doi:10.3772/j.issn.1002-0470.2018.01.008

基于深度学习的机器人抓取位置检测方法①

闫 哲②* 杜学丹*** 曹 淼*** 蔡莹皓** 鲁 涛** 王 硕**

(*哈尔滨理工大学自动化学院 哈尔滨 150080)

(**中国科学院自动化研究所复杂系统管理与控制国家重点实验室 北京 100190)

摘 要 进行了机器人的智能抓取研究,提出了一种基于深度学习的机器人抓取位置检测方法。该方法以目标物体的多模态特征作为训练数据,采用无监督学习与监督学习相结合的方式对目标物体的最优抓取位置进行学习。在无监督学习阶段,使用压缩自动编码器(CAE)对神经网络进行无监督逐层预训练;在监督学习阶段,使用反向传播算法对整个网络进行监督微调。仿真实验结果验证了该方法能够对目标物体的最优抓取位置做出精确的判断。在 Universal Robot 5 机器人上进行了抓取实验,实验结果表明该方法的抓取成功率较高,能够应用到与机器人抓取相关的任务中。

关键词 深度学习,机器人抓取,位置检测,压缩自动编码器(CAE)

0 引 言

机器人抓取研究一直以来都是机器人领域的一个重点研究方向,在工业、探索、服务、军事等方面得到了广泛的应用。然而在实际的工业生产中,机器人在执行抓取任务时,大多情况下只是简单重复地执行预设的抓取动作,当待抓取对象的状态或机器人所处工作环境发生改变时,抓取任务则会失败。针对这个问题,人们对机器人抓取提出了更高的要求,希望机器人不再依靠预定的程序,在非结构化环境中针对不同的抓取目标执行更加合理的抓取动作,从而使得抓取更加智能化。因此,机器人的智能抓取研究有着重要的理论意义和实用价值。

受到目标模型各种特性如形状、姿态、材质、重量等因素的影响[1,2],智能机器人抓取研究就变得富有挑战性。目前,大多数抓取研究工作通常侧重于使用深度学习方法[3]学习抓取特征。与传统的手工特征提取方法[4-7]相比,深度学习的优势在于特征提取环节不需要人为干预。通过监督学习、无监督学习等方式,神经网络模型可获得目标物体在机器人坐标系下的位置和姿态,进而执行抓取动作。

在机器人抓取任务研究中,抓取位置检测一直是研究重点之一。针对抓取位置检测环节,文献[8,9]分别提出使用改进的自动编码器对目标物体的最优抓取位置进行多模态特征学习,从而在目标物体上搜索出有效的抓取位置。文献[10]将抓取位置检测视为一种 18 路二分类问题,采用卷积神经网络(convolutional neural network, CNN)模型学习目标物体的最优抓取位置,最终预测出目标物体上的抓取点在二维图像坐标系下的坐标以及夹持器抓取目标物体时所需的旋转角。文献[11]提出一种无监督特征学习方法,首先重建目标物体的三维信息,其次对局部场景进行扫描并提取目标物体的三维局部几何描述子及其对应的标记信息。通过三维神经网络对描述子进行特征学习,获得目标的 6

① 国家自然科学基金(61503381)和北京市科技计划(Z171100000817009)资助项目。

② 男,博士,教授;研究方向:控制理论及应用,复杂电子系统的电磁预测;E-mail:yanzhehrb@163.com

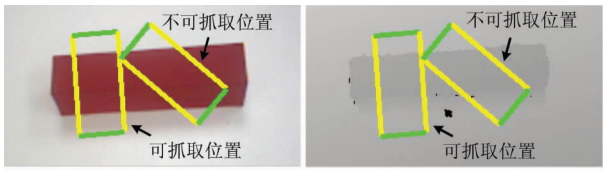
(收稿日期:2017-08-22)

自由度位姿,根据所得目标位姿最终确定目标物体的抓取位置。文献[12]首先获取多视角场景的彩色及深度图像,然后对彩色图像做二维目标分割并将分割结果整合成三维点云,整合的三维结果与目标物体的预扫描三维模型进行匹配后可获得目标物体的6自由度位姿,最后根据目标位姿执行抓取任务。本文从抓取位置检测方法的易实现性出发,提出了一种基于深度学习的机器人抓取位置检测方法。本方法对最优抓取位置的多模态特征进行学习,从而获得能够检测出最优抓取位置的深度网络模型。在抓取位置学习阶段,本方法首先使用压缩自动编码器(contractive autoencoder, CAE)对神经网络模型进行无监督逐层预训练,其次使用反向传播算法对整个网络的参数进行监督微调,采用两种手段相结合的方式,目的在于提高网络模型的鲁棒性和检测的准确性。当网络模型用于抓取位置检测时,仅需一张包含抓取目标的彩色图像及其对应的一张深度图像就能通过神经网络模型检测出目标物体的最优抓取位置。本方法所用模型结构简单,易于实现,具有较好的可移植性。

1　抓取位置检测

要将深度学习方法应用于抓取位置检测,首先必须明确如何根据抓取位置检测问题构建深度学习模型。本文将目标物体的抓取位置作为神经网络的学习对象,将目标物体的可抓取位置视为神经网络的正样本数据,不可抓取位置视为负样本数据,抓取位置的学习则可转化为一个二分类问题。抓取位置检测的过程可描述为:根据目标物体的大小选择一组不同大小不同方向的矩形框并在目标物体上滑动提取抓取位置,将每个抓取位置均输入到神经网络中进行预测,预测为正且预测概率值最高的抓取位置作为最优抓取位置。

抓取位置的学习结合了多个模态的信息,包括3通道二维彩色图像信息、1通道深度信息和3通道表面法向量信息。图1示出了目标物体上的可抓取位置与不可抓取位置。



(a) 彩色图像　　(b) 深度图像

图1　抓取位置信息

1.1　自动编码器算法的选择

典型的自动编码器(AE)通常由输入层、隐含层和重构层这3层组成,从数据输入到重构输出,其过程分为数据编码和数据解码两部分。数据编码通过映射函数f将输入数据$\boldsymbol{x}\in\boldsymbol{R}^N$映射为隐含层输出数据$\boldsymbol{y}\in\boldsymbol{R}^H$,即

$$\boldsymbol{y}=f(\boldsymbol{x})=s_f(\boldsymbol{W}\boldsymbol{x}+\boldsymbol{b}y) \tag{1}$$

其中s_f为激活函数,$\boldsymbol{W}\in R^{H\times N}$为编码器的权值矩阵,$\boldsymbol{b}_y\in\boldsymbol{R}^H$为偏差向量;数据解码通过逆映射函数$g$将隐含层数据$\boldsymbol{y}\in\boldsymbol{R}^H$映射为重构层输出数据$\boldsymbol{x}'\in\boldsymbol{R}^N$,即

$$\boldsymbol{x}'=g(\boldsymbol{y})=s_g(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{y}+\boldsymbol{b}_{x'}) \tag{2}$$

这里,解码器的权值矩阵使用编码器权值矩阵的转置。AE的训练目的在于寻找一组参数$\boldsymbol{\theta}=\{\boldsymbol{W},\boldsymbol{b}_y,\boldsymbol{b}_{x'}\}$,使得AE在训练样本上的重构误差最小,其对应最小化式

$$J_{\mathrm{AE}}(\boldsymbol{\theta})=\sum_{x\in D_n}L(\boldsymbol{x},g(f(\boldsymbol{x}))) \tag{3}$$

所示的目标函数,其中D_n为训练样本集;L为重构误差。

AE用于模型训练时容易出现过拟合现象,特别是在训练样本较小的情况下,模型的泛化性能更差。因此通常的做法是在AE中加入规则化项,对网络模型的权重进行惩罚,用以防止过拟合现象的发生。规则化项中最简单的形式是权值衰减(weight decay, WD)[13],加入权值衰减后的目标函数为

$$J_{\mathrm{AE}_{+\mathrm{wd}}}(\boldsymbol{\theta})=\left(\sum_{x\in D_n}L(\boldsymbol{x},g(f(\boldsymbol{x})))\right)+\lambda\|\boldsymbol{W}\|_2^2 \tag{4}$$

其中λ为规则化项的权重,控制权值衰减的强度。

为了对输入数据$\boldsymbol{x}$有更鲁棒性的表示,文献[13]提出了一种新的规则化项,即编码映射函数f对于输入数据$\boldsymbol{x}$的雅可比矩阵$J_f(\boldsymbol{x})$的F范数平方:

$$\| J_f(\boldsymbol{x}) \|_F^2 = \sum_{i=1}^{N} \sum_{j=1}^{H} \left(\frac{\partial f_j(x)}{\partial x_i} \right)^2 \tag{5}$$

当映射函数 f 为 Sigmoid 函数时，雅克比矩阵 $J_f(\boldsymbol{x})$ 的 F 范数平方可以进一步表示为

$$\| J_f(\boldsymbol{x}) \|_F^2 = \sum_{j=1}^{H} (f_j(\boldsymbol{x})(1 - f_j(\boldsymbol{x})))^2 \sum_{i=1}^{N} W_{ij}^2 \tag{6}$$

将其加入到 AE 的目标函数中，就能获得压缩自动编码器（CAE）的目标函数：

$$J_{CAE}(\boldsymbol{\theta}) = \sum_{x \in D_n} (L(\boldsymbol{x}, g(f(\boldsymbol{x}))) + \lambda \| J_f(\boldsymbol{x}) \|_F^2) \tag{7}$$

在式（7）中，$\| J_f(\boldsymbol{x}) \|_F^2$ 的训练目标是隐含层表示对所有输入数据 x 的偏导数都接近或等于 0，即训练目标只与偏导数不为 0 的样本有关。这样做的最终结果是当输入样本发生改变时，隐含层表示不会发生变化。而对于重构误差 L，其训练目标则是尽可能多地保留训练样本中的有用信息。因此目标函数式（7）最终的训练结果是，网络模型对训练样本有较好重构性的同时，能够捕获训练样本附近出现的微小变化，使得网络模型对训练样本附近的扰动或噪声具有较强的鲁棒性。

由于压缩自动编码器有着抑制噪声的能力，因而可以在一定程度上防止网络模型出现过拟合现象，从而提高网络模型的泛化性能。

1.2　抓取位置检测的网络结构

本文采用栈式压缩自动编码器（stacked contractive autoencoder，SCAE）的网络结构来构建本文抓取位置检测算法的神经网络模型。SCAE 由多个 CAE 堆叠而成，其结构如图 2 所示。神经网络模型由无监督学习预训练与监督学习微调相结合的方式[14]训练而成，首先使用 CAE 对网络模型进行自底向上、逐层训练的方式实现无监督学习预训练，然后使用监督学习方法训练顶层分类器并微调整个网络模型的参数。

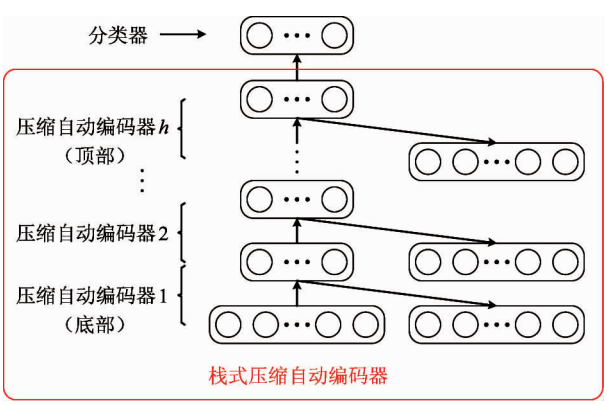


图 2　抓取位置检测的网络结构

1.3　抓取位置检测的结果表示

本方法使用一个旋转的矩形框来表示目标物体的最优抓取位置。在检测目标物体的抓取位置时，输入到神经网络的数据为一张包含目标物体的彩色图像及其对应的一张深度图像，输出为最优抓取位置四个顶点的坐标信息及抓取位置在二维图像平面内的旋转角。抓取位置表示如图 3 所示，根据人的抓取习惯，矩形框长边所在方向为机器人抓取时夹持器的张开方向；θ 表示矩形框的长边相对于图像坐标系 x 轴旋转角度的大小，箭头的指向表示矩形框的旋转方向。

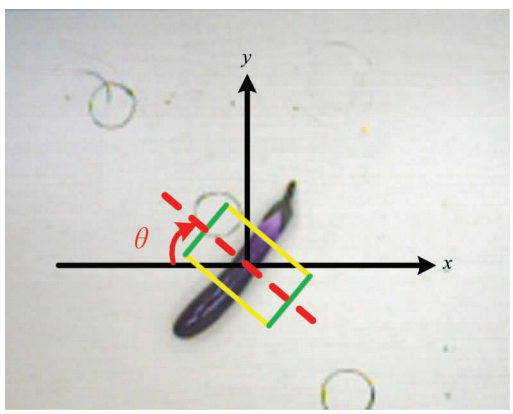


图 3　抓取位置表示

2　抓取位置检测的训练流程

本节将抓取位置检测的训练流程分为数据预处理、多模态特征预训练和监督微调这 3 部分。其中，数据预处理用于将样本图像转化为符合神经网络输入的数据格式，多模态特征预训练用于提高神经网络隐含层提取特征的能力，监督微调用于训练神经网络顶层分类器并微调网络模型的参数。

2.1　数据预处理

模型训练开始前，本文首先对原始训练数据进行预处理。数据预处理采用文献[8]所用方法，将

标记区域对应的 7 个模态数据统一转化为 24 × 24 × 7 的特征向量,其中 24 × 24 表示单模态的特征图尺寸,7 模态数据包括彩色图、深度图及基于深度图的表面法向量特征图等多模态特征数据。

2.2　多模态特征预训练

本文将数据预处理阶段得到的多模态特征向量输入到 SCAE 中做多模态特征预训练。逐层预训练过程中,每层 CAE 的训练目标均为最小化目标函数式(7),训练过程[15]如算法 1 所示。

算法 1:CAE 预训练算法

Function trainCAE($\boldsymbol{x}$, $\boldsymbol{\theta}$)

1: 随机初始化 θ
2: **for** $Iter$ = 1 to $MaxIter$ **do**
3:　前向传播:根据式(1)和式(2)计算 $\boldsymbol{y}$ 与 $\boldsymbol{x}$'
4:　通过随机梯度下降法优化式(7)
5: **end for**
6: **return** (x, $\boldsymbol{\theta}$)

在预训练算法中,本文将预训练的迭代次数限制为 $MaxIter$, 其目的在于防止模型过度学习训练样本集的特征,导致出现过拟合现象。对于由 h 层 CAE 堆叠而成的 SCAE 网络,本文采用的分层预训练过程[15]如算法 2 所示。

算法 2: SCAE 预训练算法

Function trainSCAE ($\boldsymbol{n}$, $\boldsymbol{x}$, $\boldsymbol{d}$)

输入:h,SCAE 的层数;$\boldsymbol{x} = [x_1, x_2, \cdots, x_n]^{\mathrm{T}}$,训练数据,每行为一个样本;$d$,每层的维度序列
输出:$\boldsymbol{\theta} = \{\theta_1, \theta_2, \cdots, \theta_h\}$,SCAE 的参数
1: **for** i = 1 to h **do**
2:　根据维度 d_{i-1} 和 d_i 随机初始化 $\boldsymbol{\theta}_i$
3:　($\boldsymbol{x}_i$, $\boldsymbol{\theta}_i$) = **trainCAE**($\boldsymbol{x}_{i-1}$, $\boldsymbol{\theta}_{i-1}$)
4: **end for**

对于如何处理多模态特征数据,有的做法是忽略模态间的差异性,将多模态数据视为普通数据来处理;有的是将多模态输入数据的每个模态分开进行单独训练,然后将各个模态的高维特征相融合[9,15];也有的则是将多模态数据融合后进行共同训练,并在目标函数中加入针对不同模态的规则化惩罚[8]。

对于多模态输入数据,本文在底层 CAE 的目标函数中加入多模态规则化项[8]

$$f_{\mathrm{m}}(\boldsymbol{W}) = \sum_{k=1}^{K} \sum_{j=1}^{H} I\{(\max_{i} S_{ki} \mid W_{ij} \mid) > 0\} \tag{8}$$

其中 $\boldsymbol{S}$ 为模态矩阵,大小为 $K \times N$,K 为模态数,N 为多模态特征向量的维度;S_{ki} 表示输入单元 x_i 在模态 k 中的元素。当括号中的值为真时 I 为 1,否则为 0。

2.3　监督微调

在逐层无监督预训练过程中,每层 CAE 只是去学习如何重构它的输入,即能够在最大程度上还原其输入数据的特征,还不能对输入样本进行分类。为了实现对输入样本的分类,本文在 SCAE 的最顶编码层添加一个 softmax 分类器,对最顶层编码层输出的高维特征数据进行 softmax 回归,之后采用反向传播算法对整个网络的参数进行训练。

在监督学习阶段,本文将分类层权重的学习与隐含层权重的微调相结合,训练目标为最大化式(9)所示的似然函数:

$$\begin{aligned}\boldsymbol{\theta}^{*} = \arg\max_{\theta}\Big(\sum_{t=1}^{M} \log P(\hat{y}^{(t)} \\ = y_T^{(t)} \mid x^{(t)}; \boldsymbol{\theta}) - \boldsymbol{\beta}_1 f_1(\boldsymbol{W}^{[1]}) - \cdots - \beta_h f_h(\boldsymbol{W}^{[h]})\Big)\end{aligned} \tag{9}$$

其中,θ^{*} 为似然函数最大时神经网络模型所有参数的取值;$\hat{\boldsymbol{y}} \in \boldsymbol{R}^{M}$ 为分类器对多模态输入数据的预测值;$\boldsymbol{y}_{\mathrm{T}} \in \boldsymbol{R}^{M}$ 为多模态输入数据对应的真实值;M 为分类器输出的数据维度;f_1 至 f_h 为 h 个 CAE 对应的规则化惩罚函数;β_1 至 β_h 为每个规则化惩罚函数对应的权重;$\boldsymbol{W}^{[1]}$ 至 $\boldsymbol{W}^{[h]}$ 为 h 个 CAE 对应的编码器权值矩阵。

3　实验结果分析

为了验证本文所提机器人抓取位置检测算法的有效性,分别从仿真实验与机器人抓取实验两方面进行测试。实验的系统环境为 Windows 7 操作系统,神经网络模型的训练环境及仿真实验环境为 Matlab R2013a,抓取实验的测试环境为 Visual Studio 2013。在线机器人抓取实验中,采用华硕 Xtion PRO 深度相机进行彩色图像及深度图像的采样。

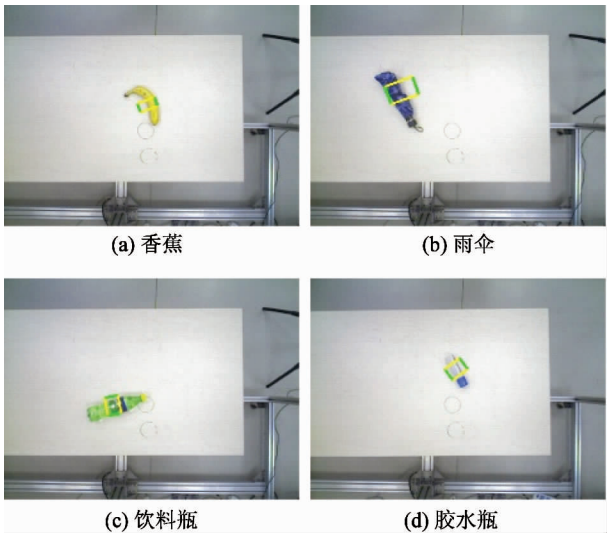
构建的神经网络模型包含两个隐含层，每个隐含层有 200 个节点，softmax 分类层有两个输出节点。本实验所需训练样本集包含不同形状、大小、方向等情形下的 233 个物体，共 885 张样本图像。每张图像中的抓取目标各标记 3 到 5 个正样本抓取位置和负样本抓取位置。

在抓取位置的二分类问题上，本文首先将目标物体与背景相分离，定位出包含目标物体的最小矩形区域并将其作为待检测区域；其次选择若干个抓取框在待检测区域上滑动提取抓取位置并对抓取位置进行分类。假设待检测区域的大小为 $W \times H$，矩形框的大小为 $w \times h$（$w < \min(W, H)$，$h < \min(W, H)$），矩形框在图像平面内的旋转方向为 $\theta(0 \leqslant \theta \leqslant \pi)$，滑动步长为 $s(s < \min(W, H))$，因而每个矩形框在待检测区域上遍历后可以提取 $((W\cos\theta + H\sin\theta - w)(H\cos\theta + W\sin\theta - h)/s^2 + 1)$ 个抓取位置。采用如上方式选择抓取框，使得本文的抓取位置检测方法可实现检测准确度与检测速度之间的权衡。

3.1　仿真实验

本研究的部分仿真实验结果如图 4 所示，图 4（a）~（d）中的目标物体分别为香蕉、雨伞、饮料瓶和胶水瓶。图 4（a）、（b）中的物体为训练样本集中出现的物体，图 4（c）、（d）中的物体未出现在训练样本集中。仿真实验结果表明，本文所提的抓取位置检测方法通过学习训练样本集中的物体的最优抓取位置，可以实现对目标物体最优抓取位置的检测。当出现新的未知的抓取目标时，本文方法同样能够检测出抓取目标的最优抓取位置，这说明本方法具有较好的泛化性。

为了进一步说明本文方法对最优抓取位置的检测性能，从两个角度分别做了相应的对比实验：（1）保持多模态输入特征不变，与带有权值衰减的自动编码器（AE_{+wd}）[13] 算法训练所得的神经网络模型作对比；（2）与传统的位置检测方法作对比，采用的传统位置检测方法为基于重心法的抓取位置检测方法。

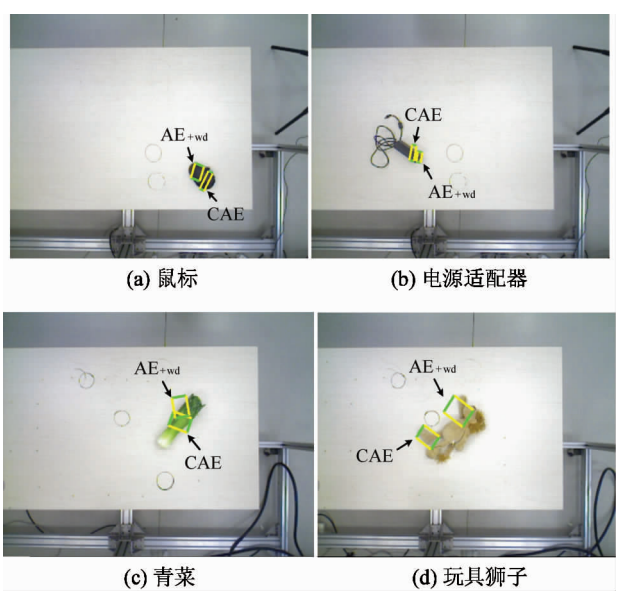


(a) 香蕉　(b) 雨伞　(c) 饮料瓶　(d) 胶水瓶

图 4　抓取位置检测结果

（1）与 AE_{+wd} 训练所得模型的对比实验

本研究采用 AE_{+wd} 对神经网络进行训练，训练所得模型用于抓取位置检测，并将检测结果与本文方法的检测结果作对比，对比实验结果如图 5 所示。图 5（a）~（d）中的目标物体分别为鼠标、电源适配器、青菜和玩具狮子，其中图 5（a）、（b）中的目标物体为训练样本集中出现的物体，图 5（c）、（d）中的目标物体没有出现在训练样本集中。对比实验结果显示，当抓取目标为训练样本集中的物体时，AE_{+wd} 和 CAE 都能较好地检测出目标物体的最优抓取位置；



(a) 鼠标　(b) 电源适配器　(c) 青菜　(d) 玩具狮子

图 5　AE_{+wd} 与本文 CAE 训练方法的检测结果比较

当抓取目标不同于训练样本中的任何一种物体时,AE_{+wd}的检测结果明显较差,CAE 仍能保持较好的检测结果。实验结果表明,使用 CAE 训练所得的神经网络模型具有较强的泛化性能,在面对未知的目标物体时能够较好地检测出目标物体的最优抓取位置。

(2) 与传统的位置检测方法的对比实验

本研究采用基于重心法的传统位置检测方法与本文方法作对比。基于重心法的抓取位置检测方法的检测过程为首先获得目标物体的位置信息,将目标物体与背景相分离并以二值图形式表示,二值图中目标物体的颜色为白,背景的颜色为黑。其次根据二值图信息求出白色区域的重心位置 (x, y) 以及与白色区域具有相同标准二阶中心矩的椭圆在二维图像坐标系下的方位角 α。

基于重心法的抓取位置检测方法的问题在于过度依赖目标物体在二维图像平面内的重心位置而忽略了目标物体的空间形态。当目标物体的重心不在物体上或者适合机器人抓取的位置位于物体边缘时,基于重心法的抓取位置检测方法将失效。本文方法用于神经网络模型训练的最优抓取位置具有一定的语义信息,因而不论适合机器人抓取的位置是否位于目标物体的重心,对本文方法的抓取位置检测没有任何影响。只要目标区域有符合的特征,都可以被视为目标物体的抓取位置。

基于重心法的传统抓取位置检测方法与本文方法的对比实验结果如图 6 所示,图 6(a) ~ (d)中的目标物体分别为刷子、玩具海豚、瓷杯和 USB 延长线。在对比实验结果中,本文将目标物体的重心位置用点标出,为了使基于重心法的抓取位置检测结果能够与本文方法的检测结果做出更加清晰的对比,将方位角 α 以旋转矩形框的形式表示出来。在图 6 (a)、(b)中,目标物体的重心都位于适合机器人抓取的区域,因而两种方法都能得到较好的检测结果;在图 6(c)、(d)中,目标物体的重心与适合机器人抓取的位置严重偏离,基于重心法的抓取位置检测方法已经不适合检测此类目标物体的抓取位置。从实验结果中可以看出,本文所提抓取位置检测方法对目标物体的特征提取能力较好,对不同形状的抓取目标具有较强的适应性。

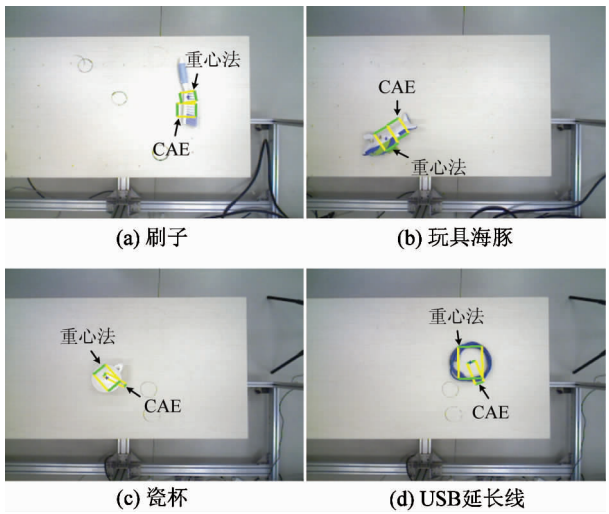


图 6　基于重心法与本文方法的检测结果比较

3.2　机器人抓取实验

本文以 6 自由度人机协作型工业机器人 Universal Robot 5 (UR5)为实验平台,进行机器人抓取实验,其中 UR5 机器人的末端夹持器采用 Robotiq 2 指夹手。机器人抓取系统结构图如图 7 所示,抓取系统由华硕深度相机 Xtion PRO、UR5 机器人和计算机三部分组成。抓取实验过程中,计算机通过深度相机获取包含抓取目标的彩色图像与深度图像,将两幅图像输入到本文算法中获得抓取目标的最优抓取位置信息,从而实现机器人的自动抓取。

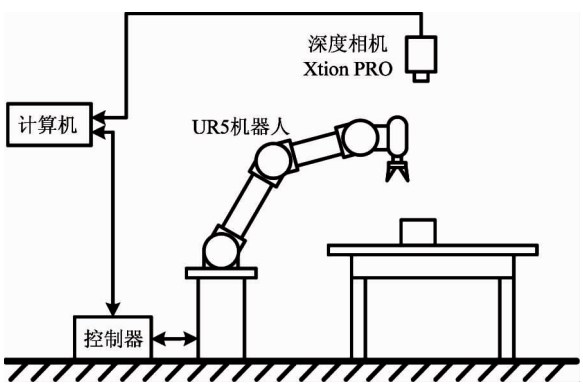


图 7　机器人抓取系统结构图

在机器人实验中，本研究用固定数量的矩形框在不同目标物体上滑动提取抓取位置。矩形框的宽 w 和高 h 的初始值分别取待检测区域短边长度的三分之一。宽 w 每隔 4 个像素点取一次值，共取 6 个不同的 w 值；高 h 每隔 4 个像素点取一次值，共取 3 个不同的 h 值，通过组合最终可以选出 18 个矩形框。本实验设定每个矩形框在待检测区域上的滑动步长 s 为 5 个像素点；矩形框的滑动角度通过预判断目标物体在图像上的大致方向，在取得的大致方向附近每隔 10°共取 5 个角度值。

本文方法得出的待抓取目标的最优抓取位置仅是其在二维图像平面内的坐标，而机器人抓取目标时需要的是最优抓取位置在机器人坐标系下的三维位姿，因而位置信息在二维坐标和三维位姿之间存在一种位姿转换关系。本实验假定机器人坐标系为世界坐标系，那么从抓取位置在图像平面内的坐标关系到抓取位置在机器人坐标系下的坐标关系的转换问题，则转化为从图像坐标系到世界坐标系的转换问题。首先根据深度相机的内参数矩阵将抓取位置从图像坐标系转换到相机坐标系中，其次通过使用外参数矩阵和抓取位置的深度信息将抓取位置从相机坐标系转化到世界坐标系中，此时即可抓取位置在机器人坐标系下的三维位置信息。关于抓取位置的三维姿态信息，本文假定所有抓取位置的 Z 轴都垂直于地面，X、Y 轴方向由抓取位置的长和宽所在的直线确定，由此就能确定抓取位置的三维姿态信息。

本研究首先进行了针对不同种物体的抓取实验，实验结果如图 8 所示。图 8 中的目标物体分别为饮料瓶、玩具松鼠和矩形盒。实验结果表明，UR5 机器人根据抓取位置检测方法得出的检测结果可以实现对不同种目标物体的有效抓取。其次，本文对在不同摆放位置下的同一物体进行了抓取实验，实验结果如图 9 所示。图 9 中的目标物体为玩具熊。实验结果表明，本文方法适用于不同摆放姿态的目标物体的抓取。

最后，本研究从图 3 至图 9 出现的目标物体中选取 7 个物体进行机器人抓取实验成功率统计，表 1 的第1列给出了7个物体的名称。本研究针对这

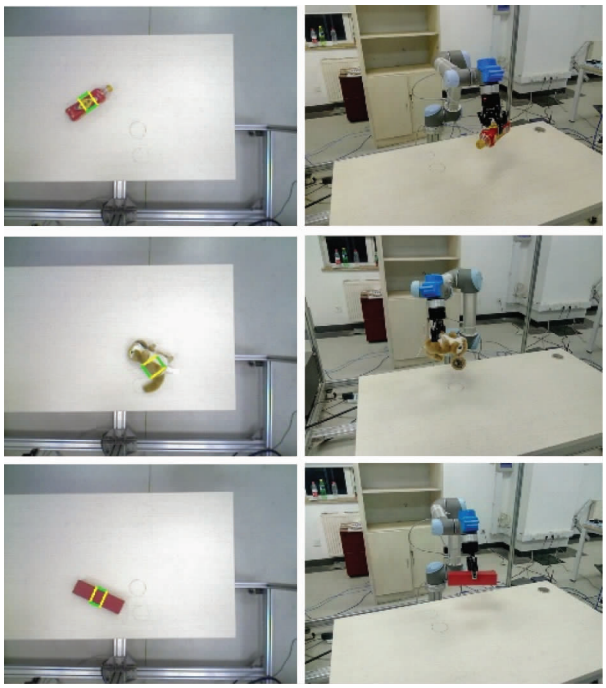


图 8　不同种类物体的抓取实验结果

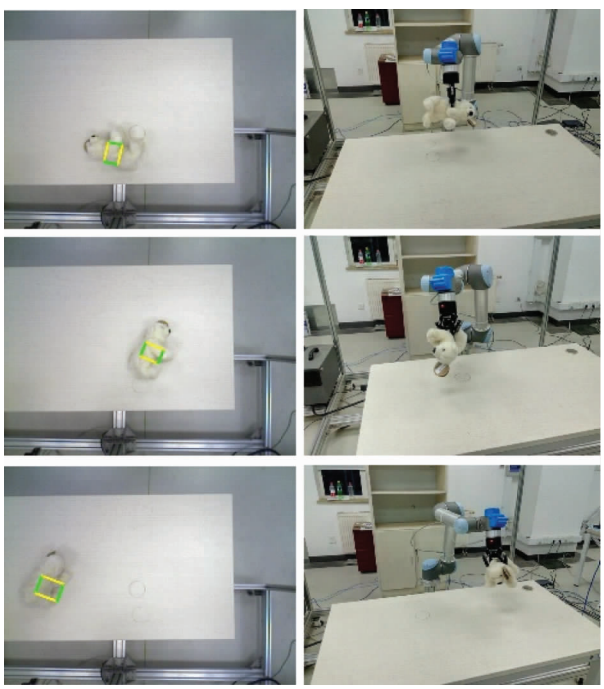


图 9　不同摆放位置的同一物体的抓取实验结果

几个物体分别从不同的摆放位置进行了共计 140 次抓取实验，并记录了各自的实验结果。在 140 次抓取实验中，位置检测算法检测出单个目标物体最优抓取位置的时间约为0.223s。从 140 次抓取实验中统计，机器人抓取目标物体的平均成功率约为

92.8%。抓取实验结果表明,本文方法能够应用在实际的机器人抓取任务中,具有一定的使用价值。

表 1　抓取实验统计结果

目标物体	抓取次数	成功次数	抓取成功率(%)
茄子	20	19	95
饮料瓶	20	18	90
青菜	20	20	100
玩具狮子	20	20	100
刷子	20	17	85
玩具松鼠	20	18	90
矩形盒	20	18	90

4　结　论

本研究针对机器人抓取问题提出了一种基于深度学习的抓取位置检测方法,并对该方法进行了仿真实验验证和机器人抓取实验验证。在仿真实验中,通过与带有权值衰减的自动编码器算法训练所得的神经网络模型作对比,实验结果表明使用压缩自动编码器算法训练所得的模型能够较好地提取目标物体的特征,在面对未知的抓取目标时具有良好的泛化性和较强的鲁棒性;通过与基于重心法的传统位置检测方法对比,实验结果表明本文所提基于深度学习的抓取位置检测方法对于不同形态的抓取目标有较强的适应性。在机器人抓取实验中,机器人根据本文抓取位置检测方法的检测结果能够对目标物体进行抓取,并取得较高的抓取成功率。

在未来的研究工作中,本文方法的改进目标是进一步完善本文所提抓取位置检测方法的网络结构,提高机器人抓取的成功率以及模型的通用性,并将本方法应用到与机器人抓取任务相关的系统中。

参考文献

[1] Bezak P, Bozek P, Nikitin Y. Advanced robotic grasping system using deep learning[J]. *Procedia Engineering*, 2014, 96: 10-20

[2] Mahler J, Liang J, Niyaz S, et al. Dex-net 2.0: deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics[EB/OL]. https://arxiv.org/pdf/1703.09312.pdf: Cornell University, 2017

[3] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. *Science*, 2006, 313(5786): 504-507

[4] Maitin-Shepard J, Cusumano-Towner M, Lei J, et al. Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding[C]. In: Proceedings of the IEEE International Conference on Robotics and Automation, Anchorage, USA, 2010: 2308-2315

[5] Ramisa A, Alenya G, Moreno-Noguer F, et al. Using depth and appearance features for informed robot grasping of highly wrinkled clothes[C]. In: Proceedings of the IEEE International Conference on Robotics and Automation, Saint Paul, USA, 2012: 1703-1708

[6] Jiang Y, Moseson S, Saxena A. Efficient grasping from rgbd images: learning using a new rectangle representation[C]. In: Proceedings of the IEEE International Conference on Robotics and Automation, Shanghai, China, 2011. 3304-3311

[7] Lin Y, Sun Y. Robot grasp planning based on demonstrated grasp strategies[J]. *International Journal of Robotics Research*, 2015, 34(1): 26-42

[8] 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

[9] 仲训杲,徐敏,仲训昱,等. 基于多模特征深度学习的机器人抓取判别方法. 自动化学报,2016,42(7): 1022-1029

[10] Pinto L, Gupta A. Supersizing self-supervision: learning to grasp from 50k tries and 700 robot hours[C]. In: Proceedings of the IEEE International Conference on Robotics and Automation, Stockholm, Sweden, 2016. 3406-3413

[11] Zeng A, Song S, Nießner M, et al. 3DMatch: learning the matching of local 3D geometry in range scans[EB/OL]. https://arxiv.org/pdf/1603.08182.pdf: Cornell University, 2016

[12] Zeng A, Yu K T, Song S, et al. Multi-view self-supervised deep learning for 6D pose estimation in the amazon picking challenge [EB/OL]. https://arxiv.org/pdf/1609.09475.pdf: Cornell University, 2016

[13] Rifai S, Vincent P, Muller X, et al. Contractive auto-encoders: explicit invariance during feature extraction[C].

In: Proceedings of the 28th International Conference on Machine Learning, Bellevue, Washington, USA, 2011. 833-840

[14] Bengio Y, Lamblin P, Popovici D, et al. Greedy layer-wise training of deep networks[C]. In: Proceedings of Neural Information Processing Systems, Vancouver, Canada, 2007. 19-153

[15] Liu Y, Feng X, Zhou Z. Multimodal video classification with stacked contractive auto-encoders[J]. *Signal Processing*, 2016, 120: 761-766

A method for robotic grasping position detection based on deep learning

Yan Zhe*, Du Xuedan*,**, Cao Miao*,**, Cai Yinghao**, Lu Tao**, Wang Shuo**

(*School of Automation, Harbin University of Science and Technology, Harbin 150080)

(**The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190)

Abstract

The research on intellignet grasping of robots is conducted, and a method for detection of the robotic grasping position of an object based on deep learning is proposed. The method learns the optimal grasping position of an object by using the mode of combining unsupervised learning and supervised learning, with the multimodal features of a target object as the training data. In the unsupervised learning process, it uses a contractive auto encoder (CAE) to pre-train the neural network layer by layer, and then the whole network is fine-tuned by using the back propagation algorithm in the supervised learning process. The simulation results verify that the proposed method can give the accurate optimal grasping position of any object. The experiment on Universal Robot 5 shows that the grasping success rate is very high, indicating that the proposed method can be applied to robotic grasping.

Key words: deep learning, robotic grasping, position detection, contractive autoencoder (CAE)