

学术型硕士(打印时删除)



同濟大學  
TONGJI UNIVERSITY

硕士学位论文

基于 RGB-D 图像的三维物体识别算法  
的研究与实现

姓名：李勇奇

学号：1531620

所在院系：电子与信息工程学院

学科门类：工学

学科专业：控制科学与工程

指导教师：陈启军 教授

二〇一八年三月





同濟大學  
TONGJI UNIVERSITY

A dissertation submitted to  
Tongji University in conformity with the requirements for  
the degree of Master of Engineering

## **3D Object Recognition and Pose Estimation Based on RGB-D Images**

Candidate : Li Yongqi  
Student Number : 1531620  
School/Department : College of Electronics and  
Information Engineering  
Discipline : Engineering  
Major : Control Science and Engi-  
neering  
Supervisor : Prof. Chen Qijun

March, 2018



## 学位论文版权使用授权书

本人完全了解同济大学关于收集、保存、使用学位论文的规定，同意如下各项内容：按照学校要求提交学位论文的印刷本和电子版本；学校有权保存学位论文的印刷本和电子版，并采用影印、缩印、扫描、数字化或其它手段保存论文；学校有权提供目录检索以及提供本学位论文全文或者部分的阅览服务；学校有权按有关规定向国家有关部门或者机构送交论文的复印件和电子版；在不以赢利为目的的前提下，学校可以适当复制论文的部分或全部内容用于学术活动。

学位论文作者签名：

年   月   日



## 同济大学学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，进行研究工作所取得的成果。除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人创作的、已公开发表或者没有公开发表的作品的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名：

年      月      日



## 摘要

三维视觉是机器人感知的重要组成部分,但其目前的技术水平难以帮助机器人有效地感知周围的三维世界。随着近几年深度学习的发展,计算机视觉领域取得了巨大的发展,尤其是在 2D 视觉领域,2D 目标的检测和分类的准确率得到了巨大的提升,但 3D 目标的检测并没有巨大的提升。因此,本文针对机器人目前三维感知的困难,通过参考深度学习在 2D 视觉上的突破,将其引入到 3D 视觉上来,提出了 3D-MRAI 算法,用于解决对 3D 目标的检测以及位姿的估计。

深度信息的质量对 3D 目标检测以及位姿估计的准确率和精度都有至关重要的影响,因此为了获取高质量的深度信息,本文针对现有 RGB-D 相机的缺点,提出了对偶 RGB-D 相机结构,通过组合两个低价的 RGB-D 相机来获取高质量的深度信息,提高了相机获取深度图的填充率并且增强了深度信息的鲁棒性。

为了能够在 RGB-D 图中检测出目标物体的种类以及位姿,本文提出的 3D-MRAI 算法分为两步,第一步在相机拍摄的三维点云中分割出目标物体点云;第二步通过点云匹配算法求解出目标的位姿。为了分割出目标物体点云,本文基于 2D 目标检测中的 Faster R-CNN 和 Mask R-CNN 两个算法,提出了 3D Faster R-CNN 和 3D Mask R-CNN 算法,3D Faster R-CNN 和 3D Mask R-CNN 通过将深度图变换为 HHA 图,有效地利用三维信息,并结合 RGB 图完成对目标物体的检测,并且为了应对目标物体的各种姿态,算法还引入了 Spatial Transformer 结构。3D Faster R-CNN 和 3D Mask R-CNN 相比 Faster R-CNN 和 Mask R-CNN 充分利用三维信息,对检测一些纹理较少(Textureless)的物体有着更高的准确率。为了求解目标的位姿,本文通过匹配目标物体点云和目标物体 3D 模型来实现,为此基于 4PCS 算法提出了 A4PCS-ICP 点云匹配算法,通过在改进 4PCS 算法的基础上引入 ICP 算法提高了匹配精度。

本文还将所提出的 3D-MRAI 算法实际应用到 Bin-Picking 问题上,设计了一个基于 3D-MARI 算法的随机分拣视觉系统,所设计的系统在实验中达到了 100% 的抓取成功率,并且算法的运算时间也完全满足实际应用。

**关键词:** RGB-D, 目标检测, 位姿估计, 随机分拣



## ABSTRACT

3D vision is an important part of robot perception, but the technology of 3D vision currently is hard to help robots effectively perceive the surrounding 3D world. With the development of deep learning in recent years, tremendous development has been made in the field of computer vision. Especially in the field of 2D vision, the accuracy of 2D object detection and classification has been greatly improved, but the detection of 3D objects has not been huge Enhance. Therefore, for the difficulty of current 3D robot perception. This paper proposed a new algorithm 3D-MRAI, which introduced deep learning into 3D vision based on the breakthrough of deep learning in 2D vision. This algorithm is proposed to solve the problem of 3D object detection and pose estimation .

High-quality depth map has a great influence on the results of 3D object detection and pose estimation. To acquire high-quality depth map, a dual RGB-D camera structure is proposed to overcome the shortcomings of the existing RGB-D cameras. Dual RGB-D camera can obtain high-quality depth map by combining two low-cost RGB-D cameras, which also increases the fill rate of depth map and enhances depth value in depth map.

To detect object and estimate pose in a given RGB-D frame, 3D-MRAI algorithm proposed this paper runs in two stage. The first stage is to get the point cloud of the object from the RGB-D frame. The second stage is to estimate the object pose by point cloud matching. To segement the point cloud of the object, 3D Faster R-CNN and 3D Mask R-CNN are proposed based on two algorithm in 2D object detection: Faster R-CNN and Mask R-CNN. 3D Faster R-CNN and 3D Mask R-CNN take full advantage of depth value by converting depth map into HHA frame and detect object by combining RGB and HHA. 3D Faster R-CNN and 3D Mask R-CNN also use Spatial Transformer to detect object in arbitrary pose. 3D Faster R-CNN and 3D Mask R-CNN have higher detection accuracy when detecting textureless objects, comparing to Faster R-CNN and Mask R-CNN. In order to estimate the pose of the target, we match the 3D model of the object to the point cloud of the object. For this purpose, a new point cloud matching algorithm call A4PCS-ICP is proposed. A4PCS-ICP algorithm has higher matching accuracy by combining ICP and modified 4PCS.

This paper also applies the proposed 3D-MRAI algorithm to solve the Bin-Picking

problem. A bin-picking vision system is designed based on 3D-MARI algorithm. The designed system achieves 100% successful picking rate, and the system's response time also fully meets the application.

**Key Words:** RGB-D, object detection, pose estimation, Bin-Picking

# 目录

第 1 章 引言 .....	1
1.1 概述 .....	1
1.2 课题研究背景及现状分析 .....	1
1.3 研究内容与论文结构 .....	3
第 2 章 RGB-D 图像的获取与融合 .....	5
2.1 3D 相机现状与分析 .....	5
2.2 RGB-D 相机 .....	6
2.2.1 RGB-D 相机原理与结构 .....	6
2.2.2 RGB-D 相机的数学模型 .....	7
2.2.3 RGB-D 相机的标定流程 .....	9
2.3 对偶 RGB-D 相机 .....	12
2.3.1 对偶 RGB-D 相机原理与结构 .....	12
2.3.2 对偶 RGB-D 相机的标定流程 .....	17
2.4 深度图质量测试实验 .....	20
2.4.1 实验流程 .....	21
2.4.2 实验原理 .....	22
2.4.3 实验结果 .....	23
2.5 本章小结 .....	25
第 3 章 基于 RGB-D 图像的目标检测算法 .....	26
3.1 3D Faster R-CNN .....	26
3.1.1 Faster R-CNN .....	28
3.1.2 HHA .....	28
3.1.3 Spatial Transformer .....	30
3.2 3D Mask R-CNN .....	33
3.2.1 特征提取网络 .....	34
3.2.2 ROIAlign .....	35
3.2.3 Mask 损失函数 .....	35
3.3 目标检测实验 .....	36
3.3.1 数据集 .....	37
3.3.2 实验内容 .....	40
3.3.3 实验结果 .....	41
3.4 本章小结 .....	43

第 4 章 基于 4PCS 的点云匹配算法 .....	44
4.1 点云匹配算法概述 .....	44
4.1.1 问题描述 .....	44
4.1.2 背景介绍 .....	45
4.2 A4PCS-ICP 算法 .....	46
4.2.1 算法框架 .....	46
4.2.2 Angle-fixed 4PCS 算法 .....	46
4.2.3 Outlier filter .....	50
4.2.4 ICP 算法 .....	52
4.3 点云匹配实验 .....	54
4.3.1 实验内容 .....	54
4.3.2 实验结果 .....	54
4.4 本章小结 .....	55
第 5 章 3D 目标位姿估计算法 .....	57
5.1 3D-MRAI 框架设计 .....	57
5.2 3D-MRAI 具体实现 .....	58
5.3 3D 目标位姿估计实验 .....	60
5.3.1 数据集 .....	60
5.3.2 实验内容 .....	61
5.3.3 实验结果 .....	61
5.4 本章小结 .....	63
第 6 章 算法应用——Bin-Picking .....	65
6.1 Bin-Picking 背景与现状 .....	65
6.2 基于 3D-MRAI 的随机分拣系统 .....	67
6.2.1 系统硬件设计 .....	67
6.2.2 系统软件设计 .....	69
6.3 随机分拣实验 .....	76
6.3.1 实验内容 .....	76
6.3.2 实验结果 .....	77
6.4 本章小结 .....	78
第 7 章 结论与展望 .....	79
7.1 结论 .....	79
7.2 进一步工作的方向 .....	79
致谢 .....	81
参考文献 .....	82
附录 A 补充资料 .....	85
个人简历、在学期间发表的学术论文与研究成果 .....	86

# 第1章 引言

## 1.1 概述

随着互联网的发展、云计算和大数据等新兴技术的不断成熟，尤其是深度学习的发展，使得计算机视觉领域 2D 目标检测的准确率得到了空前的提高，但 3D 目标检测和位姿估计的研究还较少，其识别准确率还较低，实际运用还存在大量问题。

3D 目标检测和位姿估计在工业自动化、军事侦查及医疗各个领域有着大量需求，随着近几年高性价比 RGB-D 摄像头的出现，以及深度学习的浪潮，基于深度图像的 3D 目标检测的研究思路逐渐展现出优势。从 Google 的 Project Tango 到今年 iPhone X 的 Face ID 可以看到基于深度信息的三维识别和重建技术将会越来越流行。

本课题的主要研究目标是实现一套 3D 目标检测和位姿估计的视觉算，能够识别出所需目标以及位姿。其理论意义在于区别与传统 3D 目标检测和位姿估计的方法，基于 RGB-D 图像引入深度学习，与传统视觉算法相结合，极大地提高 3D 目标检测的准确率和位姿估计的精度，对工业自动化、医疗等领域有着巨大的实际应用价值。

## 1.2 课题研究背景及现状分析

3D 目标检测和位姿估计的任务是识别出图像中有什么类型的物体，并给出物体在相机坐标系下的位姿，是对三维世界的感知理解。3D 目标检测和位姿估计的结果可直接应用于实际机器人操作环境，但国内外 3D 目标检测和位姿估计的研究还相对较少。研究 3D 目标检测和位姿估计的问题一般可以分为一下几种研究思路：

- 基于模型 (model-based) 或几何 (geometry-based) 的方法；
- 基于外观 (appearance-based) 或视图 (view-based) 的方法；
- 基于局部特征匹配的方法；
- 基于深度图像的 3D 目标检测和位姿估计；

基于模型或几何的方法需要利用有关物体的先验知识,建立模型数据库,然后从输入图像数据中获取物体描述,并与模型库中描述进行匹配,该方法的优点是比较直观和易于理解,但缺点是算法运算量较大,而且在复杂环境下、物体遮挡、噪声干扰等情况下识别率往往很低。基于外观或视图的方法其实就是利用图像识别的方法来匹配目标,该方法对数据来源要求不高,在 3D 目标检测和位姿估计和图像检索系统中有广泛的应用,如 Swain 和 Ballard 等人提出一种全局特征匹配算法 (Swain et al. 1991),用颜色直方图来表示一个物体,通过直方图的匹配来识别物体。基于局部特征匹配的方法提取局部图像块的特征用于匹配,该方法通过对视角改变局部准不变过程,来检测得到视图中三维物体的局部区域,然后通过从局部测量计算得到的不变量描述的区域集合来表示物体,该方法的关键在于局部图像区域的选择和基于这些区域的特征计算,具体有 Lowe 等人提出的尺度不变特征变换 (SIFT) 描述子 (Lowe 1999)、Mikolajczyk 和 Schmid 等人提出的 Harris 角点检测器和 Hessian 点检测器 (Matas et al. 2004)、Chum 等人提出的最稳定极值区域检测器 (Chum et al. 2004) 等,该方法的一个缺点就是不存在对于各种场景类型和图像变换类型都是最优的检测器,各种检测器之间存在互补性,结合各种检测器虽然能提高该方法的性能,但算法的计算量太大。光学数字化处理的 3D 目标检测和位姿估计的研究目前还处在仿真模拟阶段,国内四川大学的苏显渝老湿采用结构光场、莫尔条纹和基于距离像位相编码等,国外有采用相移数字全系和整体图像 (Matoba et al. 2001) 进行 3D 目标检测和位姿估计,但这些研究大多停留在计算机模拟状态,很难在实际中应用。基于深度图像的 3D 目标检测和位姿估计,是通过距离传感器如激光、红外等获取传感焦平面到目标表面的距离,深度图像仅依赖于物体的几何形状,不存在色彩图像阴影或表面投影等问题,因此利用深度图像进行三维物体的识别有一定的优势,并且由于近几年距离传感器越来越精确,基于深度图像的 3D 目标检测和位姿估计的研究逐渐多了起来。

随着近几年高性价比 RGB-D 摄像头的出现及计算机性能的提高,国外出现了基于 RGB-D 图像的 3D 目标检测和位姿估计的研究方法,RGB-D 图像结合了物体形状、颜色以及深度信息,相比单个 RGB 图像或者单个深度图像有着无法比拟的优势,但相对地,信息的处理量和复杂度也相应地增加了,但深度学习的发展给这种研究方法带来了空前的机遇。Saurabh 等人基于 RGB-D 图像提取了多种特征用于 CNN 网络的学习,加上 SVM 分类器大大提升了 3D 目标检测和位姿估计的准确率 (Gupta et al. 2014);在此基础之上, Saurabh 又引入了基于三维模型的方法,通过匹配已有的三维物体模型库,进一步提高了识别的准确率,并且识别速度也进一步增快了 (Gupta et al. 2013); Alexandre 等人将 RGB-D 四个通道分别输入卷积神经网络,并在各个通道之间使用转移学习的方法训练模型,该方法不经

提高了 3D 目标检测和位姿估计的准确率,还减少了模型训练的时间 (Alexandre 2016); Rico 和 Clemens 等人考虑具体机器人操作环境,从 RGB-D 图像中提取了六种特征,采用统计学习的方法,实现了多三维物体分类和识别 (Jonschkowski et al. 2016),并在 APC(Amazon Picking Challenge) 上取得了冠军,可以说是将基于 RGB-D 图像和学习的方法具体应用到实际机器人操作环境中的典范,是目前 3D 目标检测和位姿估计领域内的顶尖水平。

随着互联网的发展、云计算和大数据等新兴技术的不断成熟,尤其是深度学习的引入,计算机视觉领域得到了空前的发展,对于二维图像中目标检测的准确率大大地提高了。但是二维目标识别给出的结果是框住可能目标的矩形框,其中部分像素并不是目标本身,而且缺少目标的姿态、尺寸等信息,所以其结果往往难以应用于实际机器人操作环境中。实际机器人操作环境需要目标准确的位置、姿态以及尺寸等信息,机器人才能采取合适的操作。因此,三维物体的识别对于实际机器人操作具有十分重要的意义和价值。相对于二维图像中目标的检测与识别,3D 目标检测和位姿估计的相关研究,尤其是基于 RGB-D 图像的研究还相当少,因此对 3D 目标检测和位姿估计的研究十分有必要。基于 RGB-D 图像的 3D 目标检测和位姿估计的研究,引入机器学习中相关的学习方法如深度学习、支持向量机以及统计学习方法等,对 3D 目标检测和位姿估计的准确率提升、促进该问题的进一步解决有着巨大的帮助和十分重要的现实意义,有助于推动机器视觉的发展。

综上,3D 目标检测和位姿估计的研究对机器人具体应用有着十分重要的意义,是机器人乃至整个人工智能领域发展的重要组成部分,然而国内外对 3D 目标检测和位姿估计的研究还相对较少,尤其是基于 RGB-D 图像和机器学习算法的 3D 目标检测和位姿估计的研究。机器学习算法,尤其是深度学习的引入对 3D 目标检测和位姿估计的准确率的提高有着不可估量的进步空间,因此,本文基于 RGB-D 图像和机器学习算法的 3D 目标检测和位姿估计的研究十分有必要,对机器视觉领域有着十分重要的价值,对于推动该领域的进一步发展具有重大意义。

### 1.3 研究内容与论文结构

本文研究的内容是基于 RGB-D 图像的三维物体的识别,通过输入的 RGB-D 图像,给出物体的种类以及在相机坐标系下的位姿。旨在提出一种能实际应用到机器人的 3D 目标检测和位姿估计算法,基于深度学习技术和传统视觉算法以解决机器人的三维感知问题。本文的结构安排如下:

第一章即为本章,介绍 3D 视觉目标检测和位姿估计的研究意义、现状调研

与本文的研究对象和主要内容框架。

第二章主要介绍了深度信息获取的现状以及所采用的 RGB-D 相机，并针对现有 RGB-D 相机存在的问题，提出了对偶 RGB-D 相机结构。

第三章在 Faster R-CNN 和 Mask R-CNN 的基础上，通过引入 HHA 和 Spatial Transformer Network，提出了 3D Faster R-CNN 和 3D Mask R-CNN 算法，用于解决纹理缺少物体的检测。

第四章介绍了基于 4PCS 算法和 ICP 算法提出的 A4PCS-ICP 算法，A4PCS-ICP 算法通过结合 4PCS 和 ICP 算法提高了点云匹配的精度。

第五章在第二章和第三章的基础上，通过将 3D Faster/Mask R-CNN 和 A4PCS-ICP 算法结合，提出了 3D-MRAI 算法，该算法可以在输入 RGB-D 图中检测出物体的种类和位姿。

第六章介绍了所提出的三维物体目标检测和位姿估计算法 3D-MRAI 的具体应用——Bin-Picking，并通过实验评价了所设计的随机分拣视觉系统的性能。

第七章为本文的总结与进一步展望。

## 第2章 RGB-D 图像的获取与融合

RGB-D 图像是所设计的三维物体识别算法的输入, 其质量对算法结果有着至关重要的影响, 以此获取高质量的 RGB-D 图像也十分重要。本章首先分析了 3D 相机的现状, 然后详细介绍了选取的 RGB-D 相机的原理和标定流程, 并且针对其缺点提出了对偶 RGB-D 相机结构, 最后通过实验证明了对偶 RGB-D 相机可以获取更高质量的 RGB-D 图像。

### 2.1 3D 相机现状与分析

3D 相机能够获取相机到物体表面每一点的距离, 从而感知物体的形状和距离, 近几年 3D 成像技术的应用越来越多, 如用于体感游戏的 Kinect(Microsoft 2012), Google 的 Project Tango(Google 2012), 以及 Apple 公司的 iPhone X 前置摄像头的人脸识别。

目前市面上的 3D 相机要么价格昂贵, 要么精度低下, 很难找到一款性价比较高的 3D 相机。如表2.1和图2.1所示, 列举了一些市面上较为常见的 3D 相机,

品牌	价格	精度	速度
SICK	大于 30 万	高	很慢
Enshape	大于 30 万	高	较快
Ensenso	约 10 万	较高	较快
Realsense	约 1.5 千	中等	快
Kinect V2	约 1.5 千	低	快

表 2.1 市面上主要 3D 相机价格和性能

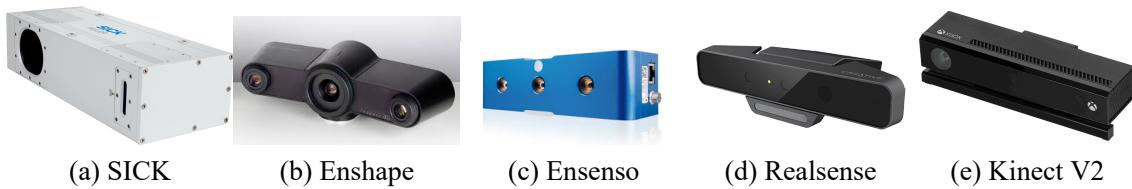


图 2.1 市面上主要 3D 相机

从表中可以发现很多精度高的 3D 相机价格及其昂贵, 并且采集速度很慢, 价格低的 3D 相机精度有相对较低。出于实际应用考虑, 我们需要相机的采集速度相

对较快,再加上成本上的限制,最终选择了精度中等、采集速度较快、价格较为便宜的 Realsense 系列相机。

## 2.2 RGB-D 相机

### 2.2.1 RGB-D 相机原理与结构

RGB-D 相机获取深度的原理大致可以分为三种:

- Structure Light
- Time of Flight(ToF)
- Stereo

Structure Light 获取深度信息的原理是通过激光发射器投射带有特定编码的结构光到物体表面后,由 IR Camera 采集,根据采集到的光信号量的变化来计算物体的深度。举一个形象的例子,将手电筒照向墙面,手电筒离墙面越远,墙面上所形成的光斑的直径就越大,所以可以通过光斑的直径来计算手电筒距离墙面的距离。ToF 获取深度信息的原理是通过专有的传感器捕捉红外光发射到接收的飞行时间来计算物体的深度。Stereo 是通过双摄像头拍摄物体,再通过特征点匹配,根据三角测量原理来计算物体的深度。

三种原理的深度相机各有其特点,采用 Structure Light 原理的深度相机一般精度比较高,但景深比较短并且受光线影响比较大,适合室内场景;ToF 原理的深度相机获取深度图的精度和分辨率一般都比较低,但帧率高,并且具有一定的抗光照性能;Stereo 获取深度精度适中,帧率相对来说较低,并且需要较强的计算性能,但抗光照能力强,适合室外场景。

本文所使用的 RGB-D 相机是 Intel 的 Realsense SR300 相机,SR300 采用的结构光的原理获取深度<sup>①</sup>,其内部结构如图2.2所示。从图2.2可以看出,SR300 内部的传感器主要有彩色摄像头(Color Camera)、红外激光发射器(Infrared Laser Projector)和红外摄像头(Infrared Camera)。Color Camera 是  $1920 \times 1080$  像素的普通针孔摄像头,用来获取彩色图像;Infrared Laser Projector 和 Infrared Camera 用来获取深度图像或者红外成像图,两种成像流程如图2.3所示。其中当 Infrared Laser Projector 投射带有编码的结构光时,Infrared Camera 可以获取深度图;当投射不带编码的红外光时,Infrared Camera 可以获取红外成像图。正常使用时,往往设置 Infrared Laser Projector 投射带有编码的结构光来获取深度信息。因此,从 RGB-D 相机的使用来看,可以忽略其内部具体结构,将其看成由一个彩色摄像头

---

<sup>①</sup> 此后所提到的 RGB-D 相机均指与 SR300 相机类似的采用结构光原理获取深度的 RGB-D 相机

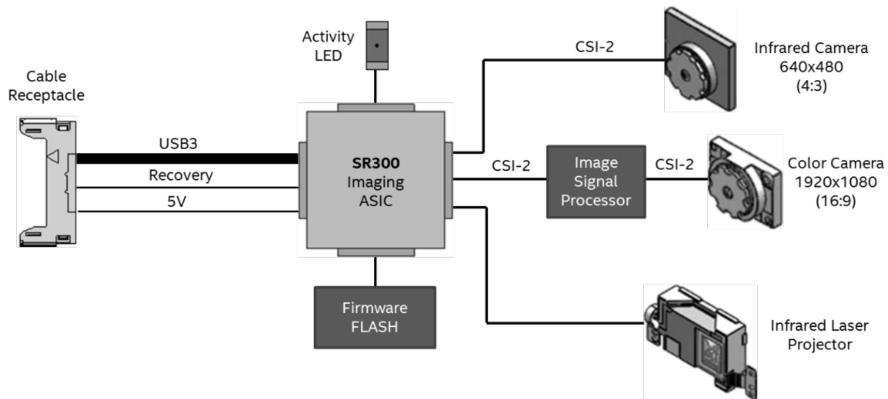


图 2.2 Realsense SR300 内部结构图

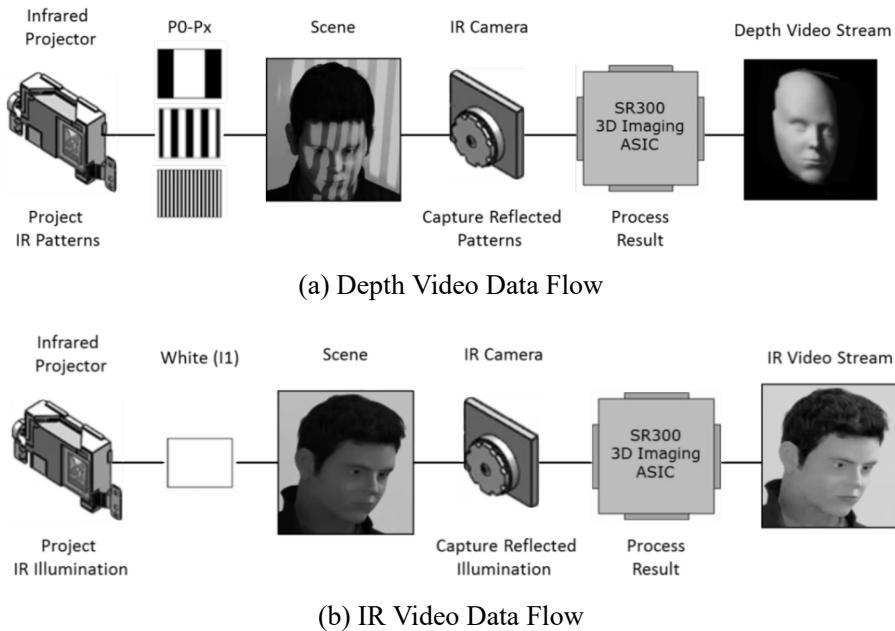


图 2.3 Realsense SR300 深度成像流程

和一个深度摄像头构成,其中彩色摄像获取彩色(RGB)信息,深度摄像头获取深度(depth)信息。

## 2.2.2 RGB-D 相机的数学模型

图2.4展示了本文所使用的RGB-D相机的基本物理模型,其中彩色摄像头和深度摄像头都使用了针孔(pin-hole)相机模型(Heikkilä 2000)。先考虑普通针孔相机的模型,相机图像坐标系下一点  $\mathbf{u} := [u, v]^T$ ,对应的三维世界中的一点在相机坐标系下表示为  $\mathbf{X} := [x, y, z]^T$ 。根据针孔相机模型有:

$$z\tilde{\mathbf{u}} = \mathbf{K}\mathbf{X} \quad (2.1)$$

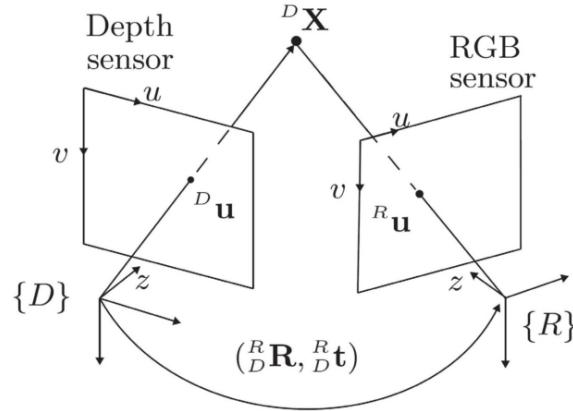


图 2.4 RGB-D 相机模型

其中  $\tilde{\mathbf{u}}$  表示  $\mathbf{u}$  的齐次变换形式, 彩色相机的内参矩阵  $\mathbf{K}$  的定义如下:

$$\mathbf{K} := \begin{bmatrix} f_u & 0 & u_0 \\ 0 & f_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \quad (2.2)$$

其中  $f_u$  和  $f_v$  分别表示彩色相机在图像坐标轴上的焦距(以像素为单位),  $u_0$  和  $v_0$  表示彩色相机光心在图像平面的投影中心。

公式2.1还未考虑镜头的畸变,为了提高相机的精度,现引入径向畸变(radial distortion)和切向畸变(tangential distortion):

- 径向畸变是由相机透镜的不完善和表面曲率存在误差造成的, 径向畸变的数学模型可以表示为:

$$\begin{cases} \hat{x} = \bar{x}(1 + k_1r^2 + k_2r^4 + k_3r^6) \\ \hat{y} = \bar{y}(1 + k_1r^2 + k_2r^4 + k_3r^6) \end{cases} \quad (2.3)$$

其中

$$\bar{x} = x/z \quad (2.4)$$

$$\bar{y} = y/z \quad (2.5)$$

$$r = \sqrt{\bar{x}^2 + \bar{y}^2} \quad (2.6)$$

$\bar{x}, \bar{y}$  表示点  $X$  在归一化平面上的坐标,  $\hat{x}, \hat{y}$  表示修正径向畸变后的的坐标,  $k_1, k_2, k_3$  表示径向畸变的参数。

- 切向畸变是由于相机透镜与图像平面不平行造成的, 其数字模型可以表示为:

$$\begin{cases} \hat{x} = \bar{x} + (2p_1\bar{x}\bar{y} + p_2(r^2 + 2\bar{x}^2)) \\ \hat{y} = \bar{y} + (p_1(r^2 + 2\bar{y}^2) + 2p_2\bar{x}\bar{y}) \end{cases} \quad (2.7)$$

其中  $p_1, p_2$  是切向畸变的参数。

- 结合公式2.3和2.7可以得到修正径向畸变和切向畸变的 Brown-Conrady 模型 (Brown 1966):

$$\begin{cases} \hat{x} = \bar{x}(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) + (2p_1 \bar{x}\bar{y} + p_2(r^2 + 2\bar{x}^2)) \\ \hat{y} = \bar{y}(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) + (p_1(r^2 + 2\bar{y}^2) + 2p_2 \bar{x}\bar{y}) \end{cases} \quad (2.8)$$

通过以上分析,根据公式2.1和2.8可以推导出带有畸变的针孔相机模型:

$$\begin{cases} u = f_u(\bar{x}(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) + (2p_1 \bar{x}\bar{y} + p_2(r^2 + 2\bar{x}^2))) + u_0 \\ v = f_v(\bar{y}(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) + (p_1(r^2 + 2\bar{y}^2) + 2p_2 \bar{x}\bar{y})) + v_0 \end{cases} \quad (2.9)$$

为方便起见,记  $\mathbf{d} := [k_1, k_2, p_1, p_2, k_3]^T$ , 定义函数

$$f_{undist}(\mathbf{d}, \mathbf{X}) := \begin{bmatrix} \bar{x}(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) + (2p_1 \bar{x}\bar{y} + p_2(r^2 + 2\bar{x}^2)) \\ \bar{y}(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) + (p_1(r^2 + 2\bar{y}^2) + 2p_2 \bar{x}\bar{y}) \end{bmatrix} \quad (2.10)$$

$$\tilde{f}_{undist}(\mathbf{d}, \mathbf{X}) := \begin{bmatrix} f_{undist}(\mathbf{d}, \mathbf{X}) \\ 1 \end{bmatrix} \quad (2.11)$$

则公式2.9可简化为:

$$\tilde{\mathbf{u}} = \mathbf{K} \cdot \tilde{f}_{undist}(\mathbf{d}, \mathbf{X}) \quad (2.12)$$

其中需要标定的参数有相机内参矩阵  $\mathbf{K}$ (包含未知参数  $f_u, f_v, u_0, v_0$ )以及畸变参数  $\mathbf{d}$ (包含未知参数  $k_1, k_2, p_1, p_2, k_3$ ),共 9 个参数。

明确了针孔相机的数学模型后,很容易推出 SR300 的相机模型:

$$\begin{cases} {}^R\tilde{\mathbf{u}} = {}^R\mathbf{K} \cdot \tilde{f}_{undist}({}^R\mathbf{d}, {}^R\mathbf{X}) \\ {}^D\tilde{\mathbf{u}} = {}^D\mathbf{K} \cdot \tilde{f}_{undist}({}^D\mathbf{d}, {}^D\mathbf{X}) \\ {}^R\mathbf{X} = {}_D^R\mathbf{R} {}^D\mathbf{X} + {}_D^R\mathbf{t} \end{cases} \quad (2.13)$$

其中左上标  $\{R\}$  表示 SR300 相机中的彩色相机(RGB), $\{D\}$  表示 SR300 相机中的深度相机(Depth), ${}_D^R\mathbf{R}$  和  ${}_D^R\mathbf{t}$  表示了彩色相机坐标系和深度相机坐标系之间的齐次变换关系。

### 2.2.3 RGB-D 相机的标定流程

根据上文所述的 RGB-D 相机的结构及数学模型,RGB-D 相机的标定主要涉及到彩色摄像头内参和畸变的标定,深度摄像头内参和畸变的标定,以及彩色摄像头和深度摄像头之间位姿变换的标定。由于 RGB-D 相机是一种较为新颖的相

机, 所以市面上基本上没有较为成熟通用的标定 RGB-D 相机的方法以及对应的工具。因此本文针对所使用的 Realsense SR300 相机, 设计了一套标定方法。

根据公式2.13可知相机需要标定的参数有彩色相机内参和畸变参数 9 个, 深度相机内参和畸变参数 9 个, 彩色相机和深度相机之间的位姿关系 6 个, 一共 24 个参数。一起标定这 24 个参数理论上是相当困难的, 考虑到普通针孔相机的标定技术已经相当成熟(如张正友的棋盘格标定 (Zhang 2002), 以及 RGB-D 相机中彩色相机和深度相机的解耦性, 因此所设计的标定方法分为三步:

Step 1 标定彩色相机内参以及畸变参数

Step 2 标定深度相机内参以及畸变参数

Step 3 标定彩色相机和深度相机之间的齐次变换关系

步骤 1 标定彩色相机内参以及畸变参数相对来说比较简单, 主要参考文献 (Zhang 2002), 但所使用的标定板是不对称圆盘标定板 (Asymmetrical Circle Board), 如图2.5是  $4 \times 11$  的不对称圆盘标定板。使用圆盘标定板而非棋盘格标定

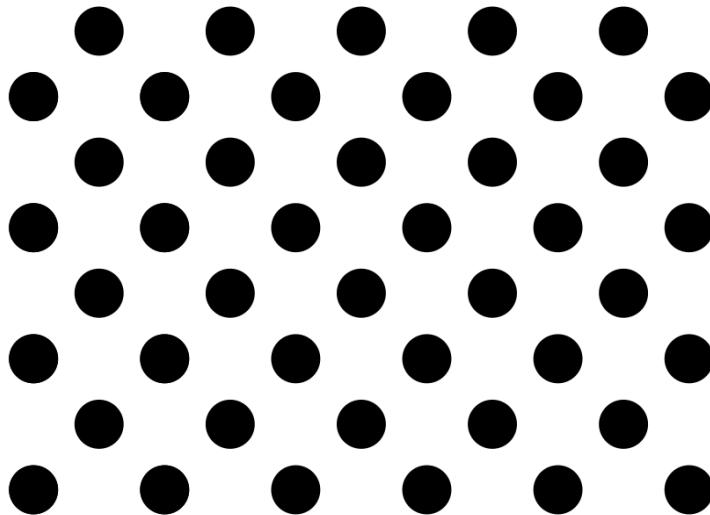


图 2.5 Asymmetrical Circle Board

板的原因是圆盘相对于棋盘格有更高的检测精度, 在某些情况下可以达到 0.1 到 0.01 像素的亚像素精度, 当然代价是相比计算棋盘格的角点, 计算椭圆(圆形经过投影变换后退化为椭圆)的中心会涉及到较为复杂的数学运算, 这也是为什么工业上大多使用圆盘作为标定板的原因。

步骤 2 标定深度相机内参以及畸变参数的方法和步骤 1 类似, 区别在于深度相机并不能直接获得颜色信息, 因此也不能直接检测图2.5所示的标定板。但是, 幸运的是, 根据前文所述的 SR300 深度相机的原理, 其本质上也是个普通的针孔相机, 只不过在其镜头上加上了滤波片, 可以认为其只对红外光成像。因此, 只要

使用图2.3中的红外成像模式获取红外成像图,在红外成像图上检测标定板。如图2.6所示,在红外成像图中检测出了标定板。

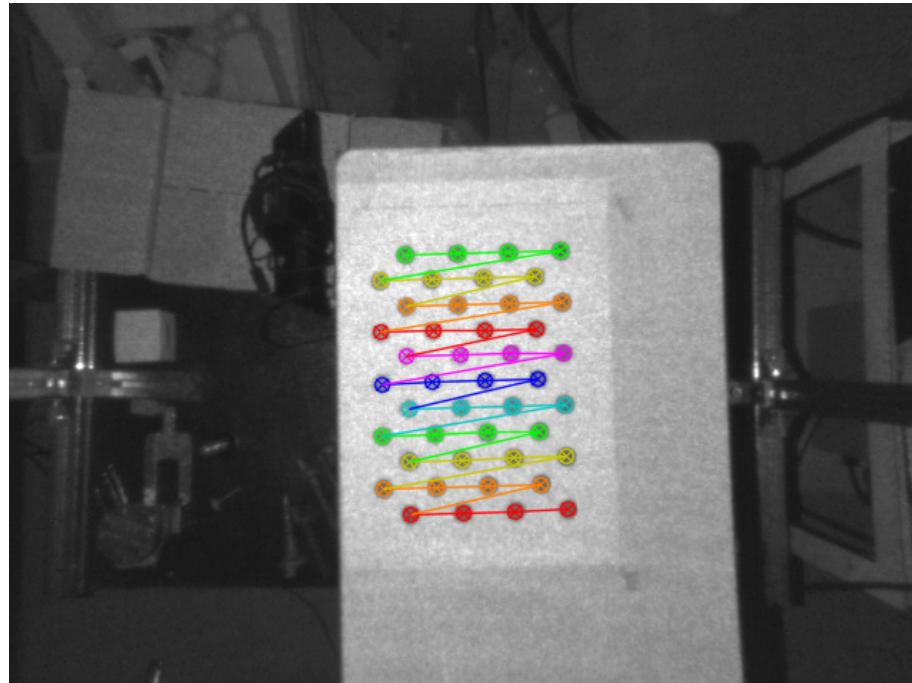


图 2.6 红外成像图中检测标定板

步骤3标定彩色相机和深度相机之间的齐次变换关系需要依赖于步骤1和步骤2中标定出的彩色相机和深度相机的内参和畸变参数,具体做法是将标定板放在彩色相机和深度相机下,使彩色相机和深度相机能够同时检测到标定板,然后分别根据各自的内参和畸变参数计算出标定板的位姿 ${}^R_B\mathbf{H}$ 和 ${}^D_B\mathbf{H}$ ,其中 ${}^R_B\mathbf{H}$ 是 $4 \times 4$ 的齐次变换矩阵,表示标定板在彩色相机坐标系下的位姿,也是彩色相机坐标系变换到标定板坐标系的齐次变换矩阵; ${}^D_B\mathbf{H}$ 也是 $4 \times 4$ 的齐次变换矩阵,表示标定板在深度相机坐标系下的位姿,也是深度相机坐标系变换到标定板坐标系的齐次变换矩阵。从而所要求的彩色相机坐标系变换到深度相机坐标系的齐次变换矩阵为:

$${}^D_R\mathbf{H} = {}^R_B\mathbf{H} {}^D_B\mathbf{H}^{-1} \quad (2.14)$$

其中

$${}^R_D\mathbf{H} := \begin{bmatrix} {}^R_R & {}^R_t \\ {}^D_R & {}^D_t \\ \mathbf{0}_{1 \times 3} & 1 \end{bmatrix} \quad (2.15)$$

当然,实际标定时,往往采取多组 ${}^R_B\mathbf{H}$ 和 ${}^D_B\mathbf{H}$ 来提高标定的精度。

### 2.3 对偶 RGB-D 相机

使用 SR300 相机时，发现相机在某些情况下，对一些反光的物体的深度图有严重的缺失，具体如图2.7所示。经过实验，发现这种缺失情况的出现和拍摄的

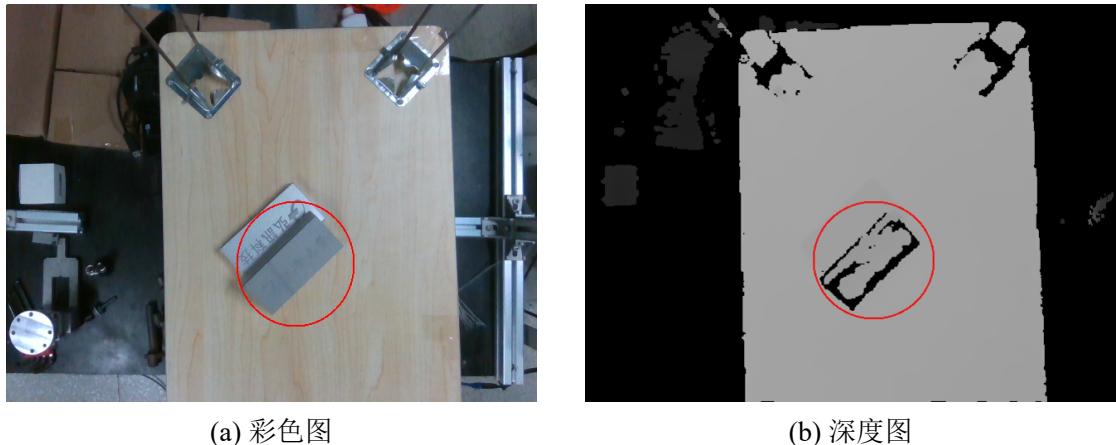


图 2.7 SR300 采集的物体深度信息部分缺失情况下的深度图

角度以及光线有关，因此本文提出一种组合相机对偶 RGB-D 相机 (Dual RGB-D Camera)。

### 2.3.1 对偶 RGB-D 相机原理与结构

对偶 RGB-D 相机在原 RGB-D 相机的基础上,通过增加一个与原相机呈 180 度夹角的 RGB-D 相机构成,实际物理结构如图2.8所示。

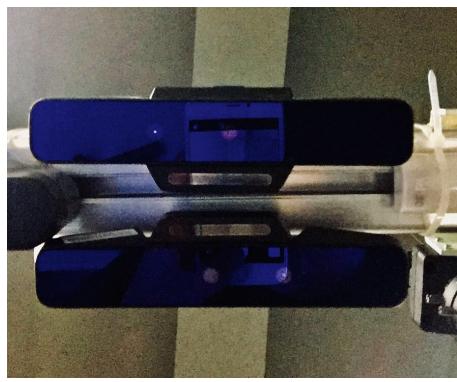


图 2.8 对偶 RGB-D 相机实际物理结构

对于对偶 RGB-D 相机,当其中一个相机深度图出现严重缺失时,另外一个相机的深度图往往不会在相同的地方深度信息出现严重的缺失,如图2.9所示<sup>②</sup>,有

② 实际下相机采集的图像与上相机采集的图像相差了 180 度,为了方便起见,都将下相机采集的图像旋转了 180 度

效的避免了单个 RGB-D 相机某些情况下深度信息严重缺失的情况。

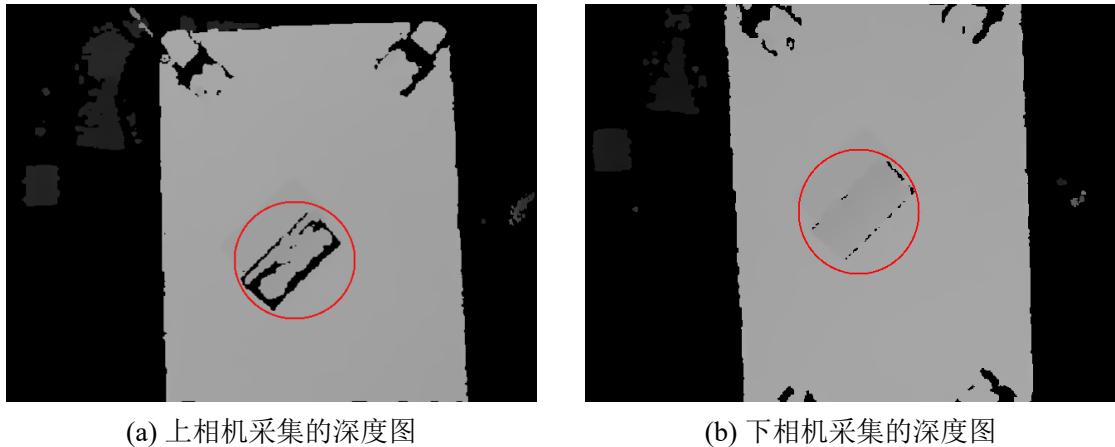


图 2.9 对偶 RGB-D 相机采集的左右两张深度图

除此之外,对偶 RGB-D 相机还可以利用两个相机的彩色图构成双目,生成第三张深度图,从而通过设计的深度的融合算法将三张深度图融合成为一张质量更高的深度图,其内部原理如图2.10所示。

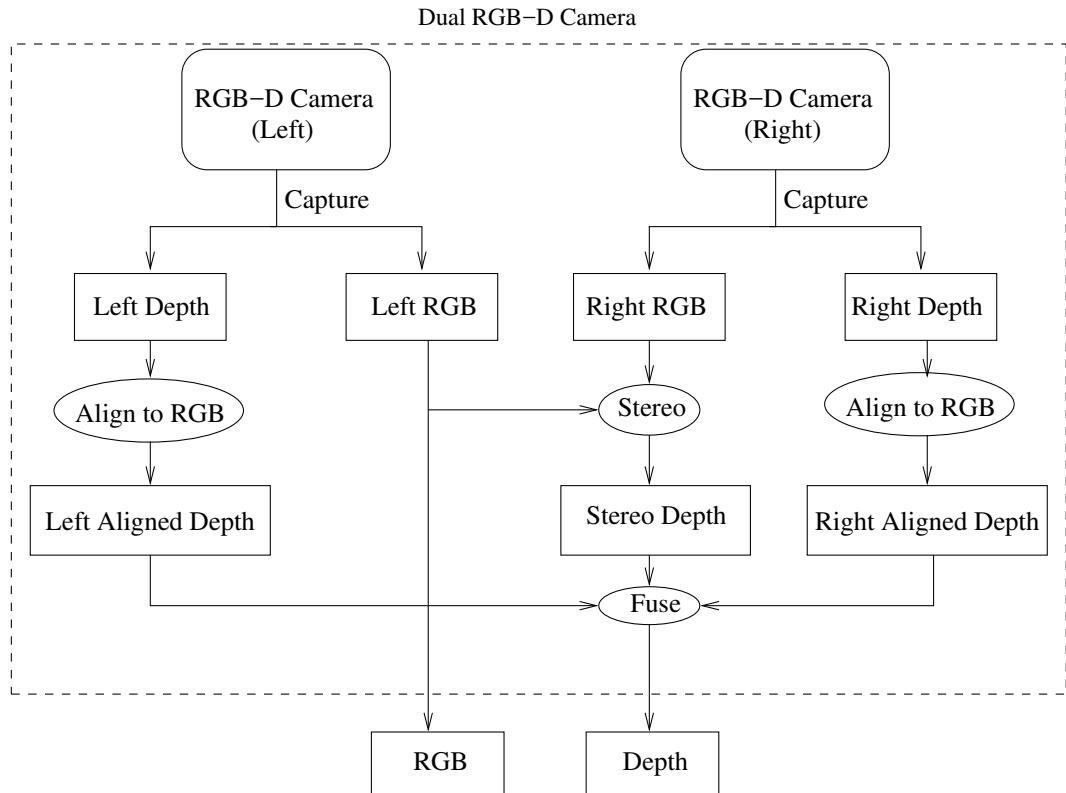


图 2.10 对偶 RGB-D 相机内部原理图

从外部使用来看,对偶 RGB-D 相机也输出一张彩色图、一张深度图。输出的彩色图就是从上相机采集到的彩色图;输出的深度图是由三张深度图融合而成,

并且与输出的彩色图相对齐,对齐的意思是彩色图和深度图相同图像坐标下的颜色信息和深度信息对应的实际物理世界中相同的一点,对齐的意义在于方便后续的一些图像处理的算法。

从内部实现来看,主要涉及到三个部分:

- 将深度图与输出的彩色图对齐 (Align to RGB)
- 利用上相机采集的彩色图和下相机采集的彩色图,通过双目匹配算法形成一张新的深度图
- 融合上相机对齐后的深度图、下相机对齐后的深度图和双目匹配得到的深度图

将深度图与彩色图对齐,相对来讲实现还是比较简单的,对齐深度图的具体流程如算法1所示。算法1主要将深度图中每个点的图像坐标利用该点的深度信息反

---

#### 算法 1: Align Depth Frame

---

**Input:** Raw Depth Frame  $Raw\_D_{dh \times dw}$

**Output:** Aligned Depth Frame  $Aligned\_D_{ch \times cw}$

```

1 for  $p$  in  $Aligned\_D$  do
2    $p = 0$ 
3 for  $dy = 1; dy <= dh; ++dy$  do
4   for  $dx = 1; dx <= dw; ++dx$  do
5     通过深度相机内参将点  $(dx, dy)$  反投影到三维空间一点  ${}^D X$ ;
6     坐标变换  ${}^R X = {}_D^R R {}^D X + {}_D^R t$ ;
7     通过彩色相机内参将点  ${}^R X$  投影变换到彩色图像坐标系下一点
8      $(cx, cy)$ ;
9     if  $cx$  in  $(0, cw]$  and  $cy$  in  $(0, ch]$  then
10     $Aligned\_D(cx, cy) = Raw\_D(dx, dy);$ 

```

---

投影变换到实际三维空间中一点,然后将该点坐标变换到彩色相机坐标系下,最后通过彩色相机的内参将该点在彩色相机坐标系下的三维坐标投影变换到彩色图像上的二维坐标。实际对齐三张深度图时,对于上相机深度图对齐到上相机彩色图,需要分别知道上相机深度相机和彩色相机的内参和畸变参数以及深度相机与彩色相机之间的齐次变换关系(通过相机标定这些参数都可以得到);双目匹配得到的深度图理论上可以有两张,一张与上相机校准后的彩色图像对齐,另一张与下相机校准后的彩色图像对齐,简单起见,选择与上相机对齐的深度图,然后通过上相机校准所使用的旋转矩阵的逆矩阵即可得到与原上相机彩色图像对齐的

深度图；对齐下相机到上相机彩色图，除了要知道下相机标定的参数外，还需要知道下相机与上相机之间的齐次变换关系（通过对偶 RGB-D 相机的标定得到）。

利用上下相机采集到的两张彩色图获取深度信息主要分为三步：

- 分别对两张原始图像进行校准
- 在校准后的两张图像上通过匹配算法得到视差图
- 通过视差图获取深度图

对两张原始图像进行校准主要通过双目相机的标定实现，使得校准后的两张图像的极线对齐，如图2.11所示，其中绿色的直线便是图像对齐后的部分极线，可以看

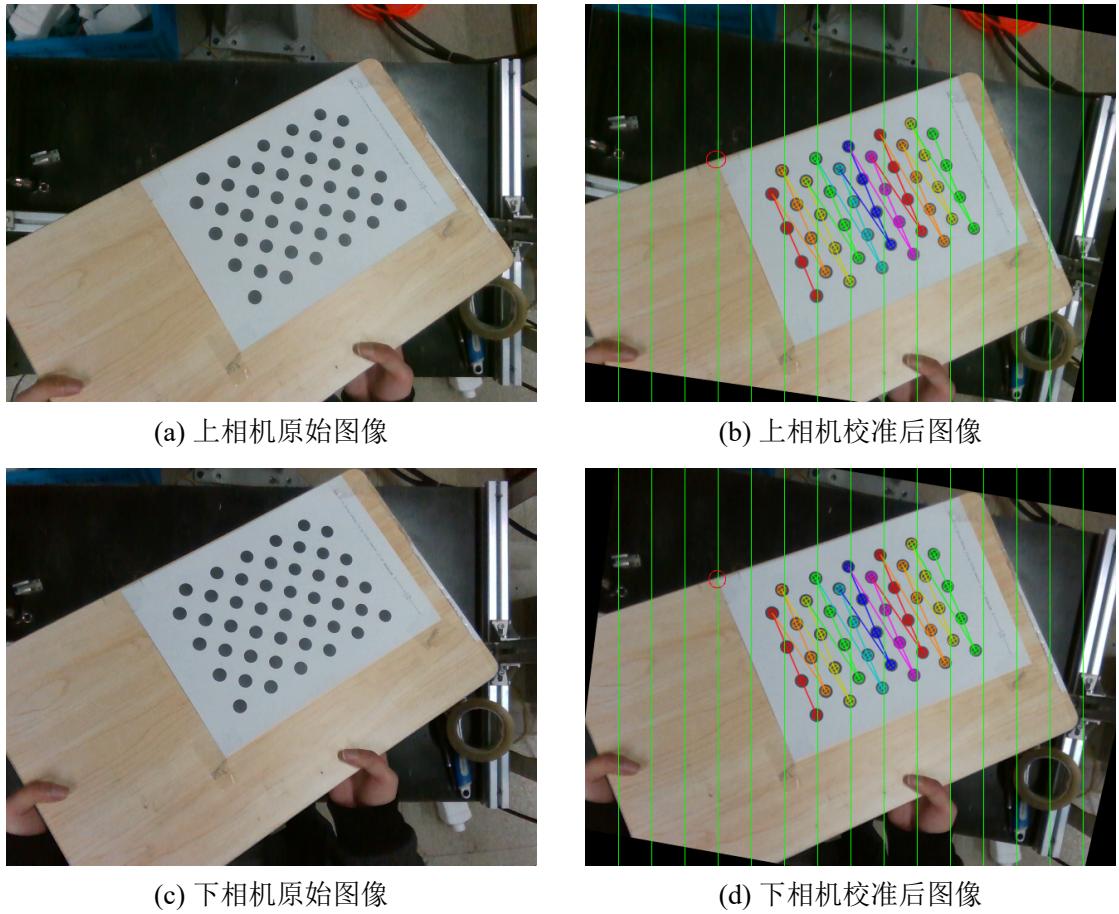


图 2.11 双目相机原始图像和校准后图像

出校准后的图像的对应点都分布在对齐的极线上（如图中用红色圈出的一对对应点所示），这样使得双目的匹配算法的搜索从二维缩小到了一维，只需要在极线上找对应点即可，能更快更稳定地在两张图中找到对应点。双目匹配算法使用的是 ELSA 算法 (Geiger et al. 2010)，通过 ELSA 算法可以从两张校准后彩色图像上得到对应的视差图，视差图到深度图的变化可以通过公式2.16得到：

$$z = \frac{\{T, R\} f B}{-(\{T, R\} v_0 - \{B, R\} v_0) + \{T, R\} d} \quad (2.16)$$

其中上标  $\{T, R\}$  (Top,RGB) 表示上相机的 RGB 摄像头,  $\{B, R\}$  (Bottom,RGB) 表示下相机的 RGB 摄像头,  $B$  表示基线长度,  ${}^{\{T, R\}}d$  表示视差。一般地, 会人为地校准过程中使得  ${}^{\{T, R\}}v_0 - {}^{\{B, R\}}v_0 = 0$ , 从而公式2.16可以简化为:

$$z = \frac{{}^{\{T, R\}}fB}{{}^{\{T, R\}}d} \quad (2.17)$$

融合上相机对齐后的深度图、下相机对齐后的深度图以及双目匹配得到的深度这三张深度图的算法首先做的是分别对这三张深度图进行预处理, 填补一些深度缺失的像素, 因为对齐后的深度图和双目匹配得到的深度图深度信息都有细微的缺失, 填补深度信息缺失的方法如算法2所示。算法2主要实现对于深度缺失的

---

### 算法 2: Fill Holes in Depth Frame

---

**Input:** Depth Frame  $D_{h \times w}$

**Output:** Filled Depth Frame  $FD_{h \times w}$

```

1 for  $y = 1; y <= h; ++y$  do
2   for  $x = 1; x <= w; ++x$  do
3     if  $valid(D_{x,y})$  then
4        $FD_{x,y} = D_{x,y};$ 
5     else
6        $FD_{x,y} = \text{NAN};$ 
7       bool leftTop =  $valid(D_{x-1,y-1})$  or  $valid(D_{x,y-1})$  or  $valid(D_{x-1,y})$ ;
8       bool leftBottom =  $valid(D_{x-1,y+1})$  or  $valid(D_{x,y+1})$  or  $valid(D_{x-1,y})$ ;
9       bool rightTop =  $valid(D_{x+1,y-1})$  or  $valid(D_{x,y-1})$  or  $valid(D_{x+1,y})$ ;
10      bool rightBottom =  $valid(D_{x+1,y+1})$  or  $valid(D_{x,y+1})$  or  $valid(D_{x+1,y})$ ;
11      if  $leftTop$  and  $leftBottom$  and  $rightTop$  and  $rightBottom$  then
12        validPoints = {};
13        for  $dy = -1; dy <= 1; ++dy$  do
14          for  $dx = -1; dx <= 1; ++dx$  do
15            if  $valid(D_{dx,dy})$  then
16              push back  $D_{x,y}$  to validPoints;
17            if  $\max(validPoints) - \min(validPoints) < 0.05$  then
18               $FD_{x,y} = \text{mean}(validPoints);$ 

```

---

点, 将检查其周围的深度信息, 当其四个角上都有有效的深度信息时, 并且周围有

效深度信息的极值小于一定阈值时,会用周围有效深度信息的均值填充该缺失的点。实际的效果如图2.12所示。分别对深度图进行预处理后,将会对三张深度图

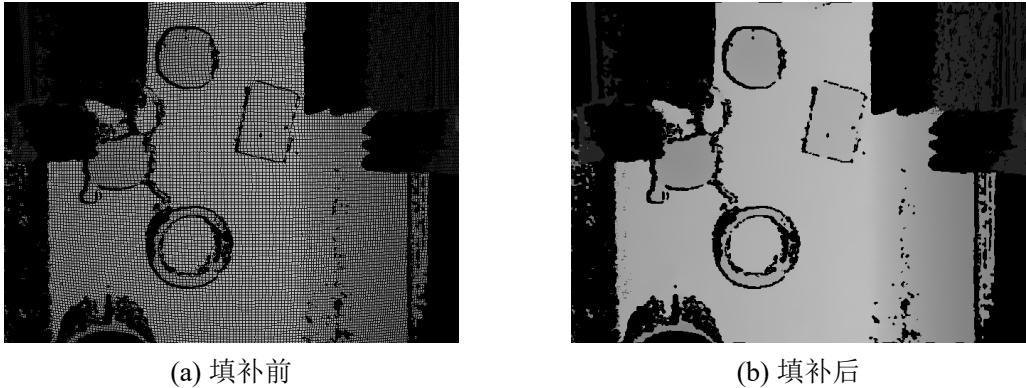


图 2.12 填补深度信息缺失算法效果图

进行线性叠加得到最终的深度图,基本叠加的公式如2.18所示。

$$d_{fuse} = \frac{w_1 d_{left} + w_2 d_{right} + w_3 d_{stereo}}{w_1 + w_2 + w_3} \quad (2.18)$$

其中  $w_1, w_2, w_3$  分别表示上相机深度、下相机深度以及双目匹配深度的权重,SR300 相机得到深度的精度比双目计算得到的深度要高,所以实际使用时  $w_1, w_2$  要比  $w_3$  大许多。融合三张深度图的理论相对简单,但实际上,三张深度图的深度信息并非都会永远有效,因此根据实际情况实际的融合算法如3所示。算法3不仅考虑了深度缺失的情况,对于深度信息差值过大的情况也进行了处理。实际处理的效果如图2.13所示。

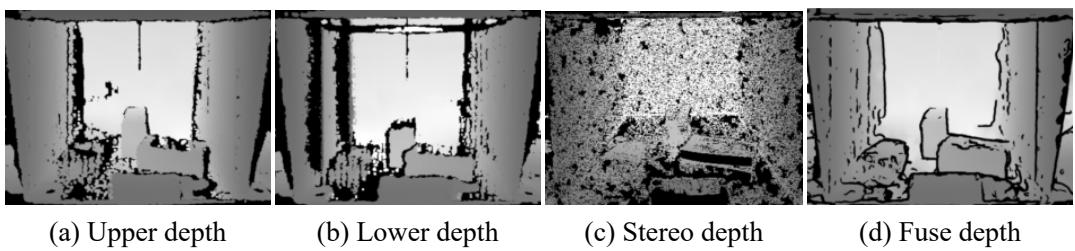


图 2.13 深度融合算法效果图

### 2.3.2 对偶 RGB-D 相机的标定流程

对偶 RGB-D 相机的标定流程可以分为三步:

Step 1 分别标定好单个 RGB-D 相机

Step 2 标定出两个彩色相机之间的齐次变换关系

Step 3 标定出矫正彩色图像的旋转矩阵以及矫正后图像的投影矩阵

---

**算法 3: Fuse Depth Frames**

---

**Input:** leftDepth, rightDepth, stereoDepth**Output:** fuseDepth

```

1 Initialize w1,w2,w3;
2 for ( $d_1, d_2, d_3, d_4$ ) in ( $leftDepth, rightDepth, stereoDepth, fuseDepth$ ) do
3   validDepth = [], validWeight = [];
4   for  $i = 1$  to  $3$  do
5     if  $d_i$  is valid then
6       push back  $d_i$  to validDepth,  $w_i$  to validWeight;
7   if size of validDepth == 0 then
8      $d_4 = \text{NAN};$ 
9   else if size of validDepth == 1 then
10     $d_4 = \text{validDepth}[1];$ 
11   else if size of validDepth == 2 then
12     if extremum of validDepth < 0.03 then
13        $d_4 = \text{validDepth} \cdot \text{validWeight} / \text{sum of validWeight};$ 
14     else
15        $d_4 = \text{NAN};$ 
16   else
17     mediumDepth = medium(validDepth);
18      $d_4 = 0, \text{sum} = 0;$ 
19     for ( $d, w$ ) in (validDepth, validWeight) do
20       if  $\text{abs}(d - \text{mediumDepth}) < 0.03$  then
21          $d_4 += d * w;$ 
22          $\text{sum} += w;$ 
23       if  $\text{sum} > 0$  then
24          $d_4 = d_4 / \text{sum};$ 
25       else
26          $d_4 = \text{NAN};$ 

```

---

单个 RGB-D 相机的标定在2.2.3小节中已经详细叙述过了, 分别标定完单个 RGB-D 相机后, 后面的步骤其实就等价于双目标定了。双目的几何结构如图2.14所示, 标定出两个彩色相机之间的齐次变换关系, 即图2.14中的  $H$ , 简单地可以通过 8

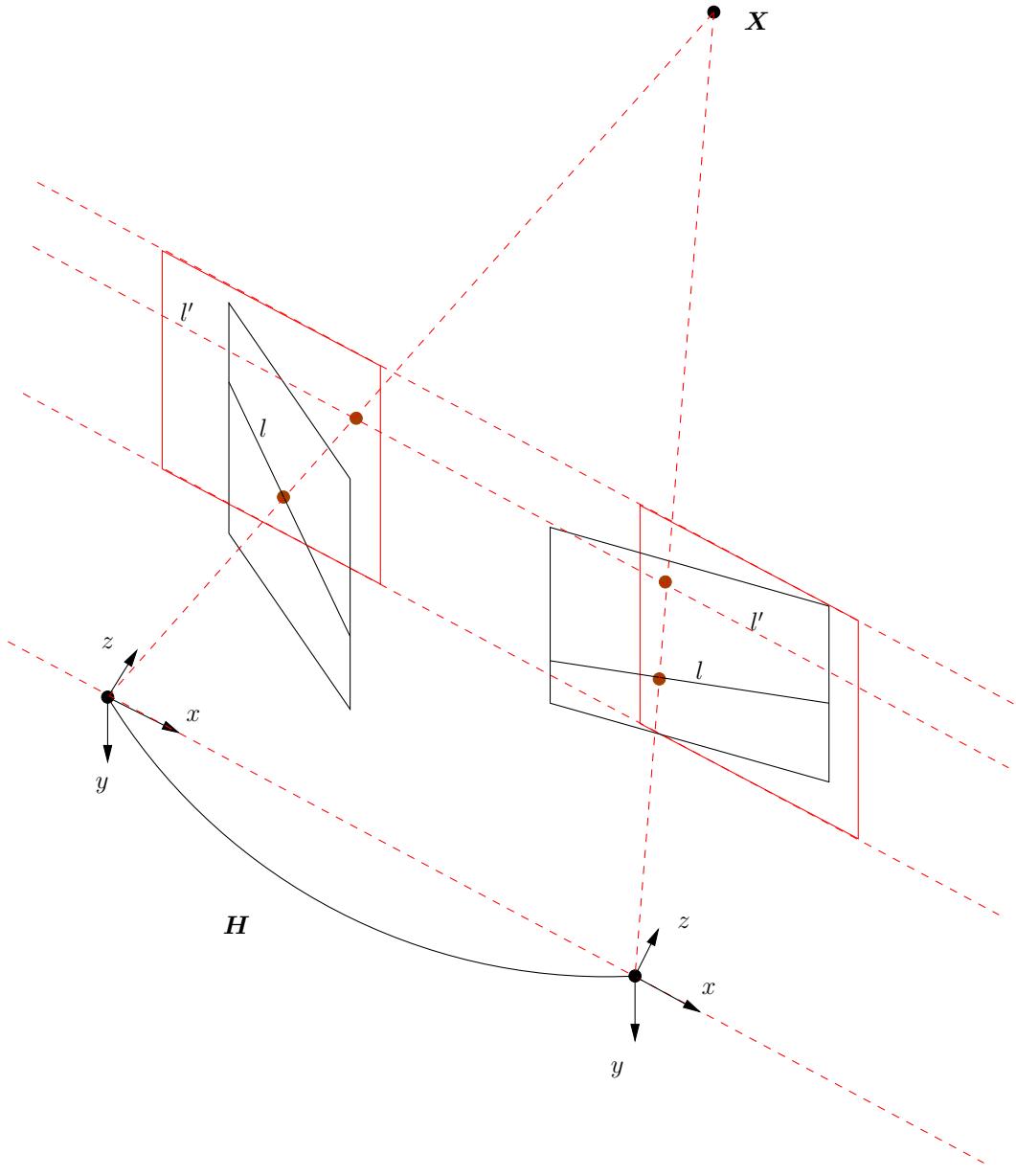


图 2.14 双目几何结构

点法 (Sur et al. 2008) 先求出基础矩阵 (Fundamental Matrix)  $F$ , 即所谓的“弱标定”, 然后根据相机的内参矩阵可求得本质矩阵 (Essential Matrix)  $E$ :

$$E = K^T F K' \quad (2.19)$$

其中  $K$  和  $K'$  分别是两个相机的内参矩阵。求得本质矩阵后可以通过奇异值分解求得齐次变换矩阵的旋转矩阵  $R$  和平移向量  $T$ :

$$\begin{cases} E &= U\Sigma V^T \\ R &= U\mathbf{R}_Z^T(\frac{\pi}{2})V^T \\ [T]_x &= U\mathbf{R}_Z^T(\frac{\pi}{2})\Sigma U^T \end{cases} \quad (2.20)$$

其中  $\mathbf{R}_Z(\theta)$  表示绕 Z 轴旋转  $\theta$  角的旋转矩阵,  $[T]_x$  的定义如下:

$$[T]_x = \begin{bmatrix} 0 & -T_z & T_y \\ T_z & 0 & -T_x \\ -T_y & T_x & 0 \end{bmatrix} \quad (2.21)$$

矫正彩色图像的旋转矩阵会将图2.14中黑色线框的图像平面变换到红色线框的图像平面上,使得对应点在两张图像的同一条极线上。矫正彩色图像的旋转矩阵的计算参考文献 (Loop et al. 1999), 此步标定完最终可以得到:

- 两个相机的矫正旋转矩阵  $R_1, R_2$
- 两个矫正坐标系下的投影矩阵  $P_1, P_2$
- 主相机<sup>③</sup>的投影变换矩阵  $Q$

其中

$$Q = \begin{bmatrix} 1 & 0 & 0 & -u_0 \\ 0 & 1 & 0 & -v_0 \\ 0 & 0 & 0 & f \\ 0 & 0 & 1/B & 0 \end{bmatrix} \quad (2.22)$$

包含了公式2.17由视差计算深度的所有参数。

## 2.4 深度图质量测试实验

RGB-D 相机相对于彩色图我们更关心其深度图的质量,因此设计实验测试了所采用的 SR300 相机深度图的质量以及改进的由两个 SR300 相机所组成的对偶 RGB-D 相机深度图的质量。通过实验,主要考察相机采集的深度在不同距离下的填充率、精度和噪声这三个指标。实验器材除了测试所用的相机,还需要沿固定方向运动的导轨,以及固定在导轨上的平板,相机通过采集平板上的深度信息来计算填充率、精度和噪声这三个指标。实际实验时,由于实验室没有沿固定方向运动的导轨,但是有六轴机械臂,所以将平板固定在机械臂末端,然后通过机

<sup>③</sup> 另外一个相机的投影变换矩阵也可以得到,但没有必要。



图 2.15 深度测试实验装置

械臂示教器控制机械臂末端沿固定方向移动，并且移动的距离可在示教器上读出，整个实验装置如图2.15所示。

#### 2.4.1 实验流程

实验流程示意图如图2.16所示，其中  $z_c$  是相机坐标系的  $z$  轴方向， $z_r$  是机械臂末端运动方向，也是平板运动方向， $z_b$  是垂直于平板的方向， $D_i := \{d_1, d_2, \dots, d_{n_i}\}$  是相机采集到平板的深度信息。具体实验步骤如下：

- 固定机械臂和相机，通过手眼标定（具体见??小节）得到相机坐标系和机器人坐标系之间的齐次变换关系
- 固定平板到机械臂末端
- 通过相机采集平板的深度信息  $D_0$ ，记录此时示教器上机械臂末端在机器人坐标系  $z$  轴上的值  $r_0$
- 控制机械臂示教器使机械臂末端沿机器人坐标系  $z$  轴运动，分别记录此时平板的深度信息  $D_i$  和机械臂末端位置在  $z$  方向上的值  $r_i$
- 重复上述步骤，直到采集满  $n$  组数据

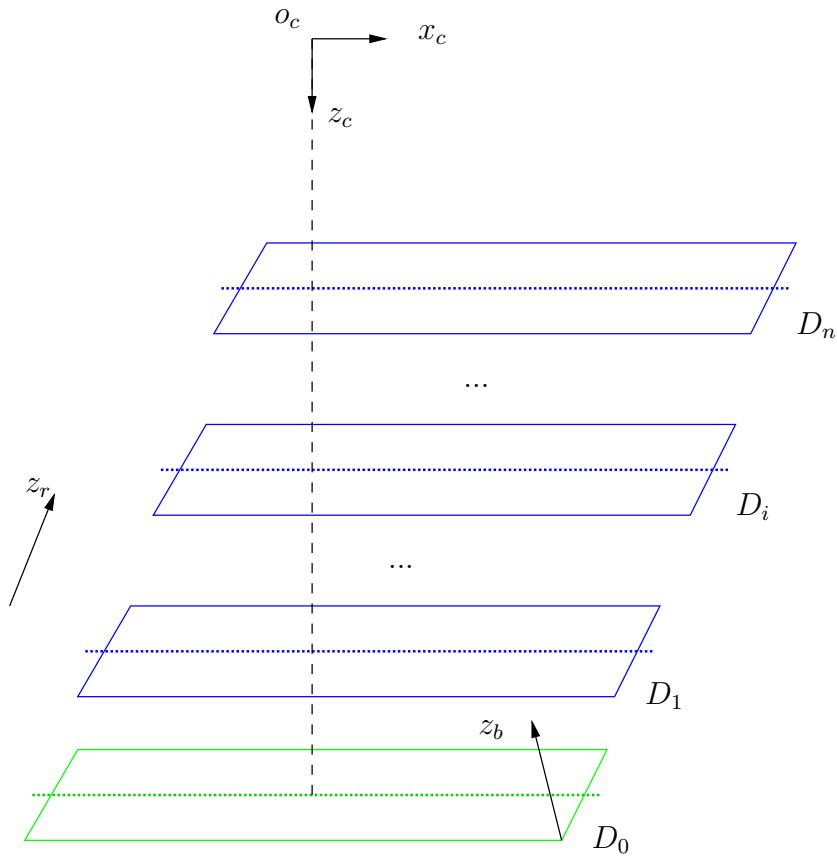


图 2.16 实验流程图

### 2.4.2 实验原理

通过上述实验步骤采集完数据后, 需要计算深度信息的填充率、精度和噪声这三个指标, 下面分别介绍这三个指标的定义和计算方式。

填充率表示深度图中有效深度信息的百分比, 计算相对简单, 首先在采集到的深度图中拟合出平板的平面方程  $\theta_i^T \bar{p} = 0$ , 其中  $\theta_i := [\theta_i(1), \theta_i(2), \theta_i(3), \theta_i(4)]$  为平面方程参数, 然后通过深度图中每个像素到平面的距离确定平板在深度图中的闭合区域。定义像素表示的点到平面的距离小于规定的阈值  $\delta$  时该像素深度信息有效, 最后统计在该闭合区域中像素的总点数  $M_i$  和有效深度信息的像素点个数  $M'_i$ , 则第  $i$  组数据测得的填充率为

$$FillRate_i = \frac{M'_i}{M_i} \times 100\% \quad (2.23)$$

精度表示深度图测量的深度的精度, 定义如下:

$$Precision_i = \frac{|\hat{d}_i - d_i|}{d_i} \times 100\% \quad (2.24)$$

其中  $\hat{d}_i$  表示相机坐标系原点到拟合平面的距离,  $d_i$  表示相机坐标系原点到实际平板的距离, 详细如图2.17所示。 $\hat{d}_i$  的计算相对简单, 根据  $D_i$  中拟合出相机坐标

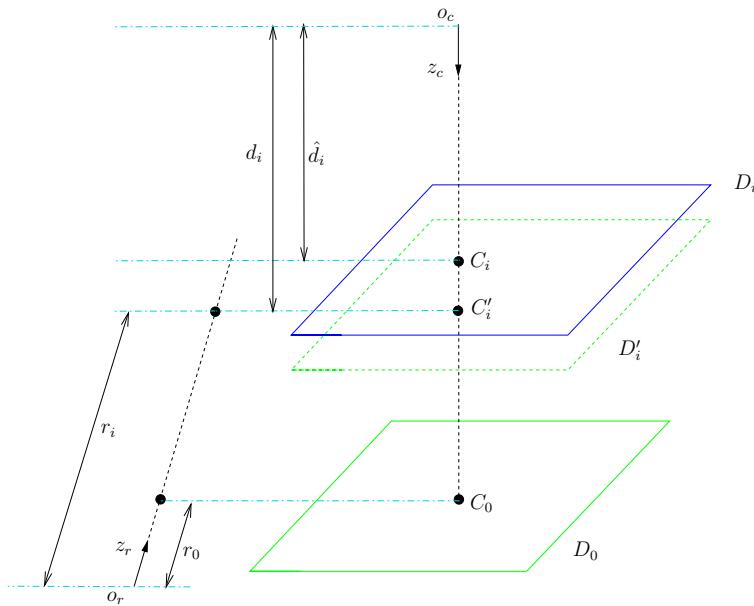


图 2.17 精度测量原理图

系下平板的平面方程的参数  $\Theta_i$ , 容易得到

$$\hat{d}_i = -\frac{\theta_i(4)}{\theta_i(3)} \quad (2.25)$$

为了获得实际平板的与相机坐标系的原点的距离, 首先以  $D_0$  平面为基准, 将其沿机器人坐标系  $z$  轴方向移动  $r_i - r_0$ , 得到新的平面方程  $\theta'_i$ , 则

$$d_i = -\frac{\theta'_i(4)}{\theta'_i(3)} \quad (2.26)$$

噪声定义为平板闭合区域内点距离平面拟合方程的距离的均方根 (RMS):

$$Noise_i = \sqrt{\frac{1}{J} \sum_{j=1}^J \delta_j^2} \quad (2.27)$$

其中  $\delta_j$  为点到拟合平面的距离。

### 2.4.3 实验结果

实验分别对 SR300 和对偶 RGB-D 相机在平板距离相机 0.2 到 1.2m 范围内采集了 1 + 10 组数据, 测得每组数据深度信息的三个指标, 填充率随距离变换的曲线如图??所示。从图中可以看出随着距离的增加, 两个相机的填充率都有所下降, 对偶 RGB-D 相机的填充率相比单个 RGB-D 相机具有更高的填充率。

深度图的精度随距离的变换曲线如图2.19所示, 从图中可以看出精度在  $d = 0.4m$  处最高, 当  $d > 0.4$  时, 越远离相机精度越低; 当  $d < 0.4$  时, 越靠近相机精度越低。对偶 RGB-D 相机与单个 RGB-D 相机在精度上略有提升, 但提升不大。

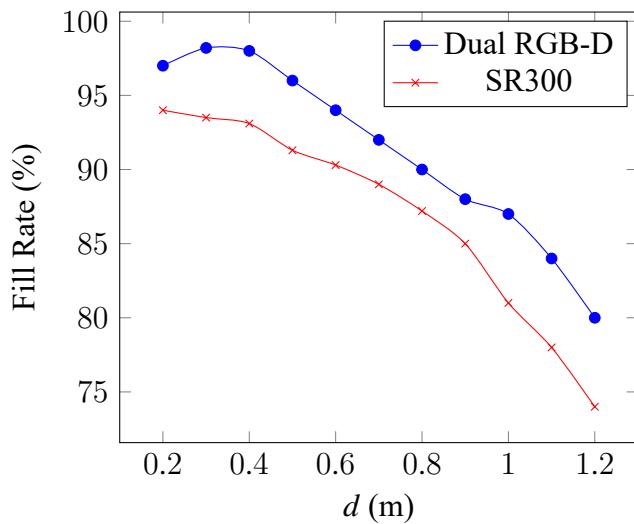


图 2.18 填充率随距离变化曲线

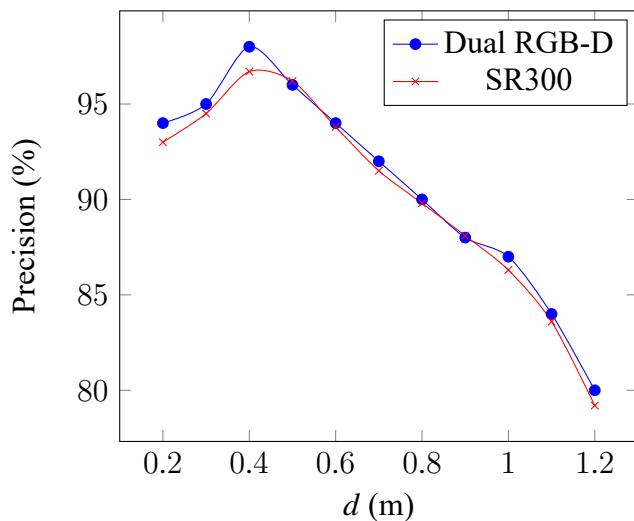


图 2.19 精度随距离变化曲线

深度信息噪声随距离的变化曲线如图2.20所示,从图中可以看出噪声随着距离的增加而增加,并且距离越大,噪声增加的越多。在同等距离下,对偶 RGB-D 相机相比单个 RGB-D 相机具有更小的噪声。

综上,所设计的对偶 RGB-D 相机相比 SR300 相机有更高的填充率,与设计时的初衷一致,毕竟结合了两个相机的深度信息,理论上深度信息的填充率就应该有所增加;对偶 RGB-D 相机的精度与 SR300 相机相比没有太显著的提升,但噪声有着明显的下降。综上可以得出对偶 RGB-D 相机所采集的深度图相比单个 RGB-D 相机有着更高的填充率、更低的噪声,深度图的质量更好。

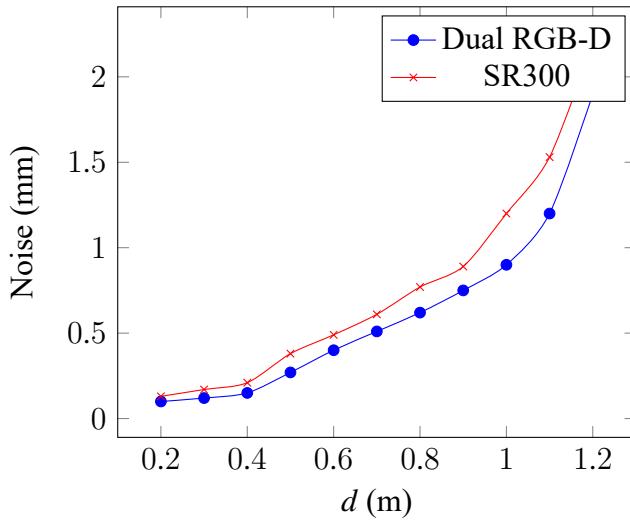


图 2.20 噪声随距离变化曲线

## 2.5 本章小结

本章首先介绍了 RGB-D 相机的现状, 然后详细介绍了以结构光为原理的 RGB-D 相机 SR300 的原理以及标定方法, 针对 SR300 对于反光物体深度信息缺失的情况, 通过组合两个 RGB-D 相机构成对偶 RGB-D 相机实现采集高质量的深度图, 并给出了对偶 RGB-D 相机的标定流程。最后设计了深度图质量测试的实验, 证明了对偶 RGB-D 相机比单个 RGB-D 相机有更高的填充率, 更低的噪声。

## 第 3 章 基于 RGB-D 图像的目标检测算法

本章主要介绍所提出的两种基于 RGB-D 图像的目标检测算法 3D Faster R-CNN 和 3D Mask R-CNN。3D Faster R-CNN 是在 Faster R-CNN(Ren et al. 2015) 的基础上, 通过引入深度图以解决单从 RGB 图难以检测缺少纹理物体 (Textureless Object) 的问题, 并且还引入了 Spatial Transformer 结构使得提取的特征具有旋转不变性。由于 3D Faster R-CNN 目标检测的结果是框出目标的 Bounding Box, 因此使得一些框住细长目标的 Bounding Box 内大部分像素并不属于该目标, 这就使得后面的点云匹配算法难以得到满意的结果。因此 3D Mask R-CNN 根据 Mask R-CNN(He et al. 2017) 对 Faster R-CNN 的改进思路, 对 3D Faster R-CNN 进行了改进, 使得其不仅能得到目标的 Bounding Box, 还能得到目标的 Mask(可以知道 Bounding Box 内属于检测目标的像素), 大大减少了后续匹配算法的难度。

### 3.1 3D Faster R-CNN

3D Faster R-CNN 算法的整体结构如图3.1所示。

相比于 Faster R-CNN, 本文所提出的 3D Faster R-CNN 主要增加对深度信息的处理和 Spatial Transformer, 分别用于解决 Faster R-CNN 在实际应用时所不能解决的问题:

- 难以检测出缺少纹理的物体
- 对物体的旋转敏感, 提取的特征不具有旋转不变性

对于缺少纹理的物体, 单从 RGB 图中很难检测出目标, 这是一个很显然的问题, 但是现在我们可以从对偶 RGB-D 相机中获取深度图, 对于纹理少的物体, 可以从深度图中提取特征检测出目标, 所以现在的关键问题是如何从深度图中提取特征, 并结合到 Faster R-CNN 中, 本文所提出的方法是将深度图转换到 HHA, 然后再使用 CNN 提取特征, 具体后文会详细介绍。

Faster R-CNN 对于物体旋转敏感的问题, 归根到底是因为 CNN 所提取的特征不具有旋转不变性, 实际出现这种问题的情况, 如图3.2所示, 其中图(b)只是将图(a)旋转了 180 度, 由于 CNN 所提取的特征不具有旋转不变性, 并且训练所实验的图片中的宠物都是头朝上的, 即使图(a)在训练集中, 将其旋转 180 度后, 也无法从中检测出目标来。解决这个问题有两个思路:

- Data Augmentation

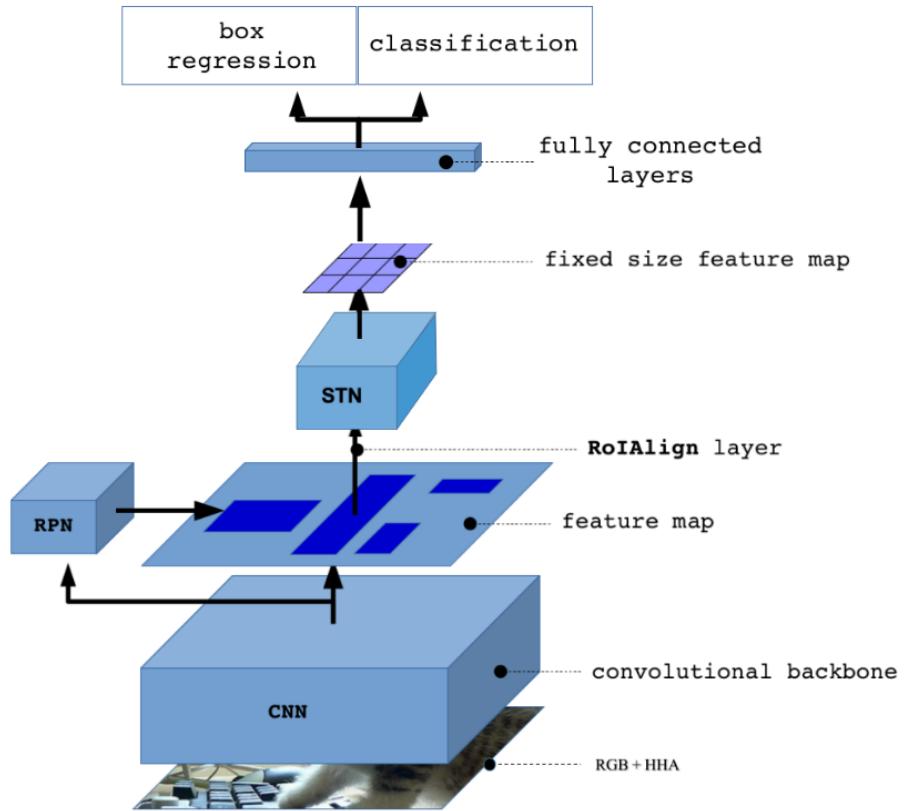


图 3.1 3D Faster R-CNN 结构

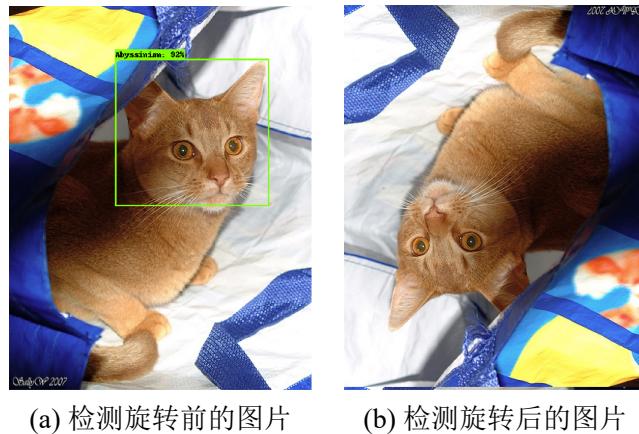


图 3.2 Faster R-CNN 检测识别宠物猫示例

- Spatial Transformer

Data Augmentation 是通过对训练集中的图片进行旋转以获取不同角度的图片, 通过这种方式增大数据集从而使得最终训练得到的模型对各种角度的图片都能识别; Spatial Transformer 是一种特殊的网络结构, 本文所使用的就这种方式, 后文会详细介绍。

### 3.1.1 Faster R-CNN

为了更好地介绍所提出的 3D Faster RCNN 算法, 先回顾一下 Faster R-CNN 算法。Faster R-CNN 的网络结构如图3.3所示, Faster R-CNN 的由两个核心模块构

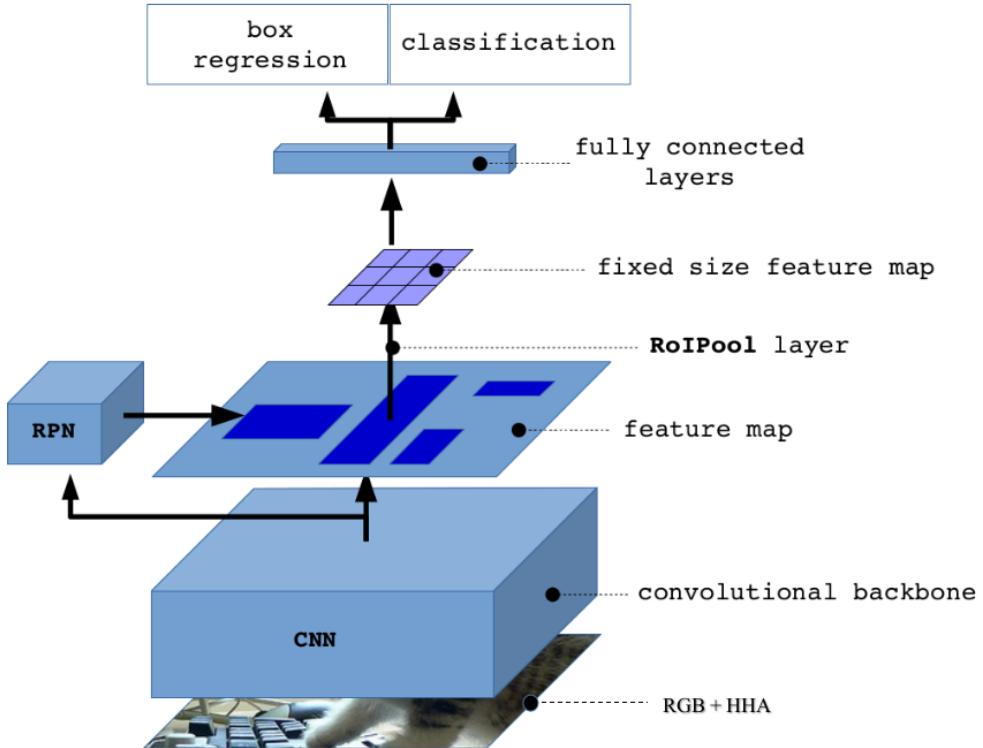


图 3.3 Faster R-CNN 结构

成:

- RPN(Region Proposal Network)
- Fast R-CNN

整个网络是一个端到端(end-to-end)的目标检测网络, 输入图片, 输出图片中检测到的目标的类别和 Bounding Box。RPN 模块输出候选框, 形象地说, RPN 模块告诉 Fast R-CNN 模块去哪里检测目标, Fast R-CNN 模块输出检测结果。

### 3.1.2 HHA

有了与彩色图对应的深度图, 如何有效地利用深度图是一个值得思考的问题。从 2012 年 AlexNet(Krizhevsky et al. 2012) 在 ImageNet(ImageNet 2011) 数据集上的应用开始, 深度学习在计算机视觉领域其准确率相比传统方法有了一个很大的提升, 因此, 本文考虑通过深度学习的方法结合深度图和彩色图进行目标检测。深度学习在彩色图上的应用已经相当成熟, 但对于深度图的应用还比较少, 如何使用 CNN 在深度图上提取特征也是一个值得探讨的问题, 是将深度图直接作

为一个通道使用CNN提取特征？还是将深度图变换到三维坐标( $x, y, z$ )，然后再在这三个通道上通过CNN提取特征？经过实验和相关调研，发现将深度图转换为HHA图后进行训练的模型有较高的准确率(Gupta et al. 2014)，因此本文将深度图转换为HHA三个通道，然后再通过CNN提取特征。HHA三个通道分别为：

- 水平方向上视差(Horizontal disparity)
- 距离地面的高度(Height above ground)
- 法向量与重力的夹角(Angle with gravity)

**Horizontal disparity:** 深度图到视差的转换相对来说十分简单，理论上视差与深度呈倒数关系，因此水平方向上的视差计算具体如算法4所示。

---

#### 算法4：计算水平方向上视差

---

**Input:** Depth Frame  $D_{h \times w}$

**Output:** Horizontal disparity Frame  $H_{h \times w}$

```

1  $h_{floor} = 1/d_{ceil}, h_{ceil} = 1/d_{floor};$ 
2 for  $y \leftarrow 1$  to  $h$  do
3   for  $x \leftarrow 1$  to  $w$  do
4      $H[y, x] = 1/D[y, x];$ 
5    $H[y, x] = (H[y, x] - h_{floor})/(h_{ceil} - h_{floor});$ 

```

---

**Height above ground:** 计算距离地面的高度首先要确定一个世界坐标系，然后得到世界坐标系到相机坐标系的旋转矩阵 ${}^W_C R$ 和平移向量 ${}^W_C T$ ，最后通过坐标变换得到距离地面的高度，具体如算法5所示。

---

#### 算法5：计算距离地面的高度

---

**Input:** Point Cloud  $P_{h \times w}$

**Output:** Height Frame  $H_{h \times w}$

```

1 for  $y \leftarrow 1$  to  $h$  do
2   for  $x \leftarrow 1$  to  $w$  do
3      $p = {}^W_C RP[y, x] + {}^W_C T;$ 
4      $H[y, x] = p.z;$ 

```

---

**Angle with gravity:** 法向量与重力的夹角的计算相对来说稍微复杂一点，重力的方向在工作区间内一般与所设的世界坐标系的 $z$ 轴负方向相同，因此原问题就是求法向量与世界坐标系 $z$ 轴负方向之间的夹角。参考文献(Gupta et al. 2013)，

首先计算深度图中每个点上的法向量, 计算点云中一点  $p_0$  的法向量  $\vec{n}$  的简单思路如下:

- 找出距离点  $p_0$  最近的  $k$  个点:  $p_1, p_2, \dots, p_k$
- 通过最小二乘在点  $\{p_i | i = 0, 1, \dots, k\}$  中拟合出平面  $Ax + By + Cz + D = 0$
- 点  $p_0$  的法向量  $\vec{n} = [A, B, C]^T$

考虑到所采集的深度图转换的点云是有序的 (Organized Point Cloud), 意味着坐标索引相近的点实际物理距离也相近, 因此找出距离点  $p_0$  最近的  $k$  个点可以通过选取点  $p_0$  坐标索引附近的点代替, 具体地, 记点  $p_0$  在深度图中图像坐标为  $(x_0, y_0)$ , 取点集  $S = \{p_i | x_0 - R \leq x_i \leq x_0 + R, y_0 - R \leq y_i \leq y_0 + R\}$ , 其中  $R$  是选取区域的半径。得到法向量后计算法向量与世界坐标  $z$  轴负方向的角度就十分简单了, 整个计算法向量与重力的夹角的算法如6所示。

---

#### 算法 6: 计算法向量与重力的夹角

---

**Input:** Point Cloud  $P_{h \times w}$

**Output:** Angle Frame  $A_{h \times w}$

```

1 for  $y \leftarrow 1$  to  $h$  do
2   for  $x \leftarrow 1$  to  $w$  do
3     Calculate surface normal  ${}^C\vec{n}$  at point  $P[y, x]$ ;
4      ${}^W\vec{n} = {}^W_C R {}^C\vec{n} + {}^W_C T$ ;
5      $A[y, x] = \arccos(-(\vec{n} \cdot \vec{o}_z) / (\|\vec{n}\| \|\vec{o}_z\|))$ ;

```

---

计算完上述 HHA 三个通道后, 为了计算和存储方便, 分别将三个通道的值线性变换到 0 到 255 之间, 可视化如图3.4所示。

### 3.1.3 Spatial Transformer

Spatial Transformer 是一个可微模块, 根据输入的特征对其进行相应空间变化, 输出变换后的特征, 如图3.5所示, 输入特征  $U$  经过 Spatial Transformer 模块后输出特征  $V$ 。Spatial Transformer 模块具体可以分为三个部分, 如图3.6。简单来讲, 第一部分是一个定位网络 (localisation network), 输入特征  $U$ , 输出需要进行空间变换的参数; 第二部分是一个网格生成器 (grid generator), 根据空间变换的参数生成输入特征中需要变换的点的网格; 第三部分是个采样器, 根据网格生成器的输出对输入特征进行采样并进行空间变换, 生成输出特征。

具体地, 记定位网络的输入为特征  $U \in \mathbb{R}^{H \times W \times C}$ , 其中  $W, H, C$  分别为长、宽和通道数, 网络的输出为空间变化  $\mathcal{T}_\theta$  的参数  $\theta$ , 参数  $\theta$  的个数由空间变换的类型决

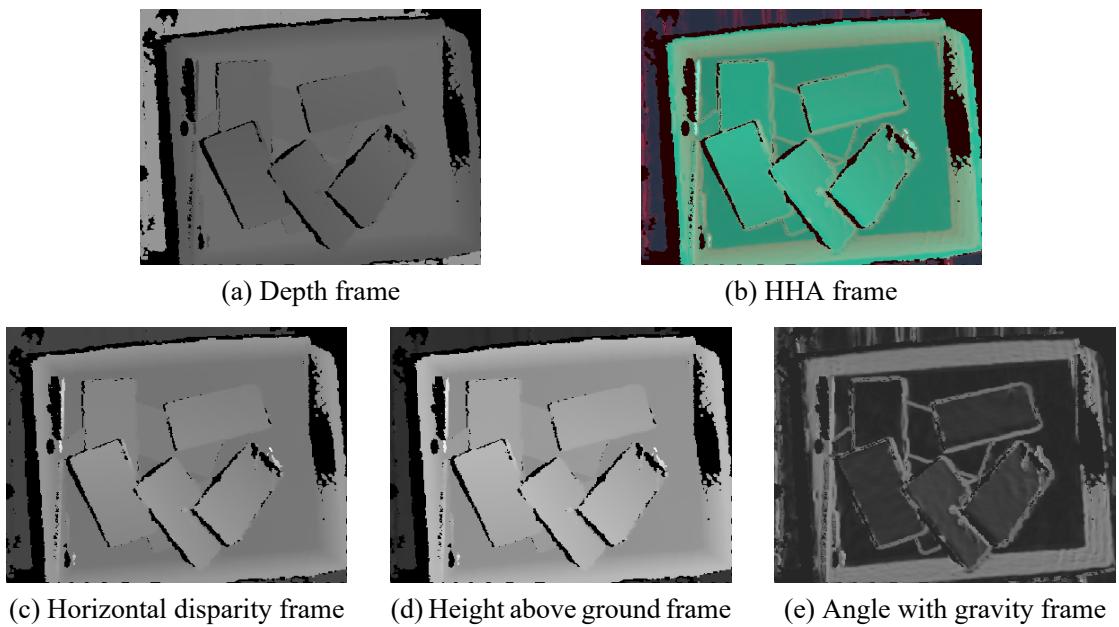


图3.4 HHA可视化效果图

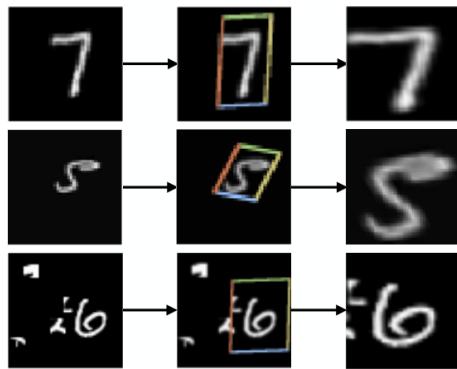


图3.5 Spatial Transformer效果图

定,本文所采用的空间变换为2D仿射变换,则

$$\mathcal{T}_\theta = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \quad (3.1)$$

定位网络内部可以由一些全连接层或者卷积层再加一个回归层组成。

网格生成器本质上就是在输入特征中选取需要进行空间变化的点,如图??中绿色点便是网格生成器所选取的点,记Spatial Transformer的输出特征为 $V \in \mathbb{R}^{H' \times W' \times C}$ ,其中 $W', H', C$ 分别为输出特征的长、宽和通道数,输出特征的通道数和输入特征的通道数相同,不能改变,并且空间变换 $\mathcal{T}_\theta$ 将分别作用于输入 $U$ 的各个通道以保证每个通道上的变换一致。并记点集 $G = \{G_i | G_i = (x_i^t, y_i^t)\}$ ,其中 $(x_i^s, y_i^s)$ 为输出特征图中点的坐标,由定位网络输出的参数 $\theta$ 和 $G$ 我们就可以在输

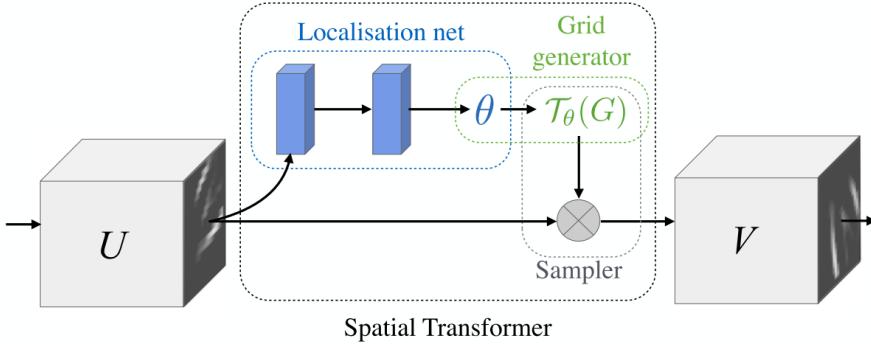


图 3.6 Spatial Transformer 结构图

入特征中确定需要进行空间变换的点的集合  $\mathcal{T}_\theta(G)$ :

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \mathcal{T}_\theta(G_i) = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \quad (3.2)$$

其中  $(x_i^s, y_i^s)$  是输入特征中点的坐标,也是图??中的绿色点。

采样器输入网格生成器生成的点集  $\mathcal{T}_\theta$ ,和输入特征  $U$ ,最终输出经过空间变换后的特征  $V$ ,具体如公式3.3所示:

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c k(x_i^s - m; \Phi_x) k(y_i^s - n; \Phi_y) \quad \forall i \in [1 \dots H'W'] \quad \forall c \in [1 \dots C] \quad (3.3)$$

其中  $\Phi_x$  和  $\Phi_y$  是采样核函数  $k()$  的参数,  $U_{nm}^c$  表示输入特征  $U$  在坐标  $(n, m)$  下第  $c$  个通道上的值,  $V_i^c$  表示输出特征在坐标  $(x_i^s, y_i^s)$  下第  $c$  个通道上的值。理论上可以使用任何采样核函数,只要可以对  $x_i^s$  和  $y_i^s$  求导,因为网络训练需要对公式3.3求导。以双线性采样核函数为例,公式3.3变为

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c \max(0, 1 - |x_i^s - m|) \max(0, 1 - |y_i^s - n|) \quad (3.4)$$

则  $V$  对  $U$  和  $G$  的梯度为

$$\frac{\partial V_i^c}{\partial U_{nm}^c} = \sum_n^H \sum_m^W \max(0, 1 - |x_i^s - m|) \max(0, 1 - |y_i^s - n|) \quad (3.5)$$

$$\frac{\partial V_i^c}{\partial x_i^s} = \sum_n^H \sum_m^W U_{nm}^c \max(0, 1 - |y_i^s - n|) \begin{cases} 0 & \text{if } |m - x_i^s| \geq 1 \\ 1 & \text{if } m \geq x_i^s \\ -1 & \text{if } m < x_i^s \end{cases} \quad (3.6)$$

$$\frac{\partial V_i^c}{\partial y_i^s} = \sum_n^H \sum_m^W U_{nm}^c \max(0, 1 - |x_i^s - m|) \begin{cases} 0 & \text{if } |n - y_i^s| \geq 1 \\ 1 & \text{if } n \geq y_i^s \\ -1 & \text{if } n < y_i^s \end{cases} \quad (3.7)$$

## 3.2 3D Mask R-CNN

Mask R-CNN 相比 Faster R-CNN 不仅可以输出目标的 Class 和 Bounding Box, 还可以输出目标的 Mask。Mask R-CNN 相比 Faster R-CNN 主要的技术要点有:

- 强化了特征提取网络
- 采用 ROIAlign 代替 ROIPooling
- Mask 的损失函数

因此 3D Mask R-CNN 也针对上述三个要点对 3D Faster R-CNN 进行改进, 其结构框架如图3.7所示。

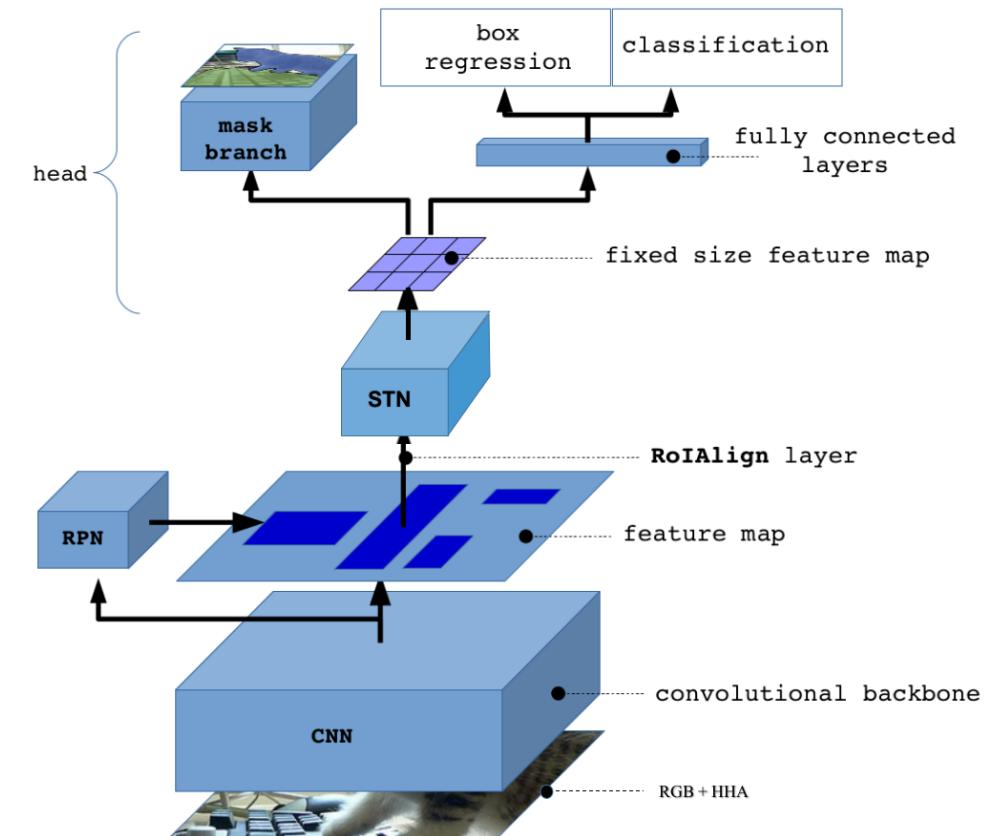


图 3.7 3D mask R-CNN 结构

### 3.2.1 特征提取网络

3D Faster R-CNN 的特征提取网络使用的是 VGG-16(Simonyan et al. 2014), VGG16 是牛津大学 VGG 组提出的。VGG16 相比最早的 AlexNet 的一个改进是采用连续的几个  $3 \times 3$  的卷积核代替 AlexNet 中的较大卷积核( $11 \times 11, 5 \times 5$ )。对于给定的感受野(与输出有关的输入图片的局部大小),采用堆积的小卷积核是优于采用大的卷积核,因为多层非线性层可以增加网络深度来保证学习更复杂的模式,而且代价还比较小(参数更少)。比如,3 个步长为 1 的  $3 \times 3$  卷积核连续作用在一个大小为 7 的感受野,其参数总量为  $3 \times 9C^2$ ,其中  $C$  是通道数,如果直接使用  $7 \times 7$  卷积核,其参数总量为  $49C^2$ 。而且  $3 \times 3$  卷积核有利于更好地保持图像性质。

3D Mask R-CNN 改用了 ResNeXt-101+FPN 网络提取特征,该网络主要由 ResNeXt(Xie et al. 2017) 和 FPN(Lin et al. 2017) 两部分构成。ResNeXt 是对残差网络 ResNet(He et al. 2016) 的改进,在介绍 ResNeXt 之前先介绍一下 ResNet。ResNet 为了解决随着网络层数增加,靠前的层梯度会很小,导致训练时学习停滞、梯度消失的问题,引入了残差模块,如图3.8所示。残差单元可以解决学习停滞问

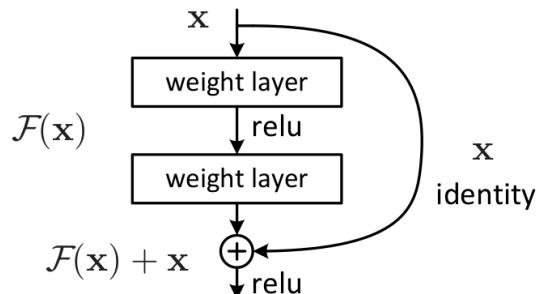


图 3.8 残差模块

题的背后逻辑在于此:想象一个网络 A,其训练误差为  $x$ 。现在通过在 A 上面堆积更多的层来构建网络 B,这些新增的层什么也不做,仅仅复制前面 A 的输出。这些新增的层称为 C。这意味着网络 B 应该和 A 的训练误差一样。那么,如果训练网络 B 其训练误差应该不会差于 A。但是实际上却是更差,唯一的原因是让增加的层 C 学习恒等映射不容易。为了解决这个退化问题,残差模块在输入和输出之间建立了一个直接连接,这样新增的层 C 仅仅需要在原来的输入层基础上学习新的特征,即学习残差,会比较容易。ResNeXt 在 ResNet 的基础上,提出 cardinality 的概念,如图3.9,其中左右两个网络结构有相同的参数个数,左边是 ResNet 的一个区块,右边的 ResNeXt 中每个分支一模一样,分支的个数就是 cardinality,其通过在大卷积核层两侧加入  $1 \times 1$  的网络层来控制核个数、减少参数个数。因此,与 ResNet 相比,相同的参数个数,ResNeXt 结果更好:一个 101 层的 ResNeXt 网络,

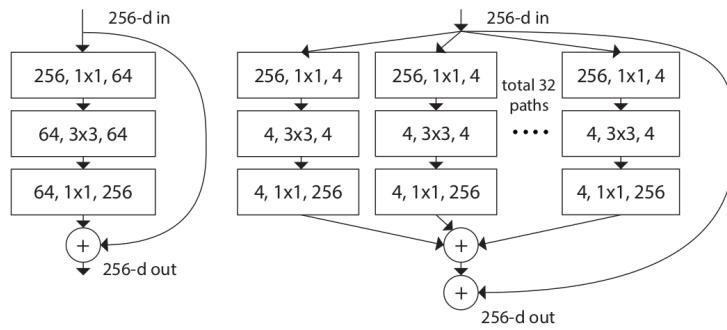


图3.9 ResNeXt对ResNet的改进

和200层的ResNet准确度差不多,但是计算量只有后者的一半。

### 3.2.2 ROIAlign

ROIPooling采用的是最近邻插值(Nearest neighbor interpolation),即在resize时,对于缩放后坐标不能刚好为整数的情况,采用四舍五入的方法,相当于选取离目标点最近的点。虽然这种处理方法对分类问题影响不大,但是现在Mask R-CNN需要预测Pixel级别的Mask,ROIPooling造成的像素不对齐问题对Mask的精确度影响很大,因此提出ROIAlign代替ROIPooling,ROIAlign使用双线性插值(Bilinear interpolation)来获得像素级别的对齐。举例来说,假设在 $8 \times 8$ 的特征图中提取 $2 \times 2$ 的输出,ROIPooling的示意图如3.10所示,ROIAlign的示意图如3.11所示。

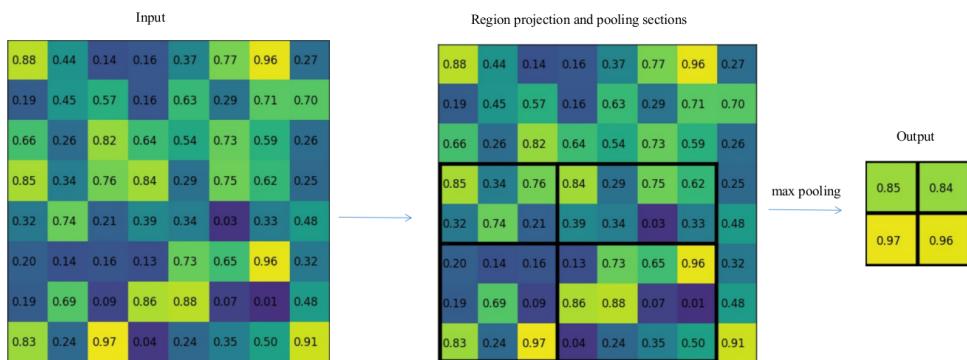


图3.10 ROIPooling示意图

### 3.2.3 Mask损失函数

为了有效的避免类之间的竞争,使得其他class不贡献损失,Mask的损失函数使用平均二值交叉熵(average binary cross-entropy)。具体的,对于每个ROIAlign的 $K \times m^2$ 维输出, $K$ 表示总类别个数, $m$ 对应mask分辨率,即输出 $K$ 个mask,定

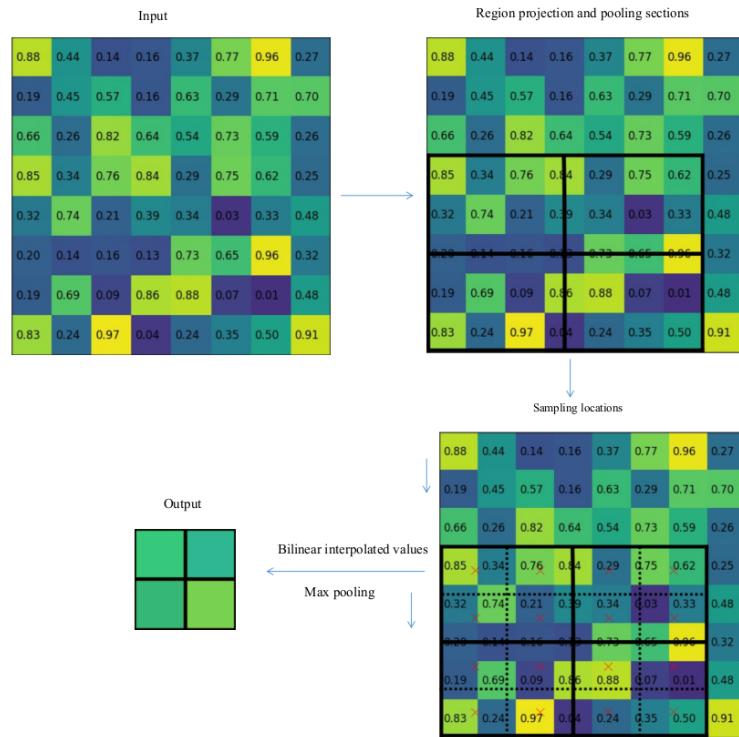


图 3.11 ROIAlign 示意图

义 Mask 的损失函数为

$$L_{mask} = -\frac{1}{K} \sum_{i=1}^K \sum_j (y'_j \lg(y_j) + (1 - y'_j) \lg(1 - y_j)) \quad (3.8)$$

其中

$$j = 1, 2, \dots, m^2 \quad (3.9)$$

表示 mask 的像素索引。通过上述损失函数的定义有效的避免了类间的竞争, 将 mask 分支与 class 分支并行区分开来, 通过 class 分支最终的输出在  $K$  个 mask 中选择对应的 mask 输出。实验表明, 相比将 mask 和 class 混在一起, 根据 mask 的结果来判断类别的方法, 这种方法对算法最终的精确度有着重要意义。

### 3.3 目标检测实验

为了评价所设计的 3D Faster R-CNN 和 3d Mask R-CNN 算法的性能, 分别在一个现有的数据集和一个自己采集的实际应用的数据集上进行了网络的训练和测试, 并与原始的 Faster R-CNN 和 Mask R-CNN 相比较, 验证了所设计算法的性能。

### 3.3.1 数据集

实验所采用的数据集一个是参加 APC(Amazon Picking Challenge) 的 MIT-Princeton 队伍所采集的数据集”Shelf & Tote” Benchmark Dataset(MIT-Princeton 2016), 此处简单记为 APC 数据集, 另外一个数据集是实际用于 bin-picking 在实验室采集的数据集, 记为 workpiece 数据集。

**APC 数据集:** 该数据集共有 39 类不同的物体, 452 个场景, 每个场景有不同的物体, 一共 7281 组图片, 通过在多个场景下, 不同的视角下使用 Intel Realsense SR300 相机所拍摄。标注的数据是每个场景下物体在世界坐标系下的位姿, 以及每个场景下相机在世界坐标系下的位姿, 是一个半自动标注的数据集, 通过物体的位姿和相机的位姿就可以得到每个物体在相机坐标系下的位姿, 数据集中部分数据如图3.12所示。

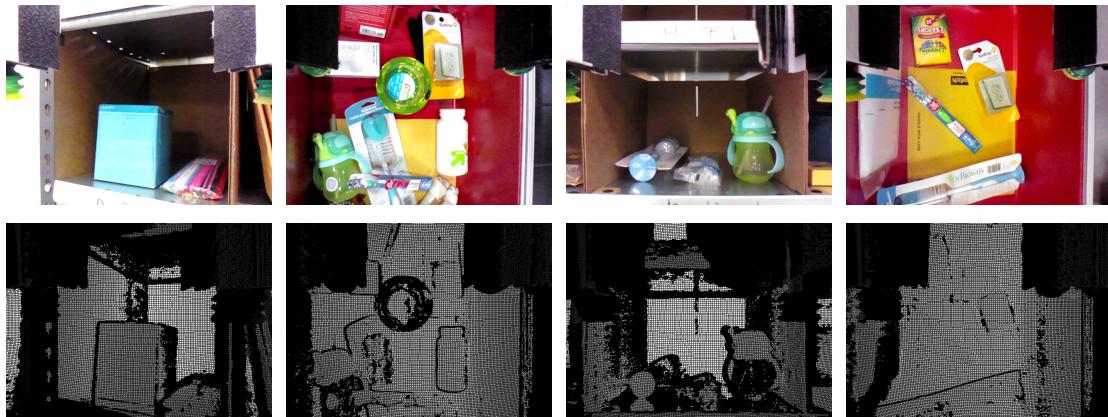


图 3.12 APC 数据集部分数据: 第一栏为彩色图像, 第二栏为与彩色图像相匹配的深度图

APC 数据集中标注的标签可以认为是每个物体在相机坐标系下的位姿和类别, 对于设计的算法来说需要的是物体的类别(class)、边界框(bounding box)和掩模(mask), 因此需要对原始标注数据进行一些处理, 因为 APC 还提供了每类物体的 CAD 模型, 并且相机的内参矩阵也在数据集中提供了, 因此可以将 CAD 模型转换为点云后齐次变换到所标注的对应物体在相机坐标系下的位姿, 然后利用相机内参矩阵将物体点云投影到图像平面, 从而获得物体的 mask, 进而可以得到物体的 bounding box。需要注意的是由于一个场景中有多个物体, 在不同相机位姿下会出现遮挡, 因此需要对被遮挡物体的 mask 进行相应的裁剪, 对于几乎被完全遮挡的物体可以去除, 判断物体是否遮挡可以通过物体点云距离相机原点的距离远近判断。将一张图中物体位姿得到 mask 和 boudning box 的处理流程如下所示:

1. 对于图中标注的每个物体:
  - 将对应物体的 3D 点云变换到物体标注的位姿

- 根据相机内参矩阵将 3D 点云投影到图像平面, 获得物体的 mask 以及 mask 对应的深度图 depth
- 遍历像素索引 i:
    - 如果在索引 i 出存在多张 mask 的值有效, 保留 depth 值最小的 mask, 将其余 mask 在索引 i 处置为无效
  - 对于每个物体的 mask:
    - 如果 mask 中有效像素点小于阈值 T, 删 除该 mask
    - 根据 mask 有效像素点的坐标计算对应的 bounding box

处理后的部分图片的 ground truth (class, mask, boudning box) 如图3.13所示。从



图 3.13 APC 数据集部分标注数据

图3.13可以看出处理后的 mask 基本覆盖了物体, boudning box 也正确框出了物体, 唯一的缺点是所生成的 mask 有时候有些缺失, 没有人工标注的完美, 如图3.13中第一张图中的瓶子(easter turtle sippy cup)标注的 mask 有很多缺失, 根本原因是所使用的物体的 CAD 模型是通过相机采集生成的, 其转换的 3D 点云质量并不是十分理想, 其 3D 点云比较稀疏并且有部分缺失, 如图3.14所示, 这个瓶子的点云有严重的缺失, 主要原因是瓶子透明, 所以相机难以采集其深度信息。模型点云的缺失, 因此将点云投影到图像平面生成的 mask 也有部分缺失, 尽管已经对生成的 mask 进行了一些滤波处理, 但部分 mask 还是有明显的缺失。

总体来说, 尽管生成的 ground truth 的质量没有人工标注的 ground truth 质量

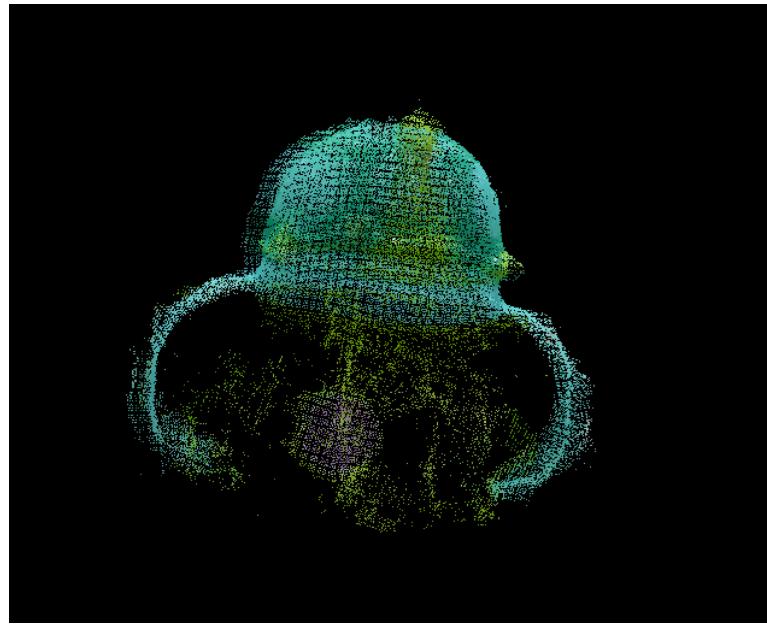


图 3.14 easter turtle sippy cup point cloud

好,但对本实验来说已经够用,并且相比人工标注这种半自动化的标注方式节省了大量时间和金钱成本。

**workpiece** 数据集: 该数据集有三类物体,共 2k 组图片。该数据集与 APC 数据集最大的不同是,同一张图片中存在大量不同位姿的同种物体,并且三类物体都缺少纹理(textureless),因此 Faster R-CNN 和 Mask R-CNN 在该数据集上的表现理论上应该大大不如 3D Faster R-CNN 和 3D Mask R-CNN。部分数据集中的图片如图3.15所示。**workpiece** 数据集的 ground truth 由人工标定,其中据测试集中

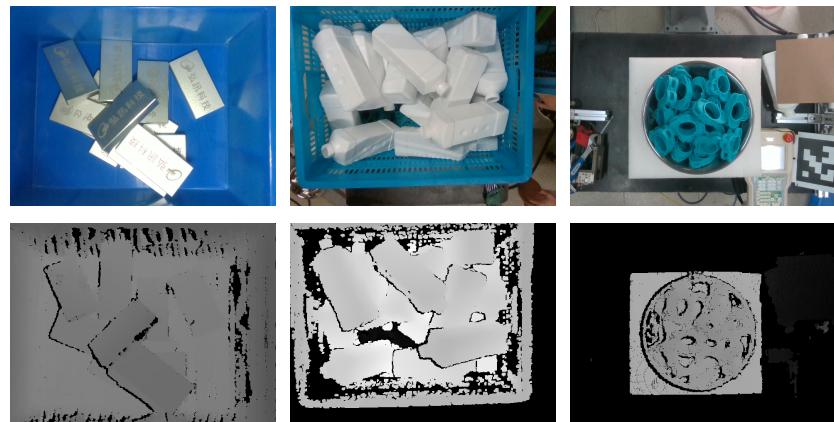


图 3.15 workpiece 数据集部分图片

有的 ground truth 不仅包括了物体的 class, mask, bounding box, 还有物体的位姿,并且由于三类物体都是工厂中的工件,因此也提供三类物体精确的 CAD 模型。

### 3.3.2 实验内容

实验在 APC 数据集和 workpiece 数据集上比较 Faster R-CNN 和 3D Faster R-CNN、Mask R-CNN 和 3D Mask R-CNN 的性能。

算法实现主要通过 Tensorflow 框架使用 python 语言实现, 详细代码见 Github 项目地址<sup>①</sup>。

评价的指标主要是检测的精确度 AP 以及算法的时间性能 FPS。FPS 是每秒能检测的图片数比较好理解, AP 是 bounding box 或者 mask 交并比的精确度。具体地, 如图3.16所示, 两个 bounding box 的交并比定义为:

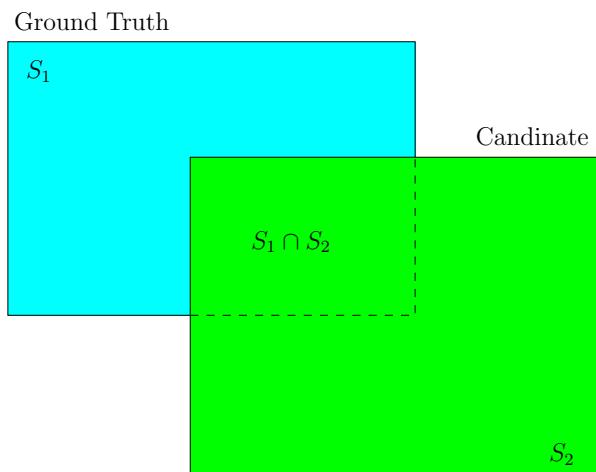


图 3.16 bounding box 交并比

$$IoU = \frac{S_1 \cap S_2}{S_1 \cup S_2} \quad (3.10)$$

$AP_{0.5}$  表示检测的结果与 ground truth 的交并比大于 0.5 的个数占总体检测个数的比例, 显然定义精度的  $IoU$  大小会影响最终评价的质量, 过小和过大的最小  $IoU$  都不能很好地反应算法的精缺度, 因此将评价的主要精确度定义如下:

$$AP = \frac{1}{10} \sum_{i=0}^9 AP_{0.5+0.5i} \quad (3.11)$$

检测结果换为 mask 精确度的定义也类似, 只需用 mask 的交并比代替 bounding box 的交并比。

模型训练在实验室的服务器上进行, 服务器有两块 Intel(R) Xeon(R) E5-2683 v3(2.00GHz) 的 CPU, 4 块 TITAN X GPU。模型训练时为了减少训练时间, 4 块 GPU 都使用了。在 APC 数据集上, 训练用了约 6k 组图片, 剩下的 1k 多组图片用于测试, 3D Faster R-CNN 训练用了 40 个小时左右, 3D Mask R-CNN 用了 48 个小

<sup>①</sup> [https://github.com/freealong/Mask\\_RCNN](https://github.com/freealong/Mask_RCNN)

时左右；在 workpiece 数据集上，训练用了约 1.6k 组图片，剩下的 400 组图片用于测试，3D Faster R-CNN 训练用了 30 个小时左右，3D Mask R-CNN 用了 36 小时左右。

### 3.3.3 实验结果

在 APC 数据集上，本文算法 Faster R-CNN 和 Mask R-CNN 的精确度如表3.1所示，在测试集上的部分图片检测结果见图??。从表3.1中可以看出在 APC

表 3.1 算法在 APC 数据集上的精确度

	input	output	$AP$	$AP_{0.5}$	$AP_{0.75}$
Faster R-CNN	RGB	bbox	33.26	56.29	34.03
<b>3D Faster R-CNN</b>	RGB+HHA	bbox	<b>34.55</b>	<b>57.99</b>	<b>34.69</b>
Mask R-CNN	RGB	mask	32.34	55.78	33.12
<b>3D Mask R-CNN</b>	RGB+HHA	mask	<b>33.94</b>	<b>56.45</b>	<b>33.99</b>

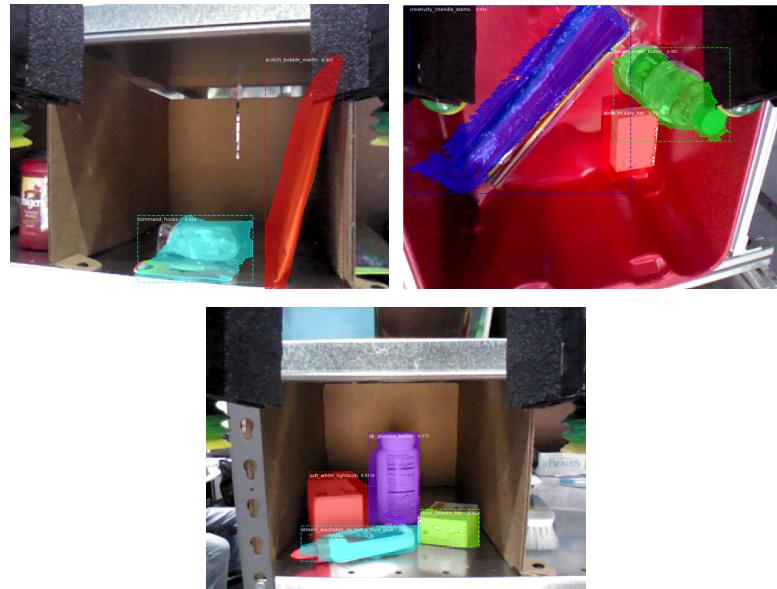


图 3.17 算法 APC 数据集上部分检测结果

数据集上 3D Faster R-CNN 相比 Faster R-CNN 的精确度提高了 1.3 个百分点左右，3D Mask R-CNN 相比 Mask R-CNN 提高了 0.8 个百分点左右。整体来说对精确度的提高并不是十分明显，究其原因，从图3.12可以看到 APC 数据集中的物体大多也是纹理丰富的，单从 RGB 图就可以训练出一个很好的模型，因此增加 HHA 通道，对模型精确度的提升十分有限，反而降低了算法的 FPS。

在 workpiece 数据集上, 本文算法 Faster R-CNN 和 Mask R-CNN 的精确度如表3.2所示, 在测试集上的部分图片检测结果见图3.18。从表3.2可以看出在

表 3.2 算法在 workpiece 数据集上的精确度

	input	output	$AP$	$AP_{0.5}$	$AP_{0.75}$
Faster R-CNN	RGB	bbox	18.78	37.49	19.46
<b>3D Faster R-CNN</b>	RGB+HHA	bbox	<b>32.39</b>	<b>56.37</b>	<b>33.54</b>
Mask R-CNN	RGB	mask	16.12	35.95	18.74
<b>3D Mask R-CNN</b>	RGB+HHA	mask	<b>30.98</b>	<b>53.74</b>	<b>32.19</b>



图 3.18 算法 workpiece 数据集上部分检测结果

workpiece 数据集上, 3D Faster R-CNN 相比 Faster R-CNN 的精确度提高了 13.6 个百分点左右, 3D Mask R-CNN 相比 Mask R-CNN 提高了约 14.8 个百分点。显然, 无论是 3D Faster R-CNN 还是 3D Mask R-CNN, 在 workpiece 数据集上精确度相比原算法有了大大的提高, 从图3.15可以发现 workpiece 数据集中的图片包含的都是一些缺少纹理的物体, 并且有大量同种物体混杂在一起, 有时候人眼也很难从中区分单个目标, 因此可能单从 RGB 图难以训练出一个准确率较高的模型来检测目标。而这些缺少纹理的大量物体在深度图, 尤其是变换后的 HHA 图上十分容易区分出来, 因此 3D Faster R-CNN 和 3D Mask R-CNN 引入 HHA 后, 增加了更多信息, 最终训练得到的模型的准确度相比原算法有了巨大的提升。

3D Faster R-CNN 和 3D Mask R-CNN 算法的时间性能见表3.3, 由于两个数据集内图片的大小都是一样的, 因此算法的时间性能在两个数据集上并不会有什么差异, 因此表3.3中直接统计了算法在两个数据集测试样本上 FPS 的平均值。从表3.3可以看出 3D Faster R-CNN 和 3D Mask R-CNN 相比原算法, 普遍具有更低的 FPS, 因为增加了 HHA 数据并且增加了 STN 模块。但考虑到本文算法的具体应用, 适当降低的 FPS 并不会对具体使用造成什么影响。

	Faster R-CNN	<b>3D Faster R-CNN</b>	Mask R-CNN	<b>3D Mask R-CNN</b>
FPS	5.5	<b>3.2</b>	4.1	<b>2.5</b>

表3.3 算法时间性能

### 3.4 本章小结

本章主要介绍了两个目标检测的算法3D Faster R-CNN和3D Mask R-CNN，算法在Faster R-CNN和Mask R-CNN的基础上通过引入深度图以解决单从RGB图难以检测缺少纹理物体(Textureless Object)的问题，并且还引入了Spatial Transformer结构使得提取的特征具有旋转不变性。最后通过在APC数据集和workpiece数据集上的目标检测实验，证明了3D Faster/Mask R-CNN相比原算法有更高的精确度，但FPS相对来说较低。

## 第 4 章 基于 4PCS 的点云匹配算法

本章主要介绍为了估计目标的位姿所提出的一种基于 4PCS 的点云匹配算法 A4PCS-ICP (Angle-fixed-4PCS-ICP), A4PCS-ICP 主要基于全局匹配算法 4PCS(Aiger et al. 2008) 和局部匹配算法 ICP(Besl et al. 1992)。为了详细介绍 A4PCS-ICP 算法, 本章首先从整体上简单介绍了 A4PCS-ICP 算法, 包括算法所要解决的问题的具体数学描述以及相关算法的背景; 然后介绍 A4PCS-ICP 算法的基础 4PCS 算法, 并分析了其不足, 进而提出 A4PCS 算法对其进行改进; 接着介绍与 A4PCS 算法相结合的 ICP 算法, ICP 算法主要用于提高最终点云匹配的精度; 最后进行了点云匹配的实验, 将本文的 A4PCS-ICP 算法与其他几个匹配算法相比较。

### 4.1 点云匹配算法概述

#### 4.1.1 问题描述

通过第 3 章中的目标检测算法可以得到目标的 bounding box 或者 mask, 根据 bounding box 或者 mask 可以在深度图中提取对应的区域, 从而获得包含目标的点云。所以现在的问题是如何通过目标的点云计算出目标的位姿, 由于可以得到目标的三维模型, 因此将目标的三维模型经过一个刚体变换  $T$ , 使之与目标点云重合, 然后目标的三维模型在相机坐标系下的位姿也是已知并且可调的, 为方便起见将三维模型坐标系与相机坐标系重合, 则目标的位姿便等于三维模型与目标点云之间的齐次变换关系, 即  $T$ 。所以, 要计算目标的位姿, 就要求解目标三维模型与相机采集到的目标点云之间的刚体变换  $T$ , 如图 4.1, 这也是 A4PCS-ICP 主要要解决的问题: 两个点云之间的匹配问题。

相机所采集到的目标点云是一组包含空间三维坐标( $x, y, z$ )以及颜色( $r, g, b$ )的点集<sup>①</sup>, 由于此处并不需要颜色信息, 因此对目标点云只保留空间位置信息, 去除颜色信息后的目标点云记为点集  $P$ 。三维模型亦可通过采样得到一组包含空间三维坐标的点集, 记为  $Q$ 。A4PCS-ICP 算法就可以简化为求解一个刚体变换  $T$  使得点集  $P$  中的点经过矩阵  $T$  变换后, 尽可能与点集  $Q$  重合。更为准确地, A4PCS-ICP 算法就可以简化为求解一个 LCP(Largest Common Pointset)问题:

---

<sup>①</sup> 对点集(point set)与点云(point cloud)不做区分, 都指包含坐标点的集合



图 4.1 位姿估计示意图

LCP 问题：给定两个点集  $P$  和  $Q$ ，在给定距离误差  $\delta$  下，求解点集  $P$  的最大子集  $P'$ ，使得  $T(P')$  和点集  $Q$  之间的距离在合适的距离度量下小于  $\delta$ ，其中  $T$  是一个刚体变换。

#### 4.1.2 背景介绍

LCP 问题并不是一个新的问题，解决该问题的算法也有很多，尤其是近些年来，随着几何扫描相关技术的发展，如何将多次扫描或者多个设备采集的三维信息统一到一个坐标系下成为研究的热点，其本质上可以归结为 LCP 问题或其衍生问题，这些问题是在计算机几何学和计算机视觉中的基础问题。

其中一个比较流行的算法是通过使用稳定的局部几何描述子来匹配得到粗略的刚体变换，然后紧接着使用 ICP 算法迭代获取较为精确的刚体变换 (Li et al. 2005)。这种算法的效果十分取决于所选取的描述子，通常一般的描述子对传感器噪声都比较敏感，尤其是一些低精度的传感器，常用的局部几何特征描述子有 SHOT(Salti et al. 2014)、FFPH(Rusu et al. 2009) 等；还有一种比较流行的方法是通过几何希哈方法从事先设置好的候选集中来选择合适的刚体变换 (Wolfson et al. 1997)；一些随机算法，如 RANSAC(Random Sample Consensus)(Bolles et al. 1981) 通常需要足够长的时间才能保证得到合适的解。

上述介绍的一些算法，有些对噪声的鲁棒性不强，有些时间复杂度极高，有些也难以处理部分重叠的情况，即点集  $P$  和  $Q$  之间只有一部分点集是相匹配的，因此难以实际直接应用到本文所需要解决的问题，其效果也难以让人满意。对此，本文基于 4PCS(4-Points Congruent Sets) 算法设计了有效解决点云匹配的算法 A4PCS-ICP。

## 4.2 A4PCS-ICP 算法

### 4.2.1 算法框架

A4PCS-ICP 算法基于 4PCS，针对 4PCS 的瓶颈，有效地降低了其时间复杂度，然后通过与 ICP 算法配合提高匹配的精度，其整体框架如图 4.2 所示。A4PCS-ICP

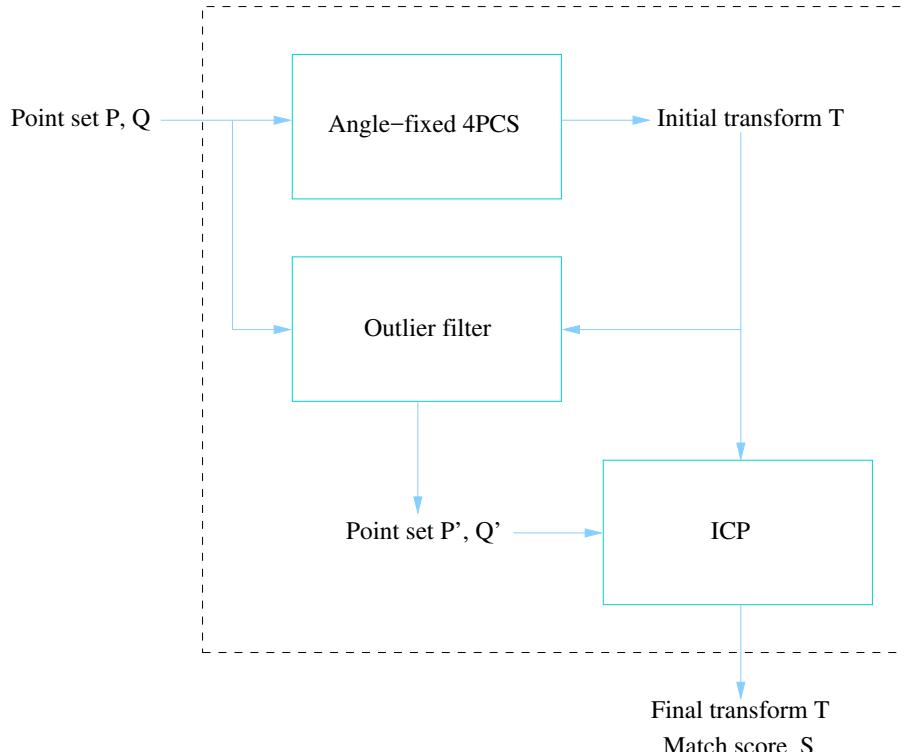


图 4.2 A4PCS-ICP 算法框架

由三个模块组成：Angle-fixed 4PCS、Outlier filter 和 ICP，Angle-fixed 4PCS 是 4PCS 算法的优化版，也是 A4PCS-ICP 算法的核心，根据输入的两个点集，输出两个点集之间的粗略的变换关系；Outlier filter 根据 Angle-fixed 4PCS 的输出对点集  $P'$  和  $Q''$  进行滤波，去掉一些离群点，用以提高下一步 ICP 算法的精度；ICP 模块通过以 Angle-fixed filter 输出的变换关系为初始值对滤波后的两个点集进行迭代求解最佳的刚体变换关系，输出最终的变换关系  $T$  和匹配的分数  $S$ ， $T$  也是目标的位姿， $S$  是点云匹配误差的倒数，实验部分会具体介绍匹配误差。

### 4.2.2 Angle-fixed 4PCS 算法

介绍 Angle-fixed 4PCS 算法之前，首先先详细介绍一下 4PCS 算法，4PCS 算法是一个对 3D 点集全局匹配的算法，即使所给的两个 3D 点集有小的重叠，4PCS 都能给出较好的结果，并且对噪声有一定的鲁棒性。这种方法对初始位姿没有任

何要求,其核心是从3D点集中提取出所有与给定平面4-points近似全等的共面4-points,该算法时间复杂度为 $O(n^2 + k)$ ,其中n是3D点集中点的个数,k是提取出的4-points个数。4PCS使用十分广泛,并且引申出许多相关的变种(Corsini et al. 2013)。

4PCS算法流程:算法流程如7所示,该算法输入两个点集 $P$ 和 $Q$ ,还有距离参数 $\delta$ ,返回两个点集之间的刚体变换 $T$ 。4PCS基于以下事实:共面点集中定义的比例在仿射变换,包括刚体运动中保持不变。举例来说,定义点集 $X := \{\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}\}$ ,其中4个点不都在同一条直线上,设直线 $ab$ 和 $cd$ 相交于点 $\mathbf{e}$ ,定义两个比例:

$$\begin{aligned} r_1 &= \|\mathbf{a} - \mathbf{e}\| / \|\mathbf{a} - \mathbf{b}\| \\ r_2 &= \|\mathbf{c} - \mathbf{e}\| / \|\mathbf{c} - \mathbf{d}\| \end{aligned} \quad (4.1)$$

则在仿射变换下所定义的 $r_1$ 和 $r_2$ 均保持不变,如图4.3。如果曲面 $S_1$ 和 $S_2$ 匹配,

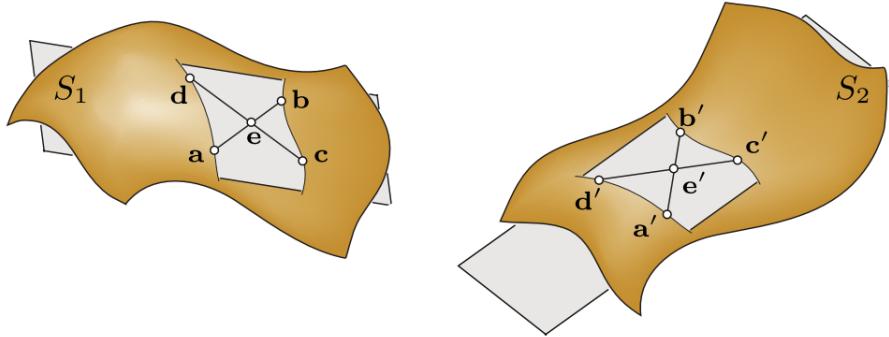


图 4.3 4-points 比例的仿射不变性

并且4-points共面基在重叠区域,则 $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$ 对应的四个点 $\mathbf{a}', \mathbf{b}', \mathbf{c}', \mathbf{d}'$ 也共面,并且

$$\begin{aligned} \|\mathbf{a}' - \mathbf{e}'\| / \|\mathbf{a}' - \mathbf{b}'\| &= r_1 \\ \|\mathbf{c}' - \mathbf{e}'\| / \|\mathbf{c}' - \mathbf{d}'\| &= r_2 \end{aligned} \quad (4.2)$$

4PCS算法另一个关键技术是使用了宽基(wide-base),相比于一般的基,宽基中基的长度更长,如图4.4所示,图片上半部分是使用宽基匹配的曲线,图片下半部分是使用一般的基匹配的曲线,显然,通过比较可以发现宽基相比普通基有更稳定的匹配结果,相关理论证明见文献(Goodrich et al. 1994)。

回到4PCS算法具体实现,算法的主体其实是一个RANSAC循环,每次循环首先会从点集 $P$ 中挑选共面的宽基 $B$ ,具体实现时,先从点集 $P$ 中随机选取3个点,然后在剩下的点中选取第四个点构成共面的四点,第四个点的选取尽可能使得每个点之间的距离最大(因为我们要使用宽基),并且与前3个点近似共面(显

---

算法 7: 4PCS 算法

---

**Input:** Point sets  $P$  and  $Q$ , measure level  $\delta$ **Output:** Rigid transform  $T$ 

```

1  $h \leftarrow 0;$ 
2 for  $i = 1$  to  $L$  do
3    $B \leftarrow \text{SELECTCOPLANARBASE}(P);$ 
4    $U \leftarrow \text{FINDCONGRUENT}(B, Q, \delta);$ 
5   forall 4-points coplannar sets  $U_i \in U$  do
6      $T_i \leftarrow$  best rigid transform that aligns  $B$  to  $U_i$  in the least square sense;
7     Find  $S_i \subseteq P$ , such that  $d(T_i(S_i), Q) \leq \delta$ ;
8      $k \leftarrow \arg \max_i \{|S_i|\};$ 
9     if  $|S_k| > h$  then
10        $h \leftarrow |S_k|;$ 
11        $T \leftarrow T_k;$ 
12   return  $T;$ 

```

---

```

13 def  $\text{FINDCONGRUENT}(B := \{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \mathbf{b}_4\}, Q, \delta):$ 
14    $d_1 \leftarrow \|\mathbf{b}_1 - \mathbf{b}_2\|;$ 
15    $d_2 \leftarrow \|\mathbf{b}_3 - \mathbf{b}_4\|;$ 
16   计算  $R_1 \equiv \{(\mathbf{p}_i, \mathbf{p}_j) \mid \mathbf{p}_i, \mathbf{p}_j \in Q\}$ , 使得  $\|\mathbf{p}_i - \mathbf{p}_j\| \in [d_1 - \delta, d_1 + \delta]$ ;
17   计算  $R_2 \equiv \{(\mathbf{p}_i, \mathbf{p}_j) \mid \mathbf{p}_i, \mathbf{p}_j \in Q\}$ , 使得  $\|\mathbf{p}_i - \mathbf{p}_j\| \in [d_2 - \delta, d_2 + \delta]$ ;
18   forall  $r_{1i} \in R_1$  do
19     计算与定量  $r_1$  和  $r_2$  相关的四个点  $\{\mathbf{e}_{1i}^1, \mathbf{e}_{1i}^2, \mathbf{e}_{1i}^3, \mathbf{e}_{1i}^4\}$ , 记  $\Pi(\mathbf{e}_{1i}^j) = r_{1i}$ ;
20     对点集  $\{\mathbf{e}_{1i}^j\}$  在  $\mathbb{R}^3$  空间建立 range tree 的数据结构;
21   forall  $r_{2i} \in R_1$  do
22     计算与定量  $r_1$  和  $r_2$  相关的四个点  $\{\mathbf{e}_{2i}^1, \mathbf{e}_{2i}^2, \mathbf{e}_{2i}^3, \mathbf{e}_{2i}^4\}$ , 记  $\Pi(\mathbf{e}_{2i}^j) = r_{2i}$ ;
23    $U' \leftarrow \emptyset;$ 
24   forall  $\mathbf{e}_{2i}^j$  do
25     在 range tree 中以  $\delta$  为领域检索点  $\mathbf{e}_{2i}^j$  附近的点, 对于每个检索到的
      点  $\mathbf{q}$ , 建立与  $B$  相对应的 4 个点的点集  $U' \leftarrow \{U', (\Pi(\mathbf{q}), \Pi(\mathbf{e}_{2i}^j))\}$ ;
26   return  $U'$ ;

```

---

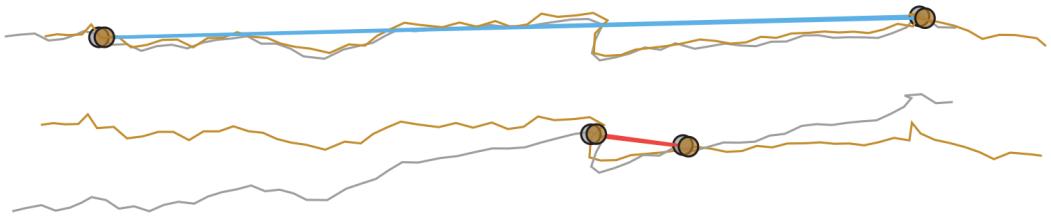


图 4.4 宽基的匹配稳定性

然由于噪声的存在,完全共面并不现实),但如果第四个点选取的过远也会出现问题,因为如果宽基超过两个点集的重叠区域则难以匹配,因此当选以最大距离取宽基造成误差变大时以  $f = 1, 0.5, 0.25, \dots$  的比率降低最大距离来选取宽基。

在点集  $P$  中选取好宽基  $B$  后, 算法下一步会在点集  $Q$  中通过 4-points 的仿射不变性找出所有与宽基  $B$  “全等”的基, 构成集合  $U$ 。在  $Q$  中选取基的方法见算法7中的 FINDCONGRUENT 函数, 函数首先使用基  $B$  中的点先定义两个仿射无关的比例, 如图4.5中左边的图所示。假设在点集  $Q$  中找到两点  $\mathbf{q}_1$  和  $\mathbf{q}_2$ , 并且

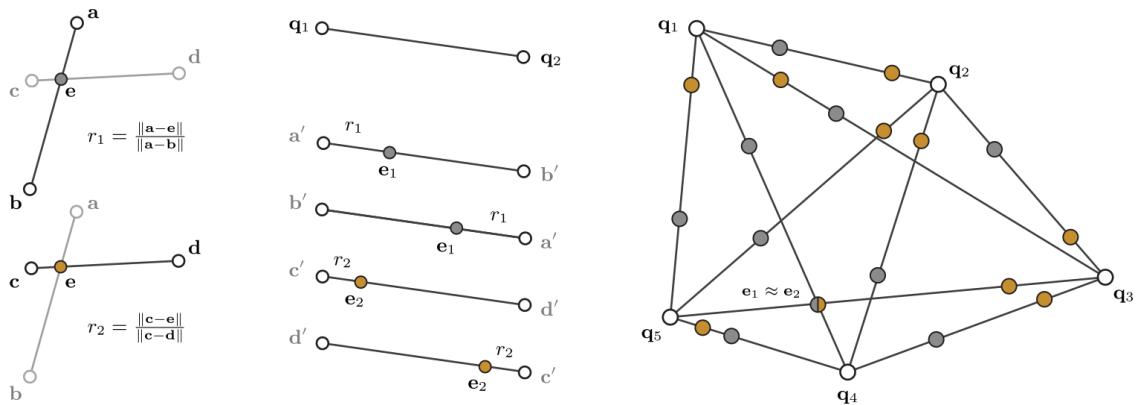


图 4.5 寻找近似“全等”的基示意图

$\|\mathbf{q}_1 - \mathbf{q}_2\| - \|\mathbf{a} - \mathbf{b}\| \leq \delta$ , 则点  $\mathbf{q}_1, \mathbf{q}_2$  有可能与点  $\mathbf{a}, \mathbf{b}$  对应, 则直线  $\mathbf{ab}$  与  $\mathbf{cd}$  相交的点  $\mathbf{e}$  的对应点可能为

$$\mathbf{e}_1 = \mathbf{q}_1 + r_1(\mathbf{q}_2 - \mathbf{q}_1) \quad (4.3)$$

或者

$$\mathbf{e}_1 = \mathbf{q}_2 + r_1(\mathbf{q}_1 - \mathbf{q}_2) \quad (4.4)$$

同理也可以根据  $\mathbf{c}, \mathbf{d}$  的对应点(设为  $\mathbf{q}_3, \mathbf{q}_4$ )求得  $\mathbf{e}$  的对应点

$$\mathbf{e}_2 = \mathbf{q}_3 + r_1(\mathbf{q}_4 - \mathbf{q}_3) \quad (4.5)$$

或者

$$\mathbf{e}_2 = \mathbf{q}_4 + r_1(\mathbf{q}_3 - \mathbf{q}_4) \quad (4.6)$$

则当  $\mathbf{e}_1 \approx \mathbf{e}_2$  时,  $\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3, \mathbf{q}_4$  就是我们所要找的一组与点  $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$  近似“全等”的基, 如图4.5中右边图中的  $\mathbf{q}_5, \mathbf{q}_3, \mathbf{q}_4, \mathbf{q}_1$ 。

具体实现时, 当我们在点集  $Q$  中找出了所有可能的  $\mathbf{e}_1$  和  $\mathbf{e}_2$  后, 找出其中近似相等的  $\mathbf{e}_1$  和  $\mathbf{e}_2$  可以通过 range 树 (Arya et al. 1998) 来实现, 对于大小为  $n$  的点集, range 树的建立时间复杂度为  $O(n \lg n)$ , 查询附近点的时间复杂度为  $O(\lg n + k)$ , 其中  $k$  是查询到点的个数。

在  $Q$  中找出所有与基  $B$  近似“全等”的基后, 下一步就是计算出最优的刚体变换  $T$ 。对于  $U$  中的每个基  $U_i$ , 我们可以利用最小二乘 (Horn 1987) 的思想计算  $B$  到  $U_i$  的刚体变换  $T_i$ 。得到刚体变换  $T_i$  后, 我们将点集  $P$  进行变换  $T_i$ , 然后对变换后的点集中的点在  $Q$  中查找最近点, 统计最近点距离小于  $\delta$  的个数  $S_i$ ,  $S_i$  也是评价  $T_i$  效果的分数, 分数越高的  $T_i$  就是我们要求的最优刚体变换  $T$ 。

**4PCS 算法时间复杂度:** 设输入的两个点集  $P, Q$  的大小分别为  $m, n$ 。算法中最耗时的部分是 FINDCONGRUENT 函数: 从点集  $Q$  中选取距离为  $d_1$  和  $d_2$  的点对, 其时间复杂度为  $O(n^2)$ , 然后建立和查询 range 树, 其复杂度为  $O(n^2 + k)$ , 其中  $k$  是满足条件的基个数, 因此 4PCS 算法整体的时间复杂度为  $O(n^2 + k)$ , 空间复杂度显然为  $O(n)$ 。

**4PCS 算法不足:** 仔细研究 4PCS 算法, 可以发现从点集  $Q$  中提取的基与  $B$  并不是全等的, 如图4.6所示, 将线段  $\mathbf{q}_1\mathbf{q}_2$  绕交点转动一定角度后便不再与原基全

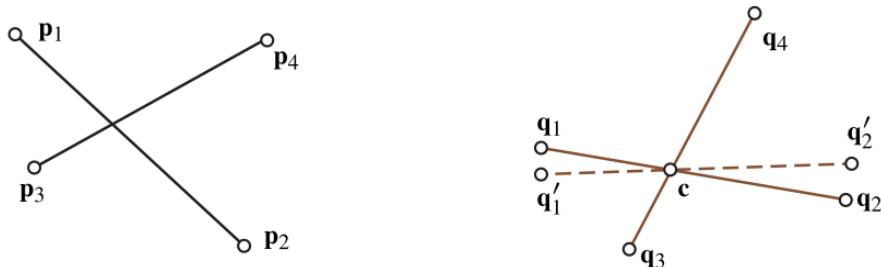


图 4.6 4PCS 中“全等”的基

等, 但是 4PCS 仍然会找出点  $\mathbf{q}'_1, \mathbf{q}'_2, \mathbf{q}_3, \mathbf{q}_4$  作为与  $\mathbf{p}'_1, \mathbf{p}'_2, \mathbf{p}_3, \mathbf{p}_4$  全等的基。这一缺点会导致 4PCS 算法需要更多的求解时间, 并且还有可能影响最终的匹配结果。因此, 对 4PCS 改进去除这些与原基不是全等的基很有必要。改进后的算法如算法8所示, 滤去不全等的基的算见其中的 FILTERCONGRUENT 函数。

#### 4.2.3 Outlier filter

从图 4.1其实可以看出目标点云有许多点并不属于目标物体, 尤其是检测结果没有 mask 只能使用 bbox 分割目标时, 不属于目标物体的离群点就尤其的多。

---

算法8: A4PCS 算法

---

**Input:** Point sets  $P$  and  $Q$ , distance tolerance  $\delta$ , angle tolerance  $\epsilon$

**Output:** Rigid transform  $T$

```

1  $h \leftarrow 0;$ 
2 for  $i = 1$  to  $L$  do
3    $B \leftarrow \text{SELECTCOPLANARBASE}(P);$ 
4    $U \leftarrow \text{FINDCONGRUENT}(B, Q, \delta);$ 
5    $U \leftarrow \text{FILTERCONGRUENT}(B, U, \epsilon);$ 
6   forall 4-points coplanar sets  $U_i \in U$  do
7      $T_i \leftarrow$  best rigid transform that aligns  $B$  to  $U_i$  in the least square sense;
8     Find  $S_i \subseteq P$ , such that  $d(T_i(S_i), Q) \leq \delta$ ;
9    $k \leftarrow \arg \max_i \{|S_i|\};$ 
10  if  $|S_k| > h$  then
11     $h \leftarrow |S_k|;$ 
12     $T \leftarrow T_k;$ 
13  return  $T;$ 

```

```

14 def FILTERCONGRUENT( $B := \{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \mathbf{b}_4\}, U, \epsilon$ ):  

15    $U' \leftarrow \emptyset;$ 
16    $a \leftarrow \overrightarrow{\mathbf{b}_1\mathbf{b}_2} \cdot \overrightarrow{\mathbf{b}_3\mathbf{b}_4};$ 
17   forall  $U_i := \{\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3, \mathbf{q}_4\} \in U$  do
18      $a' \leftarrow \overrightarrow{\mathbf{q}_1\mathbf{q}_2} \cdot \overrightarrow{\mathbf{q}_3\mathbf{q}_4};$ 
19     if  $a' \in [a - \epsilon, a + \epsilon] \cup [-a - \epsilon, -a + \epsilon]$  then
20        $U' \leftarrow \{U', U_i\};$ 
21   return  $U';$ 

```

---

因此,为了提高最终的匹配精度,除去这些离群点十分有必要,Outlier filter 模块的作用就是根据 Angle-fixed 4PCS 输出的初始刚体变换来去除点云数据中的离群点,从而提升下一步 ICP 算法匹配的精度,使最终估计的位姿精度提高。

Outlier filter 模块的核心算法如算法9所示。算法输入点集  $P$  和  $Q$ ,需要参数

---

#### 算法 9: Outlier filter 算法

---

**Input:** Point sets  $P$  and  $Q$ , Initial transform  $T$ , Distance tolerance  $\delta$

**Output:** Point sets  $P'$  and  $Q'$

```

1  $P' \leftarrow P;$ 
2  $Q' \leftarrow \emptyset;$ 
3 forall point  $\mathbf{p}_i \in P$  do
4    $\mathbf{p}_i \leftarrow T(\mathbf{p}_i);$ 
5 对点集  $P$  在  $\mathbb{R}^3$  空间建立 kd 树的数据结构;
6 forall point  $\mathbf{q}_i \in Q$  do
7   在 kd 树中检索出距离点  $\mathbf{q}_i$  最近的点  $\mathbf{p}$ ;
8    $d \leftarrow \|\mathbf{q}_i - \mathbf{p}\|;$ 
9   if  $d \leq \delta$  then
10     $Q' \leftarrow \{Q', \mathbf{q}_i\};$ 
11 return  $P', Q';$ 

```

---

初始刚体变换  $T$ ,以及允许的距离误差  $\delta$ ,由于点集  $P$  是由物体的 CAD 模型转换过来的,因此不对其进行滤波,只对点集  $Q$  进行离群点去除。具体方法是,首先使用  $T$  对点集  $P$  进行刚体变换;然后对变换后的点集建立 kd 树,建立 kd 树的目的是为了快速在点集  $P$  中找到距离某点最近的点,其查找的时间复杂度为  $O(kN^{1-1/k})$ ,其中  $k$  是所建立 kd 树的维数,显然对于三维空间中点集为 3,  $N$  是建立的 kd 树的节点个数;建立好 kd 树后,对点集  $Q$  中的每个点在 kd 树中找到与之距离最近的点,如果两点间的距离大于所设的参数  $\delta$ ,则在点集  $Q$  中去除该点。实际运行 Outlier filter 算法的效果如图4.7所示。

#### 4.2.4 ICP 算法

ICP(Iterative Closest Point)算法,即最近点迭代算法,是最为经典的数据配准算法。ICP 算法本质上是基于最小二乘法的最优配准方法。该算法重复进行选择对应关系点对,计算最优刚体变换,直到满足正确配准的收敛精度要求。由于 ICP 算法是一种迭代算法,因此只要时间允许便可以获取足够精度的解,但也正因为

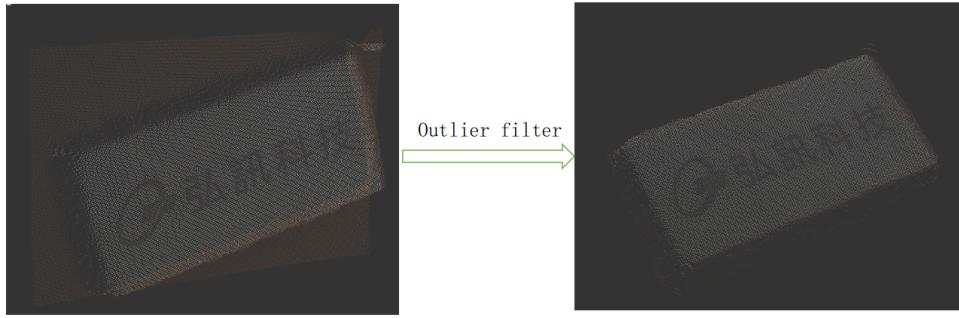


图 4.7 Outlier filter 效果图

如此, ICP 也容易陷入局部最优解。本文充分考虑了 ICP 算法的这个特点, 通过 Angle-fixed 4PCS 算法给出初始的刚体变换来避免 ICP 算法陷入局部最优解, 同时通过迭代来提高最后输出的刚体变换精度。下面介绍一下 ICP 算法的基本原理。

给定两个点集  $P_n := \{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_n\}$  和  $Q_m := \{\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_m\}$ , 以及初始旋转变换  $R$  和平移变换  $t$ , 以及迭代结束额距离误差  $\delta$ , ICP 算法迭代步骤如下:

- 步骤 1: 根据当前  $R$  和  $t$ , 对于点集  $P_n$  中的每个点在  $Q_m$  中找出距离最近的点, 构成点集  $Q_n$ ;
- 步骤 2: 计算  $P_n$  和  $Q_n$  之间的距离的均方根误差:

$$E(R, t) = \frac{1}{n} \sum_{i=1}^n \| \mathbf{q}_i - R\mathbf{p}_i - t \| \quad (4.7)$$

通过奇异值分解求得使得  $E(R, t)$  最小的  $R'$  和  $t'$ ;

- 步骤 3: 如果  $E(R, t) < \delta$ , 结束迭代; 否则  $R \leftarrow R'$ ,  $t \leftarrow t'$ , 跳转至步骤 1。

ICP 算法的迭代过程还是相对来说十分简单的, 唯一需要思考一下的是如何求得最小化  $E(R, t)$  的  $R'$  和  $t'$ , 通过奇异值分解求解  $R'$  和  $t'$  的方法如下:

首先, 记

$$\begin{cases} P'_n &= \{\mathbf{p}_i - \mu_p \mid \forall \mathbf{p}_i \in P_n\} := \{\mathbf{p}'_i\} \\ Q'_n &= \{\mathbf{q}_i - \mu_q \mid \forall \mathbf{q}_i \in Q_n\} := \{\mathbf{q}'_i\} \end{cases} \quad (4.8)$$

其中

$$\begin{cases} \mu_p &= \frac{1}{n} \sum_{i=1}^n \mathbf{p}_i \\ \mu_q &= \frac{1}{n} \sum_{i=1}^n \mathbf{q}_i \end{cases} \quad (4.9)$$

再另

$$W = \sum_{i=1}^n \mathbf{p}'_i \mathbf{q}'_i \quad (4.10)$$

然后对矩阵  $W$  进行奇异值分解:

$$W = U\Sigma V^T \quad (4.11)$$

则

$$\begin{cases} R' &= UV^T \\ t' &= \mu_p - R'\mu_q \end{cases} \quad (4.12)$$

### 4.3 点云匹配实验

为了评价所设计的算法 A4PCS-ICP, 考察其匹配精度以及时间性能, 本文设计了位姿估计的实验, 不但考察了 A4PCS-ICP 的性能, 还与其他一些算法作比较, 验证了 A4PCS-ICP 算法的匹配的精确度。

#### 4.3.1 实验内容

实验在多了点云匹配的数据集上测试了 A4PCS-ICP 算法和一个基于局部特征描述子和 RANSAC 的算法 LD-RANSAC(Li et al. 2005)。数据集中的点云数据有从激光扫描获取的、深度摄像头采集的、双目摄像头重构的, 并且里面的模型也多种多样, 有人造物、纹理缺少的物体, 光滑的物体, 粗糙的物体等。与 A4PCS-ICP 对比的 LD-RANSAC 算法使用了基于 spin-image 的描述子 (Johnson et al. 1999), LD-RANSAC 的具体实现直接使用了 PCL(Point Cloud Library)中的相关代码, 人工配置了算法的参数以达到较好的匹配效果。

实验主要考察点云匹配的误差以及算法的运行时间, 点云匹配的误差定义为两个点云之间的 RMS 误差:

$$S = \sqrt{\frac{1}{N} \sum_{i=1}^N \min_{p_j \in T(P')} \|q_i - p_j\|^2} \quad (4.13)$$

显然 RMS 误差越小, 点云的匹配效果越好, 精度越高。

#### 4.3.2 实验结果

为了考察算法的精度, 通过对输入数据增加高斯噪声, 来观察算法在不同大小噪声下的 RMS 误差, A4PCS-ICP 算法和 LD-RANSAC 算的 RMS 误差随高斯噪声方差的变化曲线如图4.8所示。图中可以看出 A4PS-ICP 比 LD-RANSAC 算法有更小的误差, 并且 A4PCS-ICP 对噪声也具有高强的鲁棒性。

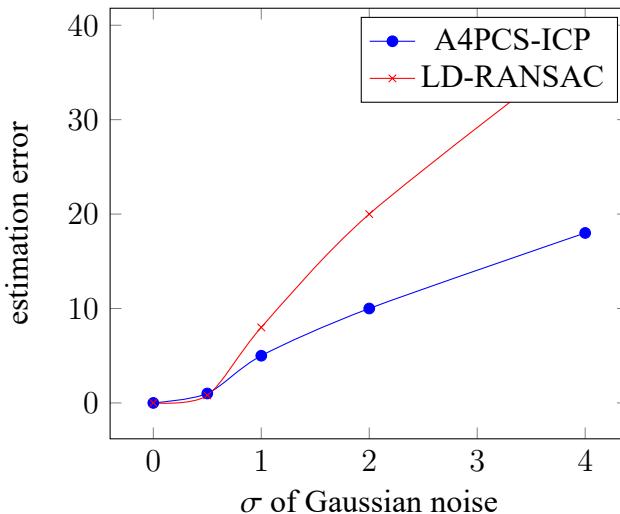


图 4.8 匹配误差随噪声变化曲线

为了考察算法的时间性能,通过对输入数据进行降采样(uniform sampling),变化 uniform sampling 的采集间距来变化输入点云的数量大小,不同点云数量大小下算法的运算时间的变化曲线如图4.9所示。图中可以看出 A4PCS-ICP 的运算

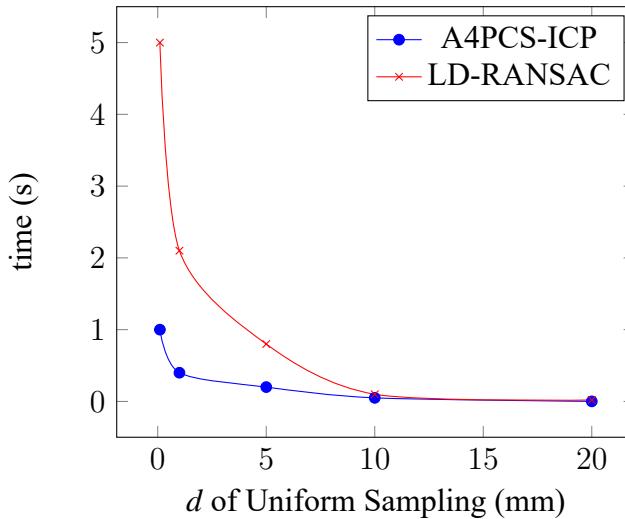


图 4.9 匹配时间随点云数量大小变化曲线

时间比 LD-RANSAC 少,并且随着点云数量的增加其运算时间的差距越来越大。

综上,A4PCS-ICP 算法相比 LD-RANSAC 算法具有更高的匹配精度,对噪声的鲁棒性也更强,并且算法的时间复杂度也更小。

#### 4.4 本章小结

本章在 4PCS 算法的基础上,针对其缺点,提出了 A4PCS 算法,减少了算法的运算时间;然后为了提高匹配精度,将 A4PCS 算法与 ICP 算法相结合

构成 A4PCS-ICP 算法。最后通过实验与 LD-RANSAC 算法相比较，实验表明 A4PCS-ICP 算法相比 LD-RANSAC 算法具有更高的匹配精度，对噪声的鲁棒性也更强，并且算法的时间复杂度也更小。

## 第 5 章 3D 目标位姿估计算法

本章提出了一种 3D 目标位姿估计算法 3D-MRAI (3D Mask R-CNN & A4PCS-ICP)，该算法根据所提供目标的 CAD 模型，可以在 RGB-D 图中检测出目标，并给出目标的位姿。3D-ODPE 算法主要基于第 3 章中基于 RGB-D 图的目标检测算法 3D Faster R-CNN/3D Mask R-CNN，以及第 4 章中的点云匹配算法 A4PCS-ICP，通过将这两个算法结合，分两步计算出 RGB-D 图中目标的位姿。为了评价 3D-MRAI 算法的性能，本章还设计了相关实验，并与同类算法做了比较。

### 5.1 3D-MRAI 框架设计

3D-MRAI 主要解决三维空间中的目标检测和位姿估计问题，根据 RGB-D 图像，给出图像中目标的种类和其在三维空间中的位姿，区别与常见的在 2D 图像中的目标检测（如第 3 章中的算法）给出的是目标的种类和其在图像坐标中的 bounding box 或者 mask。给出目标在三维空间中的位姿的意义十分巨大，尤其在机器人领域中，图像层面的结果往往难以满足要求，但难度也很大。

一些给出 3D 目标位姿的传统算法，如 3DMatch(Zeng et al. 2016)，3DMatch 通过匹配局部几何特征来计算目标的位姿，缺点是对采集的 3d 数据质量要求很高，往往需要使用激光采集，因此整个识别过程的时间很久；通过 SIFT 描述子来匹配目标位姿 (Dias et al. 2015) 也是一种方法，但其对纹理较少的物体往往难以匹配，效果很差；另外如 LINEMOD(Hinterstoisser et al. 2012) 和 MOPED(Collet et al. 2011) 这些位姿估计框架，在某些情况下如目标在平整的桌面上并且光照条件较好的情况下才能取得满意的效果。因此，急需一种鲁棒性较强，精度较高，计算时间较短的 3D 目标位姿估计算法。

3D-MRAI 的框架如图 5.1 所示，从图 5.1 可以看出算法的输入是 RGB 图像、深度图，以及目标物体的 CAD 模型，输出是图像中检测到的目标的位姿。3D-MRAI 的核心部分是 3D Faster/Mask R-CNN 和 A4PCS-ICP 算法，3D Faster/Mask R-CNN 以及在第 3 章详细介绍过，A4PCS-ICP 也在第 4 章详细介绍过了。因此，3D-MRAI 估计目标的位姿流程上也是分为两步 (two-stage)：

- **目标检测：**利用 3D Faster/Mask R-CNN 检测目标，得到目标 BBox 或者 Mask
- **点云匹配：**将 CAD 模型与目标检测对应的点云进行匹配，得到目标位姿

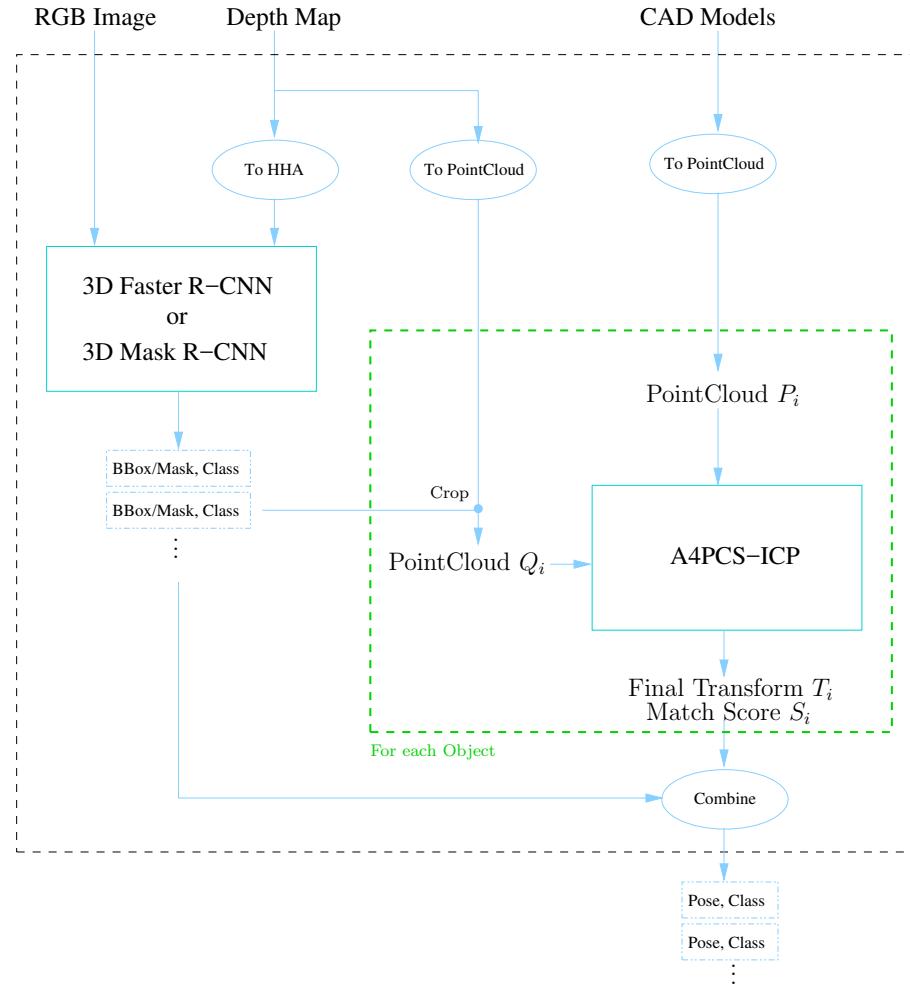


图 5.1 3D-MRAI 算法框架

## 5.2 3D-MRAI 具体实现

对与 3D-MRAI 输入的 RGB 图像和深度图,可以由对偶 RGB-D 获得,具体见第 2 章。目标物体的 CAD 模型也容易获得,对于工厂中的工件,往往是有其 CAD 模型的,对于一般物体,可以通过 3D 扫描仪重构出来,当然也可以使用本文设计的对偶 RGB-D 相机采集重构出来。

获得目标物体的 CAD 模型后，为了方便后续需处理，我们需要将其转换为点云。具体如何转换的话，基本思想是参考 Uniform Sampling 算法，Uniform Sampling 算法的核心思想是以 3D 栅格中所有点的质心代替这些点，从而达到降采样。类似地，对于 CAD 模型也建立 3D 栅格，但由于无法获得 3D 栅格总所有点，因此判断 CAD 模型是否穿过 3D 栅格，如果穿过 3D 栅格，则在该 3D 栅格中心出增加一个点。显然 3D 栅格的边长越大，转换后的点云数量越小，精度越低，考虑到所使用相机生成点云的精度，因使 CAD 模型转换后的点云的精度与相机采集的点云的精度近似，实际取 3D 栅格边长为 1mm，一个实际工件的 CAD 模型

和以1mm为边长进行采样转换后的模型点云如图??所示。此外，还需要将深度

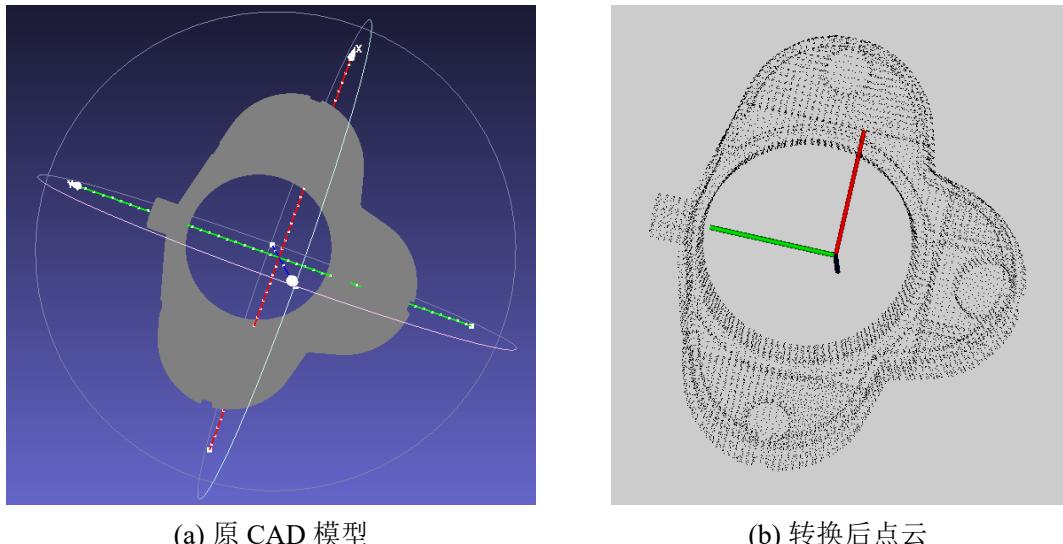


图5.2 CAD模型和转换后的点云

图转换为点云，这一步则相对简单，只要通过深度摄像头的内参矩阵反投影到三维空间即可，详细见第2章。

对于3D Faster/Mask R-CNN的输入，还需要将深度图转换为HHA，具体见3.1.2小节。3D Faster/Mask R-CNN模块的实现，由于一个深度神经网络，只要训练好后将网络模型导出成Tensorflow的pb文件，然后此处导入该网络模型，给定输入，网络输出便是图片中目标的BBox/Mask和Class。

得到目标物体的BBox/Mask后，需要从深度图对应的点云中抠出目标点云，由于深度图转换的点云是有序的，因此BBox/Mask在深度图中的索引坐标与深度图转换的点云的索引坐标是一致的，只要将点云中对应的点提取出来就行，并滤去无效的点然后适当降采样即可，尽管滤波和降采样之后的目标点云是无序的，但后续匹配算法并不需要输入点云有序，而且降采样后点云数量减少，将会减少后续匹配算法的时间。

裁剪得到目标物体的点云 $Q_i$ 后，找出对应物体的CAD模型转换的点云 $P_i$ ，将 $P_i$ 和 $Q_i$ 输入到A4PCS-ICP模型，即可得到CAD模型到目标点云之间的刚体变换 $T_i$ ，由于CAD模型坐标系与相机坐标系重叠，因此将矩阵 $T$ 转换为 $X, Y, Z, r, p, y$ 就是目标点云在相机坐标系下的位姿Pose，最后将所有匹配得到的结果与目标检测的结果组合，并滤去匹配或检测分数较低的结果。详细的算法流程如算法10所示。

---

**算法 10: 3D-MRAI 算法**

---

**Input:** RGB Image  $I$ , Depth Map  $D$ , CAD Models  $M$ **Output:** Set of Pose and Class  $Res$ 

```

1  $Res \leftarrow \emptyset;$ 
2  $P \leftarrow \emptyset;$ 
3 forall  $M_i \in M$  do
4    $P \leftarrow \{P, CAD2PointCloud(M_i)\};$ 
5  $H = Depth2HHA(D);$ 
6  $Q = Depth2PointCloud(D);$ 
7  $Mask, Class \leftarrow 3DMAS KRCNN(I, H);$            // Same with
     3DFASTERCNN
8 forall  $m_i \in Mask, c_i \in Class$  do
9    $Q_i \leftarrow Crop(Q, m_i);$ 
10   $P_i \leftarrow P(c_i);$ 
11   $T_i, S_i \leftarrow A4PCSICP(P_i, Q_i);$ 
12  if  $S_i > S_{min}$  then
13     $Res \leftarrow \{Res, [T_i, c_i]\};$ 
14 return  $Res$ 

```

---

### 5.3 3D 目标位姿估计实验

为了评价所提出的 3D-MRAI 的性能, 设计了 3D 目标位姿估计的实验, 并与文献 (Hinterstoisser et al. 2012) 所提出的基于 LINEMOD 算法的 3D 目标位姿估计框架相比。

#### 5.3.1 数据集

实验所使用的数据集是 workpiece 数据集, 在第3.3.1小节中已经部分介绍过了, 该数据集是在实验室采集的三类物体, 第 3 章中实验所用 workpiece 数据集中的 ground truth 是物体的种类、BBox 和 Mask, workpiece 数据集中测试集中的图片的 ground truth 除了物体的种类、BBox 和 Mask, 还有物体的位姿, 物体的位姿是通过在物体旁固定标定板采集的。具体方法是, 通过固定标定板在目标物体旁, 我们可以记录标定板到目标的刚体变换关系  $T_1$ , 然后我们通过彩色摄像头可以

检测出标定板的位姿  $T_2$ , 则物体的位姿可以通过下式得到

$$T = T_2 T_1 \quad (5.1)$$

### 5.3.2 实验内容

为了有效的评价 3D-MRAI 算法, 我们先定义一个合适的评价指标姿态误差:

$$m = \text{avg}_{\mathbf{x} \in M} \|(\mathbf{R}\mathbf{x} + \mathbf{t}) - (\tilde{\mathbf{R}}\mathbf{x} + \tilde{\mathbf{t}})\| \quad (5.2)$$

其中  $M$  表示算法运行结果得到的物体种类对应的 CAD 模型转换得到的点云,  $R$  和  $t$  分别表示从 ground truth 物体位姿分解得到的旋转变换和平移变换,  $\tilde{R}$  和  $\tilde{t}$  分别表示从算法运行结果得到的物体位姿分解得到的旋转变换和平移变换。显然, 如果算法运行结果和 ground truth 越接近, 所定义的姿态误差就越小。对于一些对称的物体(如圆柱体的被子), 显然不同角度下相机看到的目标物体可能近似, 会造成算法运行的结果正确的情况下与 ground truth 相差很大, 造成姿态误差很大, 与我们所定义的评价指标的宗旨相违背。因此, 针对一些对称的物体, 重新定义姿态误差为

$$m = \text{avg}_{\mathbf{x}_1 \in M} \min_{\mathbf{x}_2 \in M} \|(\mathbf{R}\mathbf{x}_1 + \mathbf{t}) - (\tilde{\mathbf{R}}\mathbf{x}_2 + \mathbf{t})\| \quad (5.3)$$

此外, 如果  $k_m d > m$ , 我们就认为目标物体准确检测到了, 并且估计的位姿正确, 其中  $d$  是目标物体对应模型的直径,  $k_m$  是系数。因此, 还可以定义一个正确检测目标并正确估计目标位姿的准确率。

实验在 workpiece 数据集的测试集上分别运行了 3D-MRAI 和文献 (Hinterstoisser et al. 2012) 中的基于 LINEMOD 算法的检测框架, 运行实验的计算机有两块 Intel(R) Xeon(R) E5-2683 v3(2.00GHz) 的 CPU, 4 块 TITAN X GPU, 由于 3D-MRAI 有深度神经网络所以使用了一块 GPU 和一块 CPU, Hinterstorisser 等人的算法不需要 GPU, 只使用了一块 CPU。

### 5.3.3 实验结果

分别统计 3D-MRAI 和 LINEMOD 在测试集上的运行结果, 变化系数  $k_m$  统计算法的准确率如表5.1所示。表中  $k_m$  从 5% 变化到 15%, 表示物体直径的占比,  $k_m$  越大, 允许的姿态误差就越大, 因此准确率就越高。实际实验时, 发现  $k_m \approx 10\%$  时基本上肉眼可以看出匹配的姿态误差。将表5.1绘制成图如5.3所示, 从图中可以发现本文所提出的 3D-MRAI 算法的准确率大大超过了 Hinterstorisser 等人的

$k_m [\%]$	5	7	9	11	13	15
Hinterstoisser et al.	75.63	83.84	89.13	93.48	96.83	98.12
<b>3D-MRAI</b>	95.12	97.35	98.10	98.69	99.22	100.00

表 5.1 3D-MRAI 和 Hinterstoisser 等人的算法准确率

算法,在  $k_m = 13\%$  是 3D-MRAI 算法的准确率已经接近 100% 了,以肉眼可以看出匹配的姿态误差  $k_m = 9\%$  为标准时 3D-MRAI 算法的准确达到了 98.10%,比 Hinterstoisser 等人提出的算法高了大约 9 个百分点。

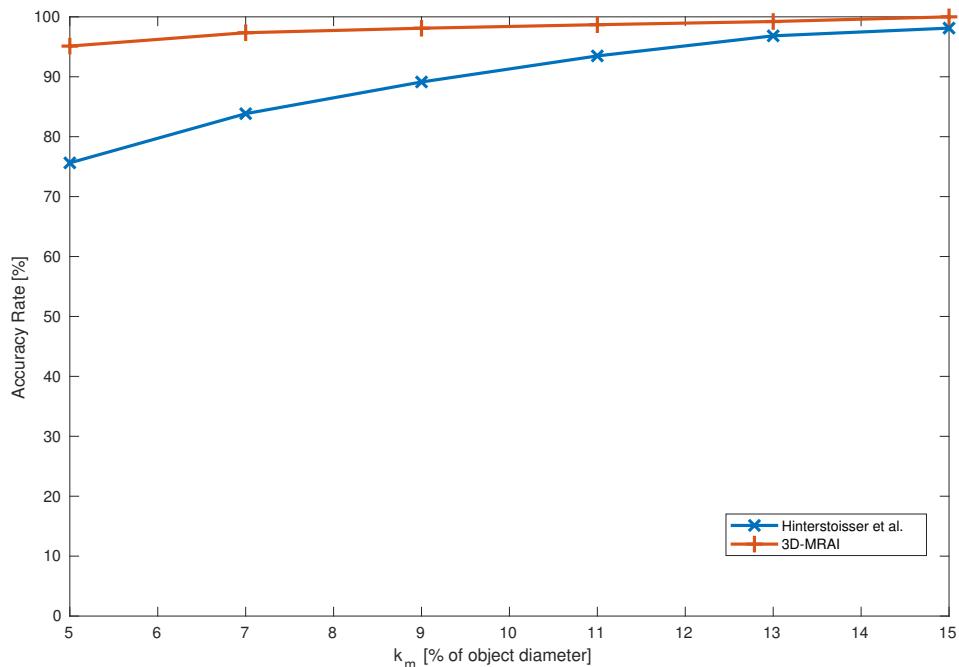


图 5.3 3D-MRAI 和 Hinterstoisser 等人的算法准确率曲线

为了进一步观察算法给出位姿的精度,取  $k_m = 9\%$  时 3D-MRAI 准确检测的例子,将目标物体的位姿转换为  $X, Y, Z, r, p, y$  六个直观的变量三维位置和姿态欧拉角,然后于 ground truth 相比较,得到算法结果在距离和角度上的误差的频率直方图如图5.4所示。从图中可以看出 3D-MRAI 正确检测和估计目标位姿的情况下,在  $X, Y, Z$  方向下的位置误差大部分分布在  $0 \sim 1mm$  之间,其距离精度在  $1mm$  左右;在  $r, p, y$  三个角度下的角度误差也大部分分布在  $0 \sim 1deg$  之间,其角度精度在  $1deg$  左右。统计图5.4中的数据,可以算出距离误差和角度误差的均值和方差为:

$$\begin{aligned} e_d &= 0.82 \pm 0.21mm \\ e_a &= 0.91 \pm 0.29deg \end{aligned} \tag{5.4}$$

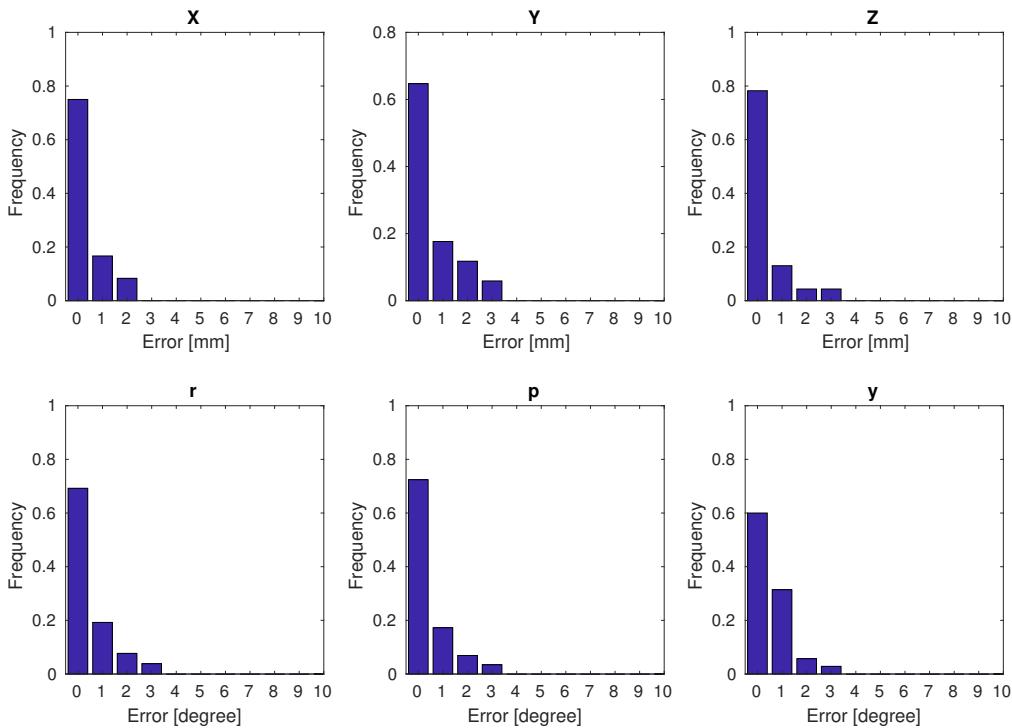


图 5.4 3D-MRAI 算法精度

除了关心算法的准确率和精度，算法的运行时间也是我们所关心的。在实验所用的计算机上，本文所设计的 3D-MRAI 算法与 Hinterstoisser 等人设计的基于 LINEMOD 算法的平均帧率如表 5.2 所示。表中可以看出 3D-MRAI 的 FPS 相

Hinterstoisser et al. 3D-MRAI		
FPS	7.6	2.2

表 5.2 3D-MRAI 和 Hinterstoisser 等人的算法准确率

比 Hinterstoisser 等人的算法的 FPS 相对较低，由于使用了深度神经网络相关的算法，涉及到较大的计算量，因此较低的 FPS 也在情理之中。

综上所述，在  $k_m = 9\%$  时，3D-MRAI 算法的准确率为 98.10% 左右，其估计的位姿的距离精度为  $0.82 \pm 0.21\text{mm}$ ，角度精度为  $0.91 \pm 0.29\text{deg}$ 。

## 5.4 本章小结

本章介绍了一种基于 RGB-D 图像的 3D 目标检测和位姿估计算法 3D-MRAI，3D-MARI 通过结合 3D Faster/Mask R-CNN 和 A4PCS-ICP 算法实现对目标位姿的估计。并通过实验与 Hinterstoisser 等人的算法比较，得出 3D-MRAI

算法具有更高的检测准确率，并且位姿精度也更高，但在时间性能上略逊于 Hinterstoisser 等人的算法。

## 第6章 算法应用——Bin-Picking

本章主要介绍 3D-MRAI 算法的实际应用——Bin-Picking。首先介绍一下 Bin-Picking 相关背景,以及近些年的具体研究与发展;然后详细介绍基于 3D-MRAI 算法所开发的一套解决 Bin-Picking 问题的视觉系统,包括硬件的选型、开发环境、系统的框架设计以及算法的具体实现;最后介绍了针对所开发的 Bin-Picking 视觉系统,进行了随机抓取的实验,测试了系统抓取的成功率以及系统的抓取速度。

### 6.1 Bin-Picking 背景与现状

使用机器人分拣散乱的工件的问题,在学术上我们称之为 Bin-Picking, Bin-Picking 并不是一个崭新的问题,学术上对这个问题的研究以及有了五十多年的历史。典型的 Bin-Picking 系统主要包括三部分:机器人、视觉检测模块和计算机控制模块。其中,视觉检测模块是整个 Bin-Picking 系统的核心部分,通过视觉检测模块对存放散乱工件的物料箱进行分析,获取目标工件的位姿,计算机控制模块根据视觉检测模块的检测结果规划机器人的运动路径,然后机器人执行完成工件的抓取。

传统的 Bin-Picking 中检测估计目标工件的算法大致可以分为两类:一类是基于特征匹配的算法,另一类是基于模板匹配的算法。基于特征匹配的算法,通过某些特征描述目标工件,如边角、空洞等特征,然后通过分析特征在空间中的旋转变换和平移变换来估计目标零件的位姿。这一类方法受工件纹理或者结构以及传感器的精度影响很大。另外一类基于模板匹配的方法的精度受限与模板的数量,要获得较高的精度就需要大量的模板,而大量的模板会造成算法运行时间过长。

工业上用于解决 Bin-Picking 问题的视觉系统也有许多,如图6.1所示。日本的 Fanuc 公司推出了基于 iRVision 的 Bin-Picking 系统,该系统通过四个相机进行三维视觉重建,然后进行目标定位 (Connolly 2007)。德国的 ISRA Vision 公司推出了 3D Shape Scan 系统,丹麦的 Scape Technologies 公司推出了 Scape-Tech Discs 系统,德国的 Sick 公司退出了 PLB-500 系统,诸如类似的 Bin-Picking 系统还有许多,这些 Bin-Picking 系统的价格大多二十万以上,并且抓取成功率和速度也往往难以满足客户需求,因此工业上还是缺少成熟的、价格便宜的、抓取成功率高、速

度快的 Bin-Picking 解决方案。

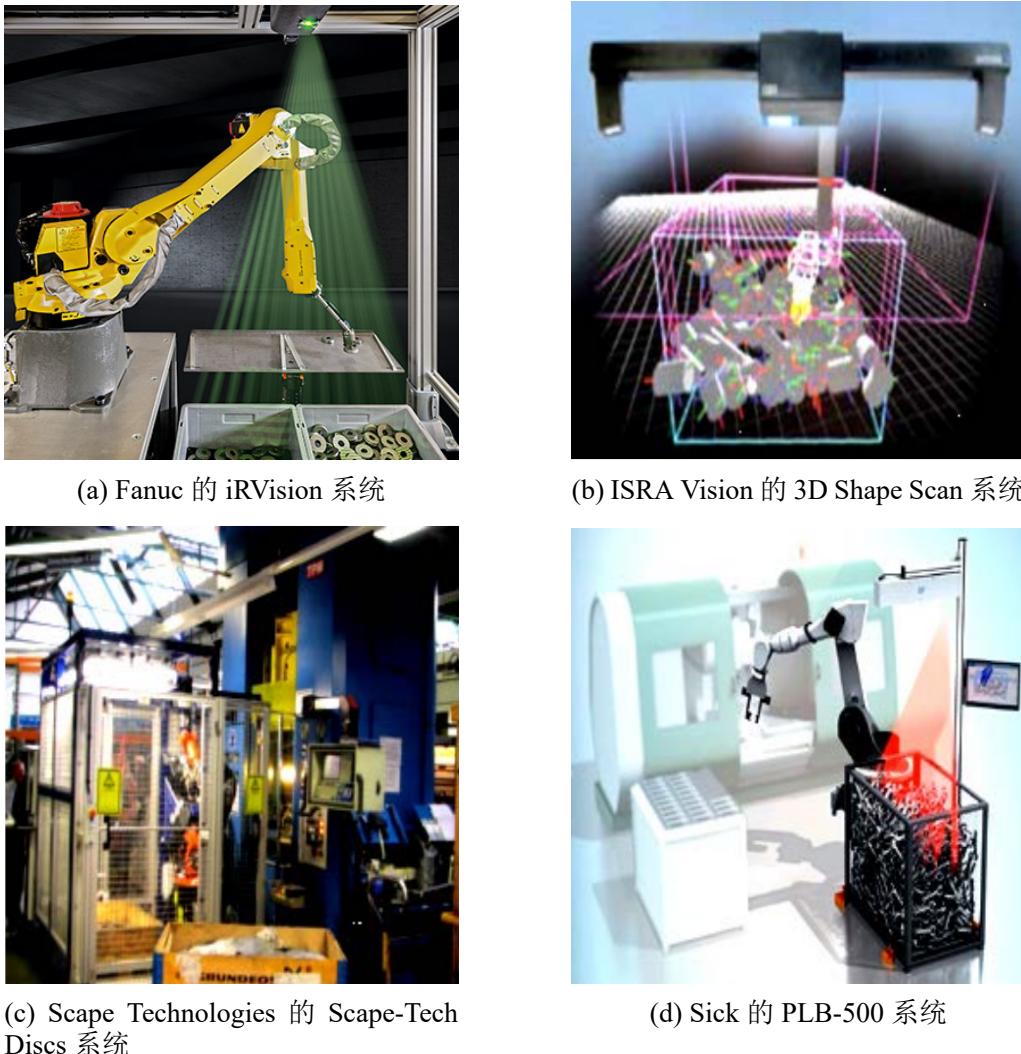


图 6.1 工业上典型的 Bin-Picking 解决方案

随着近几年一些高性价比的 3D 相机的出现,如微软的 Kinect 系列、Intel 的 RealSense 系列,加上近几年深度学习的巨大发展,使得开发一种性价比高的、抓取成功率高的、速度快的 Bin-Picking 系统成为可能。但尽管深度学习在计算机视觉领域(Computer Vision)有了大量的研究,但在机器人感知(Robot Perception)领域的应用还比较少,因此本文将深度学习在计算机视觉领域内的成果通过一些改进引入到 Robot Perception 领域,再结合传统的全局点云匹配算法,剔除了 3D-MRAI 算法,可以用于解决 Bin-Picking 相关问题。此外,本文基于 Intel 的 RealSense SR300 相机所提出的对偶 RGB-D 相机构建也为整个 Bin-Picking 系统提供了高性价比的相机解决方案。

## 6.2 基于 3D-MRAI 的随机分拣系统

### 6.2.1 系统硬件设计

本文所设计的基于 3D-MRAI 的随机分拣系统的硬件系统如图6.2所示。从

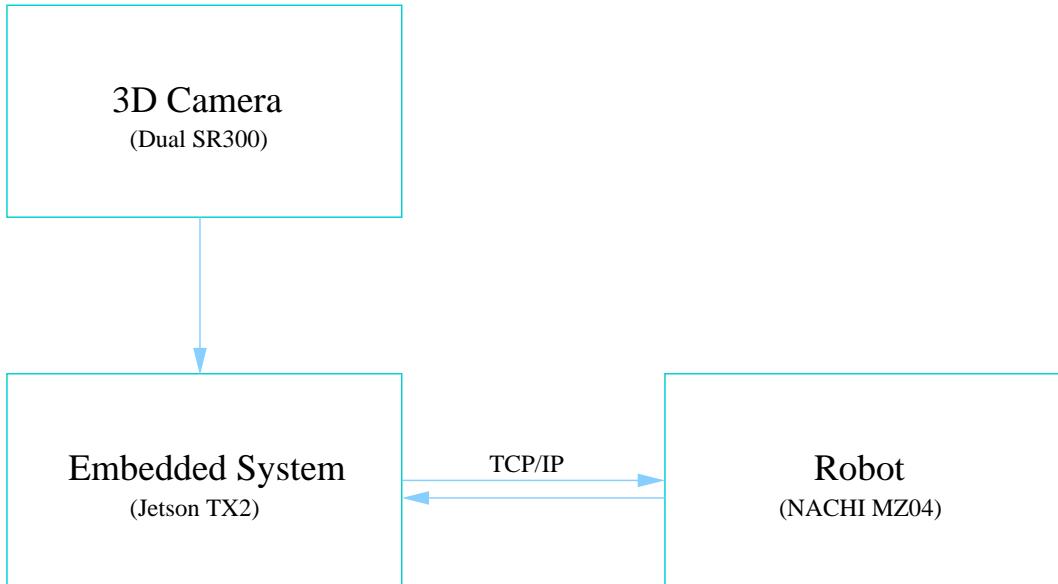
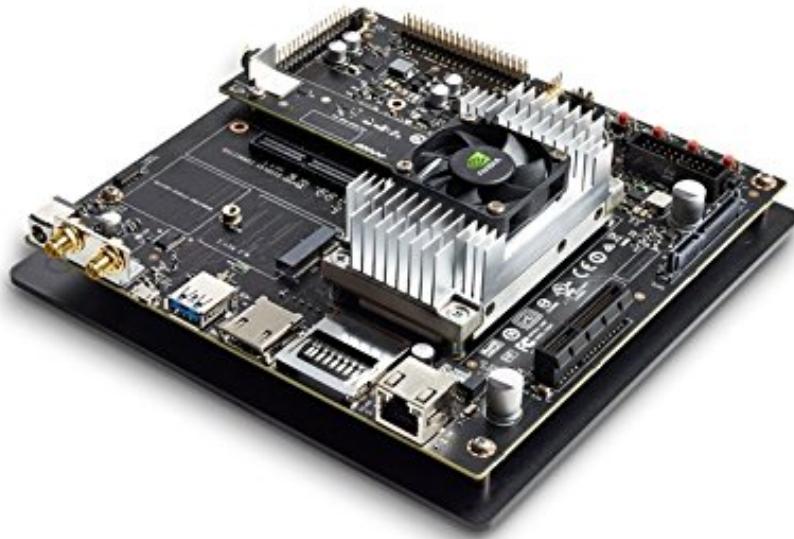


图 6.2 基于 3D-MRAI 的随机分拣系统硬件框架

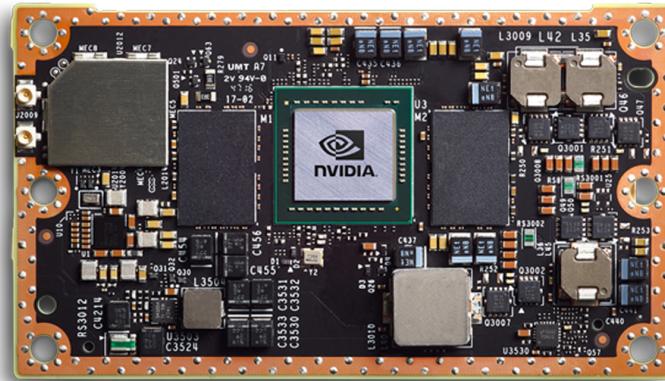
图6.2可以看出,所设计的随机分拣系统的硬件系统由三个部分构成: 相机模块、嵌入式计算模块以及机器人模块。对于相机模块,根据 3D-MRAI 算法的输入,需要相机能采集 RGB-D 图像,并且考虑整个系统的响应时间以及价格因此,希望相机模块的采集时间尽可能短,性价比尽可能高,因此选用了以结构光为原理的 3D 相机,并根据第 2 章所设计的对偶 RGB-D 相机,用两个 SR300 相机构成了对偶 RGB-D 相机模块。

嵌入式计算模块选用了搭载了 NVIDIA 公司的 Jetson TX2 模块的嵌入式计算机,如图6.3(a), Jetson TX2 如图6.3(b)所示。由于 3D-MRAI 算法使用了深度神经网络,因此所选用的计算机最好要搭载一块 GPU,当然由于模型的训练可以在服务器上完成,只需要在选用的计算机上跑模型的 Interface,因此其 GPU 性能也不需要特别好。另外,考虑到系统需要长时间运行,因此选用了低功耗的嵌入式计算机。所选用的搭载 Jetson TX2 模块的嵌入式计算机拥有一块 Pascal 架构的 GPU, 256 个 CUDA cores, CPU 是 HMP Dual Denver 加四块 ARM A57, 内存 8G (LPDDR4), 还拥有 1 Gigabit Ethernet, 802.11ac WLAN 以及 Bluetoothd, 在系统计算资源、功耗以及通信上完全满足整个 Bin-Picking 系统的要求。

机器人模块选用了 NACHI 的六轴机械臂 MZ04, 如图6.4所示。由于一般正常的随机分拣系统所要抓取的工件各种位姿都有,意味着所要抓取的工件有六个



(a) 搭载 Jetson TX2 模块的嵌入式计算机



(b) Jetson TX2 模块

图 6.3 嵌入式计算模块

自由度,因此所选用的机器人末端至少也要有六个自由度,不然难以完成各种姿态工件的抓取任务,因此选用了工业上常见的六轴机械臂,至于为何选用 NACHI 的 MZ04 这个型号,是出于合作方的需要,并不由个人意志决定,当然何种机械臂也不是本文的重点,所设计的视觉算法对机械臂也没什么特殊的要求,因此此处不作详细介绍。

实际搭建随机分拣系统环境如图6.5所示,图中相机固定在支架上,与机械臂构成了 eye-to-hand 的形式,当然也可以将相机固定在机械臂末端构成 eye-in-hand 形式,两种固定相机的形式略有不同,但对视觉识别算法那没有影响,只与控制流



图 6.4 六轴机械臂 NACHI MZ04

程和相机与机器人之间的标定有关系,后文会具体介绍到。嵌入式计算机通过相机采集物料箱内的图像,然后运行基于 3D-MRAI 的视觉系统,得出目标工件位姿,然后规划机械臂路径,通过 TCP/IP 通信,控制机械臂完成抓取任务,并根据机械臂的运动状态控制整个系统的流程。

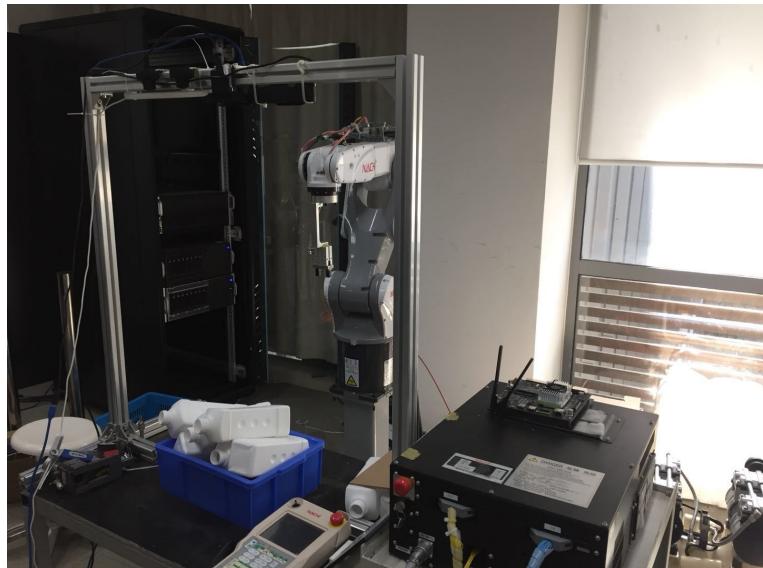


图 6.5 实际搭建随机分拣系统环境

### 6.2.2 系统软件设计

需求分析: 基于 3D-MRAI 的随机分拣系统的软件运行在所采用的嵌入式计算机上,主要需要以下几点功能:

- 图像的采集与处理
- 与机器人的通信
- 3D-MRAI 算法的实现
- 机器人相关的处理
- 可视化界面

针对以上几点需求,整个软件分为五个模块,如图6.6所示。Camera 模块主要实现

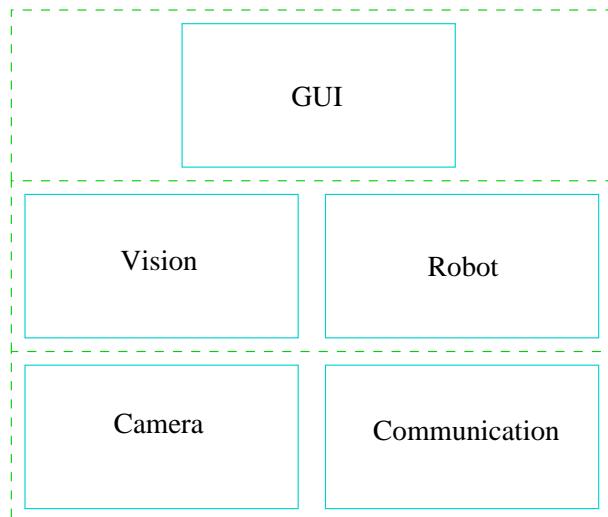


图 6.6 系统软件模块设计

对偶 RGB-D 图像的匹配与合成,旨在提供高质量的 RGB-D 图像;Communication 模块主要实现一个 TCP 服务器,并且定义了与机器人的通信协议,旨在提供与机器人高效稳定的通讯服务;Vision 模块主要实现了 3D-MRAI 算法,旨在根据输入的 RGB-D 图像,输出目标工件的位姿;Robot 模块主要实现与机器人相关的一些路径规划,根据目标工件的位姿生成机械臂抓取的位姿;GUI 模块主要实现对检测结果的可视化,以及一些简单的流程控制。每个模块具体的实现后文会详细介绍。

**开发环境:** 所采用的嵌入式计算机使用的是嵌入式 Linux 操作系统,当然由于嵌入式计算机特性,软件的开发主要在通用计算机上,通用计算机也使用了 Linux 操作系统,这样在编写完成的软件可以无缝拷贝到嵌入式 Linux 操作系统上,但由于通用计算机和嵌入式计算机的 CPU 架构不同,将程序拷贝到嵌入式计算机上后,需要重新编译。当然也可以考虑直接在通用计算机上交叉编译嵌入式计算机上的可执行程序,但考虑到调试方便,并且所使用的嵌入式计算机性能强劲,并未使用交叉编译的方式。考虑到系统的性能以及算法的复杂性,整个系统软件使用 C++ 11 编写,Clang 作为编译器,CMake 作为自动化编译工具,git 作为版本控制,具体如表6.1所示。

---

操作系统	Ubuntu 16.04
编译器	Clang 3.8.0
构建系统	CMake 3.5.1
版本控制	git 2.7.4

---

表 6.1 系统开发环境

软件依赖：系统软件的依赖如下所示：

- librealsense < 2.0.0
- OpenCV >= 3.0.0
- PCL(Point Cloud Library) >= 1.7.0
- Tensorflow >= 1.2.0
- Glog >= 0.3.4
- glfw >= 3.1.2

上述所有的软件都是跨平台、开源的软件，librealsense 是所使用的 RGB-D 相机 SR300 的驱动以及 SDK，系统主要使用它获取相机采集的图片；OpenCV 是一个基于 BSD 许可（开源）发行的跨平台计算机视觉库，系统主要使用 OpenCV 完成对相机采集图像的处理以及对偶 RGB-D 相机图像的合成与匹配算法；PCL 是一个通用的开源点云库，它实现了大量点云相关的通用算法和高效数据结构，涉及到点云获取、滤波、分割、配准、检索、特征提取、识别、追踪、曲面重建、可视化等。支持多种操作系统平台，可在 Windows、Linux、Android、Mac OS X、部分嵌入式实时系统上运行，系统主要使用 PCL 完成一些点云相关的算法；TensorFlow 是谷歌基于 DistBelief 进行研发的深度学习框架，系统主要使用 Tensorflow 完成 3D-MRAI 算法中的深度神经网络；Glog 是谷歌开发的一个 C++ 语言的应用级日志记录框架，提供了 C++ 风格的流操作和各种助手宏，系统主要使用 Glog 完成软件的日志记录；glfw 是一个 OpenGL 图形库，系统主要使用 glfw 完成 GUI 模块的设计。

*Camera module:* Camera 模块的主要接口是一个虚基类 Camera，如图6.7所示。相机模块通过虚基类 Camera 定义了一些通用的接口函数，其他具体的相机通过继承这个基类来实现，对于外部使用者来说并不需要关心这些继承 Camera 类的具体实现，只需要调用 Camera 中定义的接口即可。整个模块具有很强的扩展性，如增加一个新的相机可以通过增加一个继承 Camera 的类，外部调用的模块无需改写。实际上，使用何种相机通过设置配置文件可以由用户选择。

*Communication module:* Communication 模块主要使用 C++ 构建了一个 TCP 服务器，并且规定了通信的协议。讲道理，系统的通讯并不复杂，并且数据量也很小，基本上就视觉系统告诉机器人运动到哪，然后机器人告诉视觉系统是否运动

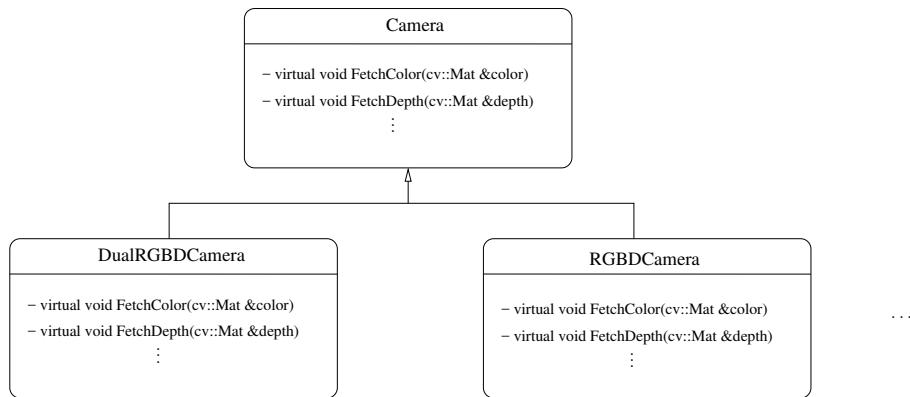


图 6.7 Camera module UML

到了目的地这些简单的信息交流。在仔细研究机器人上具体编程后,由于机器人上编程比较单调且繁琐,因此通讯服务的服务器运行在嵌入式计算机上,并且,定义了接收和发送两类信息,发送信息指的是从嵌入式计算机发送到机器人控制器上的信息,接收信息类似。具体定义了一个类模板,如代码6.1所示。

Listing 6.1 TCP server template

```

template <typename RecvMsgT, typename SendMsgT>
class SyncTCPServer {
public:
    SyncTCPServer(std::string address = "127.0.0.1", unsigned
                  short port = 8000);

    void WaitingClient() {
        acceptor_ ->accept(*socket_);
    }

    /**
     * Receive message from client
     * @param msg the received message
     * @return read bytes size
     */
    int RcvMsg(RecvMsgT &msg);

    /**
     * Send message to client
     * @param msg the message will be sent
     * @return write bytes size
     */
    int SendMsg(const SendMsgT &msg);

private:
    boost::asio::io_service io_service_;
    boost::shared_ptr<boost::asio::ip::tcp::acceptor> acceptor_;
    boost::shared_ptr<boost::asio::ip::tcp::socket> socket_;
};
  
```

*Vision module:* Vision 模块的主要类的 UML 图如图6.8所示。Vision 类提供在

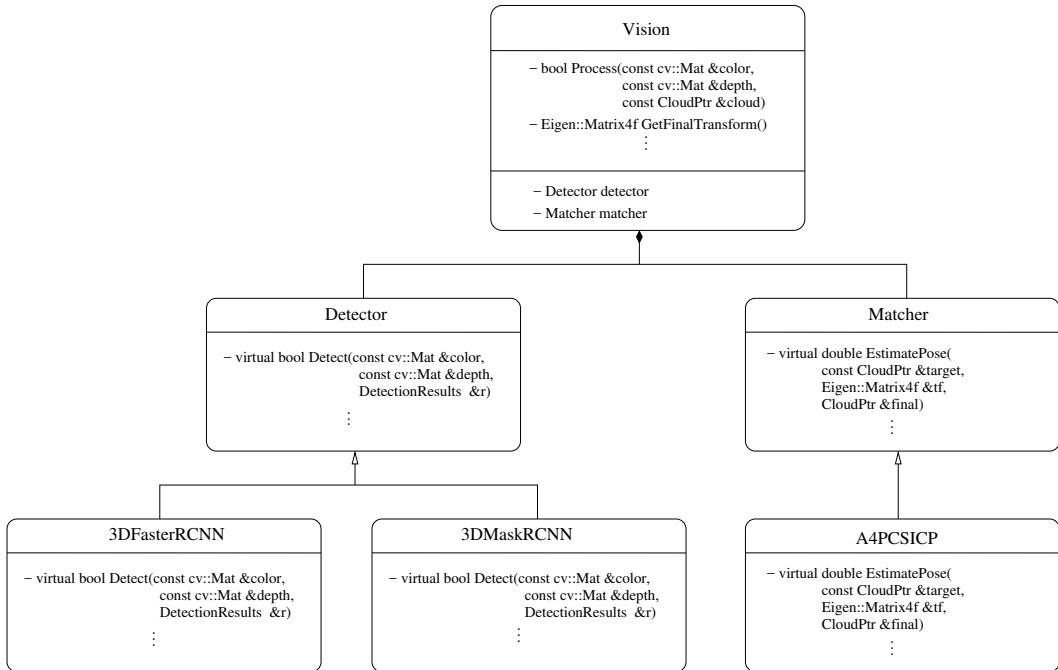


图 6.8 Vision module UML

图像中找出目标工件位姿的接口，其核心是 3D-MRAI 算法，因此也有两个模块：Detector 和 Matcher。Detector 和 Matcher 的设计思想和 Camera 模块类似，通过定义接口屏蔽具体实现，从而提高了程序的扩展性，因为显然可以有多种检测的方法，比如本文就有 3D Faster R-CNN 和 3D Mask R-CNN 两种实现，通过这种方式可以在不改动程序代码的情况下，通过配置文件快速切换所想要使用的算法。

Vision 类的核心就是 3D-MRAI 算法，算法具体内容已经在第5 章中详细叙述了，但此处运用于 Bin-Picking 系统，为了提高系统的效率，针对 Bin-Picking 这个系统，在实现细节上对 3D-MRAI 做了些改动，或者说 trick。其中最主要的 trick 是不再对 3D Faster/Mask R-CNN 中的每个检测结果都去运行 A4PCS-ICP 算法，取而代之的是通过对每个 BBox/Mask 提取的点云根据距离工作台的高度排序，从位置最高的点云开始运行匹配算法，满足条件就返回。这么做的原因是，Bin-Picking 系统每次抓取只能抓取一个工件，并且，理论上位于物料堆最上面的工件显然最好抓取。增加这个 trick 后的 3D-MRAI 算法的流程如算法11所示，这么做大大减少了视觉系统的运算时间，提高了整个系统的抓取工作效率。

*Robot module:* Robot 模块根据 Vision 模块输出的工件位姿，将其变换到机器人坐标系下，然后生成轨迹，发送给机器人。由于 Vision 模块与 Robot 模块之间是解耦的，Vision 模块输出的工件位姿是在相机坐标系下的，为了将相机坐标系

**算法 11: 3D-MRAI with Tricks****Input:** RGB Image  $I$ , Depth Map  $D$ , CAD Models  $M$ **Output:** Set of Pose and Class  $Res$ 


---

```

1  $P \leftarrow \emptyset;$ 
2 forall  $M_i \in M$  do
3    $P \leftarrow \{P, CAD2PointCloud(M_i)\};$ 
4  $H = Depth2HHA(D);$ 
5  $Q = Depth2PointCloud(D);$ 
6  $Mask, Class \leftarrow 3DMAS KRCNN(I, H); // \text{ Same with } 3DFASTERRCNN$ 
7  $Q_{sorted} \leftarrow \emptyset;$ 
8 forall  $m_i \in Mask, c_i \in Class$  do
9    $Q_i \leftarrow Crop(Q, m_i);$ 
10   $Q_{sorted} \leftarrow Q_{sorted}, [Q_i, c_i];$ 
11  $SORTBAS EDONHEIGHT(Q_{sort});$ 
12 forall  $[Q_i, c_i] \in Q_{sort}$  do
13    $P_i \leftarrow P(c_i);$ 
14    $T_i, S_i \leftarrow A4PCSI CP(P_i, Q_i);$ 
15   if  $S_i > S_{min}$  then
16     return  $[T_i, c_i];$ 

```

---

下的位姿变换到机器人坐标系下, 还需要进行相机和机器人之间的标定(手眼标定), 所设计的系统的相机固定在支架上, 与机器人分离, 因此是一个典型的 eye-to-hand calibration。具体的, 记机器人(Robot)基坐标系为  $\{R\}$ , 机器人末端执行器(End Effector)坐标系为  $\{E\}$ , 相机(Camera)坐标系为  $\{C\}$ , 标定板(Board)坐标系为  $\{B\}$ , 则 eye-to-hand 标定要求的就是机器人基坐标系到相机坐标系的齐次变换矩阵  ${}^R_C H$ 。标定的流程如下:

- 将标定板固定在机器人末端执行器上;
- 多次变化机器人末端位姿, 记录机器人末端位姿  ${}^R_E H$  以及标定板在相机坐标系下的位姿  ${}^C_B H$ ;
- 根据多组  ${}^R_E H$  和  ${}^C_B H$  求解  ${}^R_C H$ ;

根据多组机器人末端位姿和标定板位姿求解机器人基坐标系与相机坐标系之间的关系的原理等价与求解矩阵方程  $AX = XB$ 。具体的, 由于标定板固定在机器人末端执行器上, 因此齐次变换矩阵  ${}^B_E H$  恒定, 另外  ${}^R_C H$  也恒定, 对于任意两组

$\{{}^R_E \mathbf{H}_i, {}^C_B \mathbf{H}_i\}, \{{}^R_E \mathbf{H}_j, {}^C_B \mathbf{H}_j\}$  存在等式

$${}^R_E \mathbf{H}_i {}^R_C \mathbf{H} {}^C_B \mathbf{H}_i = {}^R_R \mathbf{H}_j {}^R_C \mathbf{H} {}^C_B \mathbf{H}_j \quad (6.1)$$

将等式6.2两边右乘  ${}_C^B \mathbf{H}_i$ , 再左乘  ${}_E^R \mathbf{H}_j$ , 则等式6.2变为

$$({}^R_E \mathbf{H}_j {}^R_R \mathbf{H}_i) {}_C^R \mathbf{H} = {}_C^R \mathbf{H} ({}^C_B \mathbf{H}_j {}^B_C \mathbf{H}_i) \quad (6.2)$$

即  $\mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{B}$  形式, 矩阵方程  $\mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{B}$  的求解可以参考文献 (Daniilidis 1999)。对于相机在机器人末端的手眼标定也类似, 本文不再详细叙述。

将工件位姿从相机坐标系变换到机器人坐标系后, 便可得机器人抓取的位姿, 机器人抓取的位姿与工件的位姿是事先标定好的, 并且为了提高抓取的成功率, 对一个工件设了多组抓取位姿, 根据工件的位姿选取合适的抓取位姿。

*GUI module:* 为了可视化视觉系统的检测情况, 设计了 GUI 模块实时展示检测结果, 由于系统中存在 3D 的点云, 因此采用 OpenGL 库在三维空间中可视化结果, 并且可以对程序进行简单的控制。三维可视化的界面如图6.9所示。三维可视

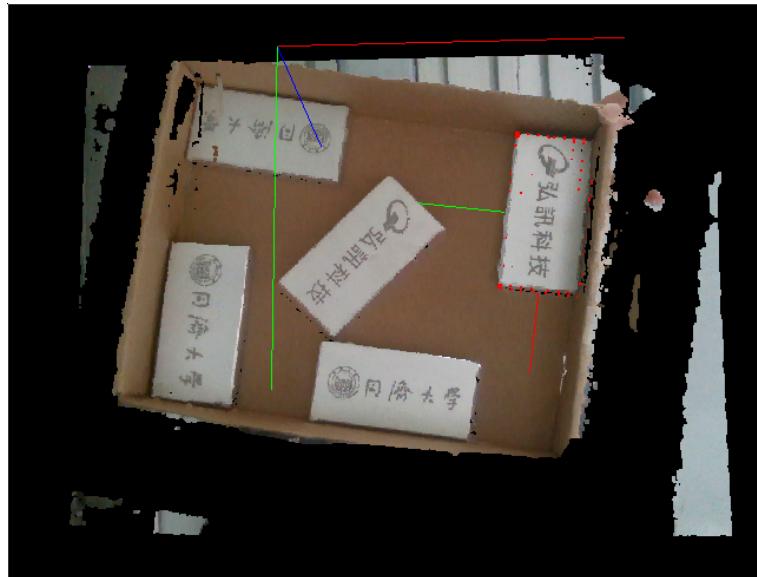


图 6.9 视觉系统三维可视化

化界面中标出了相机坐标系、检测出要抓取的工件的位姿, 整个场景是相机采集到的 3D 数据, 红色部分点云是将目标工件 CAD 模型变换到所计算得到的工件位姿, 因此红色点云与相机采集到的 3D 目标点云相重合说明视觉系统的检测正确。整个界面是 3D 的, 可以通过鼠标放大、缩小、旋转, 也可通过键盘控制视角的前进后退左右移动, 全方位 360 度观察当前检测结果, 如图6.10所示, 在不同视角下观察检测结果。除此之外, 还可以通过键盘保存当前所展示的点云和其他一些数据, 方便进一步分析和观察。

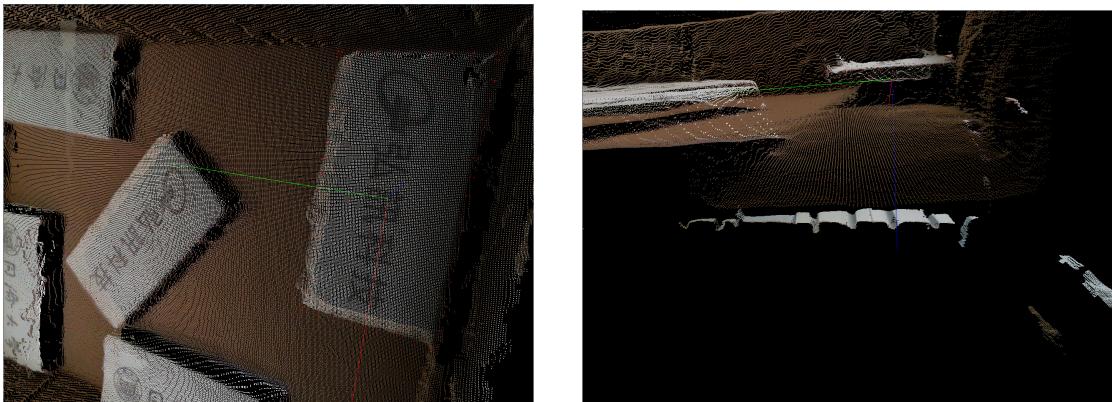


图 6.10 不同视角下观察检测结果

### 6.3 随机分拣实验

#### 6.3.1 实验内容

设计的随机分拣实验在所搭建的工作台上进行，在物料箱中随机放慢物料，然后运行整个随机分拣系统，将物料中的全部工件分拣到另外一个箱子中，分拣完一箱后，再随机填充物料，一共统计十箱物料的抓取结果。

评价系统的指标主要是系统的抓取成功率：

$$R = \frac{m}{n} \times 100\% \quad (6.3)$$

其中  $n$  表示总的抓取次数， $m$  表示成功抓取次数。一次成功的抓取是指机械臂成功将一个物料从物料箱中取出，然后放置到另外一个箱子中。除了系统的成功抓取率，为了考察系统的快速性，定义系统的响应时间  $T_r$  为从机器人请求开始抓取到机器人收到工件位姿，这个时间也是整个软件的响应时间，主要包括了：

- 相机采集时间
- 视觉算法计算时间
- 通讯时间

另外再定义机械臂从开始抓取到将物料抓出箱子的时间为抓取时间  $T_1$ ，机械臂从将物料抓出箱子到放置完物料回到抓取起始点的时间为放置时间  $T_2$ ，如果机械臂只要回到抓取起始点就立即请求下一个抓取位姿，则一个工作周期的总时间为

$$T = T_r + T_1 + T_2 \quad (6.4)$$

但是，考虑到系统的相机是固定在支架上的，因此，只要机械臂将物料抓取出箱子

就可以请求下一个抓取位姿,所以一个工作周期的总时间可以缩减为

$$T = T_1 + \max(T_r, T_2) \quad (6.5)$$

对于评价视觉系统来说,我们更关心系统响应时间  $T_r$ ,对于评价整个 Bin-Picking 系统来说,显然工作周期  $T$  更重要。

### 6.3.2 实验结果

所设计的随机分拣系统的硬件系统抓取完十箱物料后,统计每箱的成功率、平均响应时间、平均抓取时间、平均放置时间和平均工作周期,绘制成表6.2以及图6.11。从表中可以发现所设计的基于 3D-MRAI 的随机分拣系统的平均抓取成

	成功率	响应时间 $T_r$	抓取时间 $T_1$	放置时间 $T_2$	工作周期 $T$
1	100%	711ms	7.6s	4.2s	12.5s
2	100%	729ms	6.8s	4.2s	11.0s
3	100%	708ms	6.5s	4.2s	10.7s
4	100%	701ms	8.1s	4.2s	12.3s
5	100%	713ms	9.3s	4.2s	13.5s
6	100%	722ms	6.6s	4.2s	10.8s
7	100%	693ms	7.9s	4.2s	12.1s
8	100%	732ms	9.1s	4.2s	13.3s
9	100%	718ms	8.9s	4.2s	13.1s
10	100%	723ms	6.9s	4.2s	11.1s
Avg.	100%	715ms	7.77s	4.20s	12.04s

表 6.2 随机分拣实验结果

功率为 100%,但理论上随着抓取次数的增加,个人认为系统一定会出现抓取失败的情况,并且实验中所使用的物料也相对单一,不同的物料理论上也会影响视觉的识别效果,因此换一种物料系统的抓取成功率也可能达不到 100%。但总体上来说,此次实验 100% 的成功率充分说明了所设计的 Bin-Picking 视觉系统完全能满足一般随机分拣任务,成功率高。

另外,视觉系统的响应时间为 715ms 左右,分拣系统的工作周期为 12.04s 左右,从表中数据可以发现工作周期基本上为抓取时间和放置时间之和,也就是完全是机械臂运动的时间,除了一开始系统启动时会增加额外的响应时间,但这对真个工作时间来说可以忽略不计,因此,所设计的基于 3D-MRAI 算法的 Bin-Picking 视觉系统在时间上完全满足一般分拣任务的要求,实时性高。

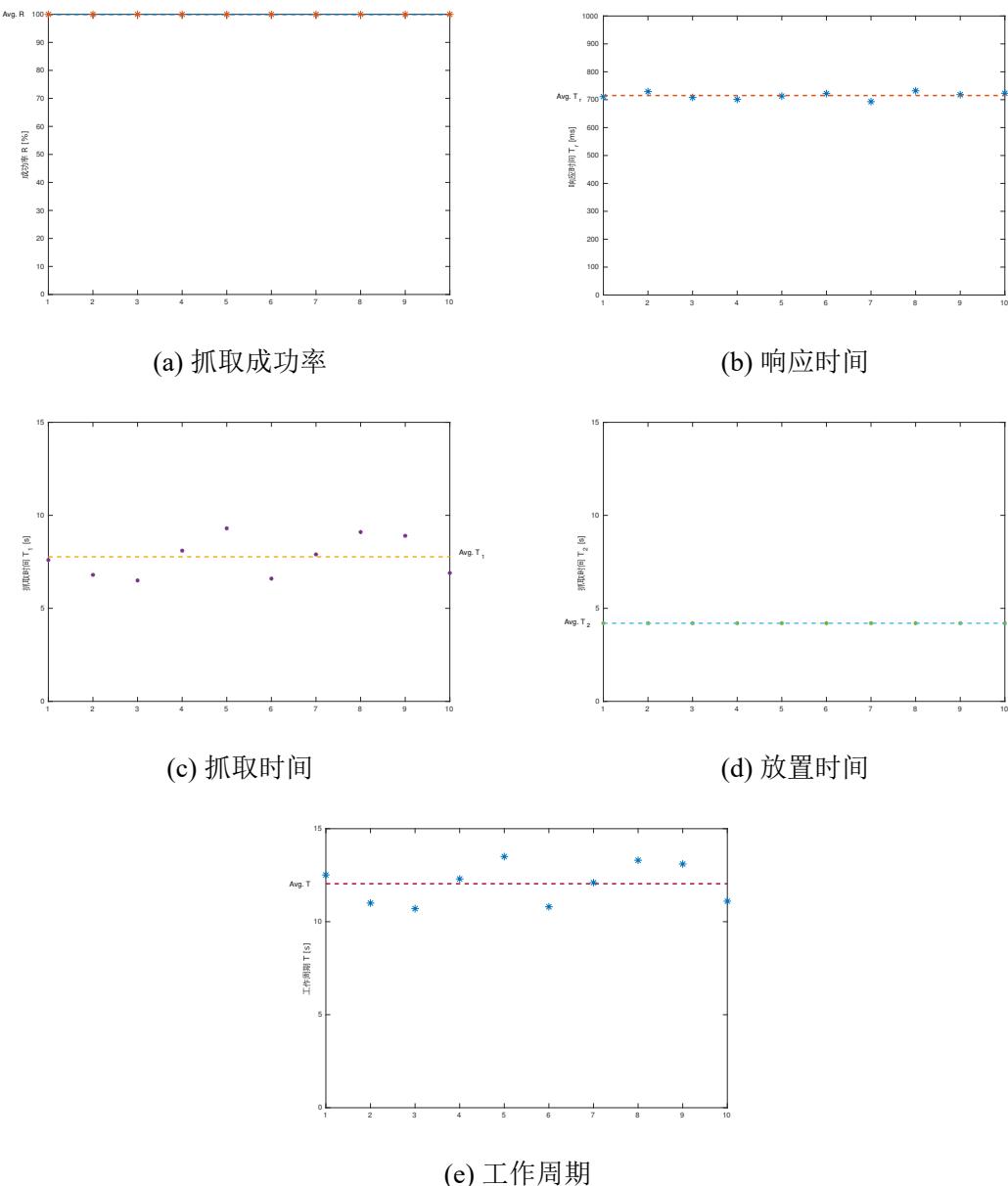


图 6.11 随机分拣实验结果

## 6.4 本章小结

本章将 3D-MRAI 算法应用到实际机器人上,用以解决 Bin-Picking 随机分拣问题,设计并实现了一个 Bin-Picking 视觉系统,并进行了抓取实验,实验结果表明,所设计的视觉系统有很高的抓取成功率,并且其响应速度也完全能满足分拣任务的要求。

## 第 7 章 结论与展望

### 7.1 结论

对于机器人三维感知面临巨大困难,本文基于 RGB-D 图像,通过引入深度学习,和传统算法相结合,旨在 3D 目标检测和位姿估计问题上有所突破,以推动整个 3D 视觉的发展,解决机器人的感知问题。本文主要有三个方面的贡献:第一,本文从 3D 视觉的深度信息的获取入手,对于当前 3D 相机性价比低的现状,提出了对偶 RGB-D 相机结构;第二,将深度学习和传统视觉算相结合,提出了 3D-MRAI 算法实现对目标物体的检测和位姿估计;第三,将 3D-MRAI 算法应用到实际机器人领域,解决 Bin-Picking 问题。

对偶 RGB-D 相机结构针对单个 RGB-D 相机采集的深度图缺失问题,通过增加一个旋转了 180 度的 RGB-D 相机与原相机构成对偶,然后结合两个相机的深度图以及通过两个相机彩色图构建的深度图来获取高质量的深度图,实验表明对偶 RGB-D 相机再结合了三张深度图后,生成的深度图比单个相机采集的深度图具有更高的填充率,深度信息的精度也更高。

3D-MRAI 算法针对 3D 目标检测和位姿估计问题的特点,分别对深度学习算法 Faster/Mask R-CNN 和传统视觉算法 4PCS 进行改进并结合,从而实现 3D 目标检测和位姿估计。实验表明,3D-MRAI 在所采集的 workpiece 数据集上相比传统算法表现出了较高的检测准确率,估计的位姿精度也更高。

针对 Bin-Picking 问题,本文将对偶 RGB-D 结构与 3D-MRAI 算法实际应用,开发出的 Bin-Picking 视觉系统相比与传统 Bin-Picking 视觉系统,具有更高的抓取成功率、更快的响应速度,以及更低的成本。在所设计的抓取实验中达到了 100% 的抓取成功率、约 0.7s 的响应时间。

### 7.2 进一步工作的方向

对于深度信息的获取,在不考虑成本的情况下,当然高价的 3D 相机获取的深度图质量高,本文使用两个低价的 RGB-D 相机构成对偶 RGB-D 相机也是无奈之举,低价 RGB-D 相机获取的深度信息在算法上再如何改进也无法对深度图质

量产生质的改变,因此针对这一部分,还是寄希望于广大相机厂家推出高性价比的 3D 相机吧。

对于本文提出的 3D-MRAI 算法,进一步工作方向可以有以下三点:第一,3D-MRAI 虽然能较好的检测目标并估计目标位姿,但其算法 FPS 只有 2.2,对于实时性要求较高的情况如避障难以满足要求,因此进一步工作可以针对算法的运算时间进行改进,比如,通过对 3D-MRAI 中的深度网络进行裁剪,因为算法的运算时间主要耗费在深度神经网络的运算上。第二,3D-MRAI 算法检测目标估计位姿分为两步:先检测出目标的 BBox/Mask,然后再匹配 3D 模型和目标点云得到目标位姿。直觉上感觉不是那么完美,能否设计一个 end-to-end 的深度神经网络,一步计算出目标的位姿和种类。第三,3D-MRAI 算法依赖于目标的 3D 模型,需要知道目标的 3D 模型才能得到目标位姿,大大限制了算法的实际应用,因此,进一步工作可以尝试解决 3D 模型未知情况下目标位姿的估计。

## 致谢

行文至此，学生感慨万分。回想起最初开始课题时的不知所措，到现在初有成果，这其中学生所学到的，不仅是专业知识与实践技能，更对我热爱的专业，对于科研有了全新的认识与领悟。转瞬间，在同济求学的三年时间已近尾声，成为一名“具有科学素养的工程师”是我在入学之初记住的，今后学生也并将铭记于心，我想这不仅是同济人的格调，更是我们看待世界的方法。

首先，我要衷心的感谢导师陈启军教授对我的教诲与启迪。我仍记得第一次见到陈老师时，关于科研的问题，他对我的教导：任何成果都不在于冥想，不在于他人，而在于自己实践过程的点滴。这句话为我的科研与生活指明了方向。感谢陈老师对我知识上的教诲，科研过程中的点拨，在我迷惘无助时给予我的关心与指引。每次与陈老师的交流都让学生受益匪浅。您的治学态度，处事原则，是学生今后学习的楷模。

诚挚的感谢朱劲老师，张皓老师，王祝萍老师，马小峰老师，刘成菊老师，对我的帮助与指导，是您们的传道授业解惑让学生更好的完成研究生的学业与科研工作。

感谢张奎师兄对我的帮助与支持，感谢您给予我的每一次沟通与交流，不仅在科研上，更在生活上给予我的关心与支持。感谢课题组的尹小川博士，王香伟博士在课题上给予我的帮助，不厌其烦地解疑答惑，给我的论文提出了宝贵的意见。更要感谢与我的同门与挚友王继民、孙旭和宁静，你们的帮助与陪伴是我收获的宝贵财富。感谢实验室的熊峰博士，杜明晓博士以及给予我帮助的师弟师妹，让我在同济的生涯永远难忘。

最后，感谢我敬爱的父亲、母亲，你们的爱与陪伴是我永远的港湾和前行的动力。

感谢所有给予我帮助的人，感谢同济给予我的美好时光。

2018年3月

## 参考文献

- [1] AIGER D, MITRA N J, COHEN-OR D, 2008. 4-points congruent sets for robust pairwise surface registration[M]//ACM Transactions on Graphics (TOG): volume 27. [S.l.]: ACM: 85.
- [2] ALEXANDRE L A, 2016. 3d object recognition using convolutional neural networks with transfer learning between input channels[M]//Intelligent Autonomous Systems 13. [S.l.]: Springer: 889–898.
- [3] ARYA S, MOUNT D M, NETANYAHU N S, et al. An optimal algorithm for approximate nearest neighbor searching fixed dimensions[J]. Journal of the ACM (JACM), 1998, 45(6): 891–923.
- [4] BESL P J, MCKAY N D, 1992. Method for registration of 3-d shapes[M]//Sensor Fusion IV: Control Paradigms and Data Structures: volume 1611. [S.l.]: International Society for Optics and Photonics: 586–607.
- [5] BOLLES R C, FISCHLER M A, 1981. A ransac-based approach to model fitting and its application to finding cylinders in range data.[M]//IJCAI: volume 1981. [S.l.: s.n.]: 637–643.
- [6] BROWN D C. Decentering distortion of lenses[J]. Photogrammetric Engineering and Remote Sensing, 1966.
- [7] CHUM O, MATAS J, OBDRZALEK S, 2004. Enhancing ransac by generalized model optimization[M]//Proc. of the ACCV: volume 2. [S.l.: s.n.]: 812–817.
- [8] COLLET A, MARTINEZ M, SRINIVASA S S. The moped framework: Object recognition and pose estimation for manipulation[J]. The International Journal of Robotics Research, 2011, 30(10): 1284–1306.
- [9] CONNOLLY C. A new integrated robot vision system from fanuc robotics[J]. Industrial Robot: An International Journal, 2007, 34(2): 103–106.
- [10] CORSINI M, DELLEPIANE M, GANOVELLI F, et al. Fully automatic registration of image sets on approximate geometry[J]. International journal of computer vision, 2013, 102(1-3): 91–111.
- [11] DANIILIDIS K. Hand-eye calibration using dual quaternions[J]. The International Journal of Robotics Research, 1999, 18(3): 286–298.
- [12] DIAS A S, BRITES C, ASCENSO J, et al. Sift-based homographies for efficient multiview distributed visual sensing[J]. IEEE Sensors Journal, 2015, 15(5): 2643–2656.
- [13] GEIGER A, ROSER M, URTASUN R. Efficient Large-Scale Stereo Matching[J]. Accv, 2010.
- [14] GOODRICH M T, MITCHELL J S, ORLETSKY M W, 1994. Practical methods for approximate geometric pattern matching under rigid motions:(preliminary version)[M]//Proceedings of the tenth annual symposium on Computational geometry. [S.l.]: ACM: 103–112.
- [15] GOOGLE, 2012. Tango[EB/OL]. <https://developers.google.com/tango>.
- [16] GUPTA S, ARBELAEZ P, MALIK J, 2013. Perceptual organization and recognition of indoor scenes from rgb-d images[M]//Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. [S.l.]: IEEE: 564–571.

- [17] GUPTA S, GIRSHICK R, ARBELÁEZ P, et al., 2014. Learning rich features from rgb-d images for object detection and segmentation[M]//European Conference on Computer Vision. [S.l.]: Springer: 345–360.
- [18] HE K, ZHANG X, REN S, et al., 2016. Deep residual learning for image recognition[M]// Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.]: 770–778.
- [19] HE K, GKIOXARI G, DOLLÁR P, et al., 2017. Mask r-cnn[M]//Computer Vision (ICCV), 2017 IEEE International Conference on. [S.l.]: IEEE: 2980–2988.
- [20] HEIKKILÄ J. Geometric camera calibration using circular control points[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(10): 1066–1077. DOI: 10.1109/34.879788.
- [21] HINTERSTOISSE S, CAGNIART C, ILIC S, et al. Gradient response maps for real-time detection of textureless objects[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(5): 876–888.
- [22] HINTERSTOISSE S, LEPETIT V, ILIC S, et al., 2012. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes[M]//Asian conference on computer vision. [S.l.]: Springer: 548–562.
- [23] HORN B K. Closed-form solution of absolute orientation using unit quaternions[J]. JOSA A, 1987, 4(4): 629–642.
- [24] IMAGENET, 2011. Imagenet[EB/OL]. <http://www.image-net.org>.
- [25] JOHNSON A E, HEBERT M. Using spin images for efficient object recognition in cluttered 3d scenes[J]. IEEE Transactions on pattern analysis and machine intelligence, 1999, 21(5): 433–449.
- [26] JONSCHKOWSKI R, EPPNER C, HÖFER S, et al., 2016. Probabilistic multi-class segmentation for the amazon picking challenge[M]//Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on. [S.l.]: IEEE: 1–7.
- [27] KRIZHEVSKY A, SUTSKEVER I, HINTON G E, 2012. Imagenet classification with deep convolutional neural networks[M]//Advances in neural information processing systems. [S.l.: s.n.]: 1097–1105.
- [28] LI X, GUSKOV I, 2005. Multiscale features for approximate alignment of point-based surfaces.[M]//Symposium on geometry processing: volume 255. [S.l.: s.n.]: 217.
- [29] LIN T Y, DOLLÁR P, GIRSHICK R, et al., 2017. Feature pyramid networks for object detection[M]//CVPR: volume 1. [S.l.: s.n.]: 4.
- [30] LOOP C, ZHANG Z, 1999. Computing rectifying homographies for stereo vision[C]// Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.: volume 1. [S.l.]: IEEE: 125–131.
- [31] LOWE D G, 1999. Object recognition from local scale-invariant features[M]//Computer vision, 1999. The proceedings of the seventh IEEE international conference on: volume 2. [S.l.]: Ieee: 1150–1157.
- [32] MATAS J, CHUM O, URBAN M, et al. Robust wide-baseline stereo from maximally stable extremal regions[J]. Image and vision computing, 2004, 22(10): 761–767.
- [33] MATOBA O, TAJAHUERCE E, JAVIDI B. Real-time three-dimensional object recognition with multiple perspectives imaging[J]. Applied optics, 2001, 40(20): 3318–3325.

- [34] MICROSOFT, 2012. Kinect[EB/OL]. <https://www.xbox.com/en-US/xbox-one/accessories/kinect>.
- [35] MIT-PRINCETON, 2016. "shelf & tote" benchmark dataset for 6d object pose estimation [EB/OL]. <http://apc.cs.princeton.edu/#shelf-and-tote-benchmark-dataset>.
- [36] REN S, HE K, GIRSHICK R, et al., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks[M]//Advances in neural information processing systems. [S.l.: s.n.]: 91–99.
- [37] RUSU R B, BLODOW N, BEETZ M, 2009. Fast point feature histograms (fpfh) for 3d registration[M]//Robotics and Automation, 2009. ICRA'09. IEEE International Conference on. [S.l.]: IEEE: 3212–3217.
- [38] SALTÌ S, TOMBARI F, DI STEFANO L. Shot: Unique signatures of histograms for surface and texture description[J]. Computer Vision and Image Understanding, 2014, 125: 251–264.
- [39] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
- [40] SUR F, NOURY N, BERGER M O, 2008. Computing the Uncertainty of the 8 point Algorithm for Fundamental Matrix Estimation[C/OL]//Proceedings of the British Machine Vision Conference 2008. 96.1–96.10. <http://www.bmva.org/bmvc/2008/papers/269.html>. DOI: 10.5244/C.22.96.
- [41] SWAIN M J, BALLARD D H. Color indexing[J]. International journal of computer vision, 1991, 7(1): 11–32.
- [42] WOLFSON H J, RIGOUTSOS I. Geometric hashing: An overview[J]. IEEE computational science and engineering, 1997, 4(4): 10–21.
- [43] XIE S, GIRSHICK R, DOLLÁR P, et al., 2017. Aggregated residual transformations for deep neural networks[M]//Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. [S.l.]: IEEE: 5987–5995.
- [44] ZENG A, SONG S, NIESSNER M, et al. 3dmatch: Learning the matching of local 3d geometry in range scans[J]. arXiv, 2016, 1603.
- [45] ZHANG Z. A Flexible New Technique for Camera Calibration (Technical Report)[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 22(11): 1330–1334. DOI: 10.1109/34.888718.

## 附录 A 补充资料

## 个人简历、在学期间发表的学术论文与研究成果

### 个人简历

李勇奇,男,1992年12月生。

2015年6月毕业于南京理工大学自动化专业获工学学士学位。

2015年9月入同济大学攻读控制科学与工程专业的硕士学位。

### 已发表论文

- [1] XU, Jing, Chenxiao CAI, Yongqi LI, and Y. Zou. "Dual-loop path tracking and control for quad-rotor miniature unmanned aerial vehicles." *Control Theory & Applications* 32, no. 10 (2015): 1335-1342.

### 科研竞赛获奖

- [1] 2016 RoboCup 机器人世界杯标准平台组八强, 莱比锡, 德国
- [2] 2016 RoboCup 机器人世界杯中国赛标准平台组冠军, 安徽, 中国