POBIL A P. Vision-based computation of three-finger grasps on unknown planar objects [C]//Proceedings of the International Conference on Intelligent Robots and Systems. Lausanne, Switzerland: IEEE, 2002: 1711-1716.

[4] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-basedlearning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[5] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]//Proceedings of the Neural Information Processing Systems. Lake Tahoe, USA: MIT Press, 2012: 1097-1105.

[6] SIMONVAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [J]. arXiv preprint, 2014: arXiv: 1409.1556.

[7] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016: 770-778.

[8] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015: 1-9.

[9] SZEGEDY C, VANHOUCKE V, IOFFE S. Rethinking the inception architecture for computer vision [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016: 2818-2826.

[10] SZEGED C, LOFFE S, VANHOUCKE V, et al. Inception-v4, inception-resNet and the impact of residual connections on learning [C]// Association for the Advancement of Artificial Intelligence. San Francisco, USA: ACM, 2017: 4-12.

[11] LENZ I, LEE H, SAXENA A. Deep learning for detecting robotic grasps [J]. The International Journal of Robotics Research, 2015, 34(4-5): 705-724.

[12] REDMON J, ANGELOVA A. Real-time grasp detection using convolutional neural networks [C]// Proceedings of the Conference on Robotics and Automation. Seattle, USA: IEEE, 2015: 1316-1322.

[13] WANG Z C, LI Z Q, WANG B, et al. Robot grasp detection using multimodal deep convolutional neural networks [J]. Advances in Mechanical Engineering, 2016, 8(9): 1-12.

[14] IOFFE S. Batch normalization: Accelerating deep network training by reducing internal covariate shift [J]. arXiv preprintv, 2015: arXi: 1502.03167

[15] CANNY J. A computational approach to edge detection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986: 679-698.

[16] BALLARD D H. Generalizing the Hough transform to detect arbitrary shapes[J]. Pattern Recognition, 1981, 13(2):: 111-122.

[17] KINGMA D P, BA J. Adam: A method for stochastic optimization [J]. arXiv preprint arXiv, 2014: 1412.6980.