

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

在实际工程结构的服役过程中,由于非线性与随机性的耦合作用,工程结构特别是混凝土结构的非线性反应具有不可精确预测的性质。因此,从概率密度演化的角度考察工程结构的非线性性状是准确把握结构非线性性能的必由之路。本文基于随机结构反应概率密度演化的思想对于随机结构分析理论进行了深入的探讨,初步建立了随机结构反应概率密度演化的基本图景。

结构静力非线性分析是评价结构抗震性能的重要手段。对于具有双线型广义随机本构关系材料的结构,其塑性截面分布状态的演化过程即非线性损伤构形状态转移过程反映了结构内力演化的性质。无记忆特性结构的非线性损伤构形状态转移过程具有马尔可夫性,通过结构的力学分析可建立风险率函数与状态转移速率之间的关系,进一步考虑状态之间的逻辑关系,即可得到概率转移速率矩阵。对于有记忆特性结构及力-状态联合演化过程,可通过引入相应的记忆变量构造向量马尔可夫过程,并采用次序分析方法建立其确定性的概率密度演化方程。关于简单结构的情况进行了解析求解,并据以探讨了结构非线性构形状态演化的若干特征,发现了在实际应用中可能具有重要意义的稳定构形现象。讨论了力-状态的解耦问题。基于非线性构形状态本身的性质以及演化过程的规律,初步研究了可能的简化与近似方法。

.....

最后,关于进一步工作的方向进行简要的讨论。

关键词: 随机结构,马尔可夫过程,非线性构形状态,差分方法

ABSTRACT

In practical engineering, the structures usually exhibits strong nonlinearity coupled with randomness of the involved parameters. This makes it almost impossible to exactly predict nonlinear response of the structures, particularly for the concrete structures. To tackle the difficulty, it is necessary to capture the nonlinear performance of the structures in the sense of probability, instead of purely deterministic standpoint. The present thesis is the result of the efforts devoted to developing the probability density evolution method for analysis of nonlinear stochastic structures.

.....

In the finality, the problems requiring further studies are discussed.

Key Words: stochastic structure, Markov process, nonlinear configuration state, difference method

目录

第 1 章 引言	1
第 2 章 RGB-D 图像的获取与融合	2
2.1 3D 相机现状与分析	2
2.2 RGB-D 相机	3
2.2.1 RGB-D 相机原理与结构	3
2.2.2 RGB-D 相机的数学模型	4
2.2.3 RGB-D 相机的标定流程	6
2.3 对偶 RGB-D 相机	9
2.3.1 对偶 RGB-D 相机原理与结构	9
2.3.2 对偶 RGB-D 相机的标定流程	14
2.4 深度图质量测试实验	17
2.4.1 实验流程	17
2.4.2 实验原理	18
2.4.3 实验结果	20
2.5 本章小结	20
第 3 章 基于 RGB-D 图像的目标检测算法	21
3.1 3D Faster R-CNN	21
3.1.1 Faster R-CNN	23
3.1.2 HHA	23
3.1.3 Spatial Transformer	25
3.2 3D Mask R-CNN	28
3.2.1 特征提取网络	29
3.2.2 ROIAlign	30
3.2.3 Mask 损失函数	30
3.3 目标检测实验	31
3.3.1 数据集	32
3.3.2 实验内容	35
3.3.3 实验结果	36
3.4 本章小结	37
第 4 章 基于点云的位姿估计算法	38
4.1 ICP	38
4.2 Super4PCS	38

第 5 章 实验验证	39
第 6 章 结论与展望	40
6.1 结论	40
6.2 进一步工作的方向	40
致谢	41
参考文献	42
附录 A 补充资料	43
个人简历、在学期间发表的学术论文与研究成果	44

符号说明

GNU	GNU's Not Unix /'gnu:/
GFDL	GNU Free Documentation License
GPL	GNU General Public License
FSF	Free Software Foundation

第1章 引言

本文 (Knuth 1989)

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

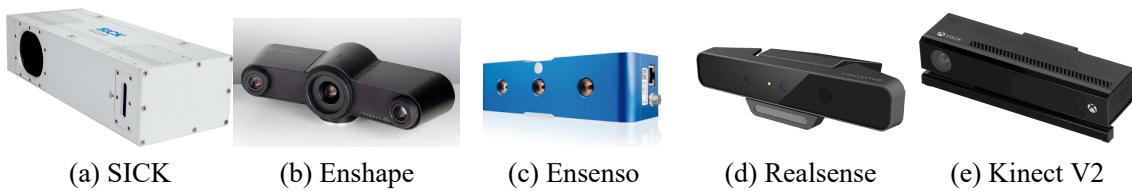


图 2.1 市面上主要 3D 相机

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^①,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 1920×1080 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

^① 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

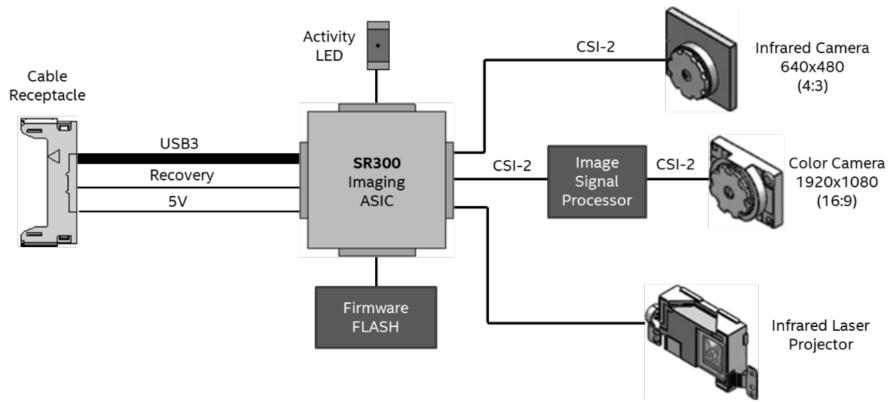


图 2.2 Realsense SR300 内部结构图

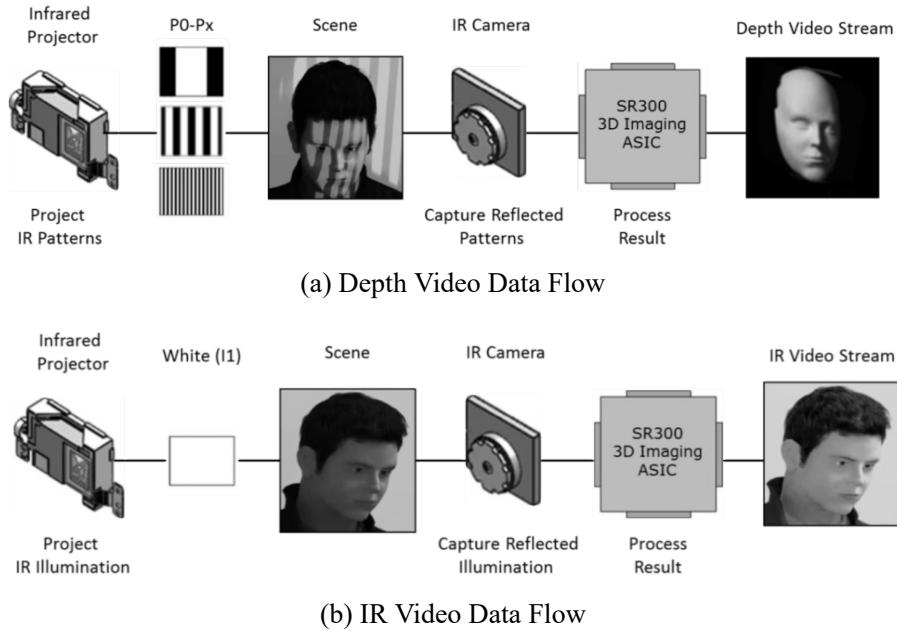


图 2.3 Realsense SR300 深度成像流程

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:

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

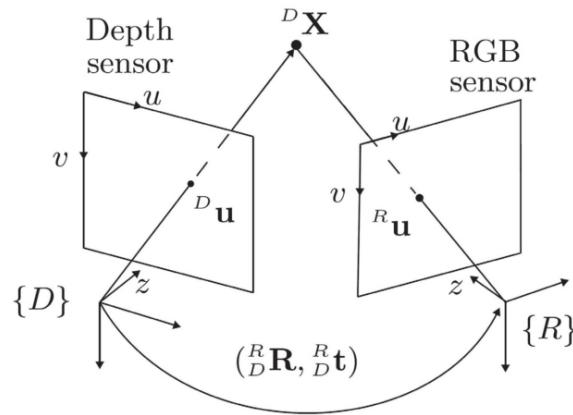


图 2.4 RGB-D 相机模型

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:

$$\begin{cases} \hat{x} = \bar{x}(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) \\ \hat{y} = \bar{y}(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) \end{cases} \quad (2.3)$$

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 4×11 的不对称圆盘标定板。使用圆盘标定板而非棋盘格标定

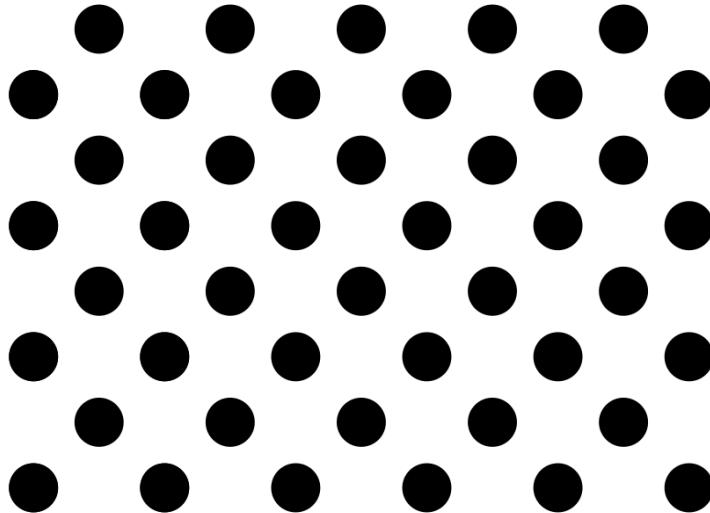


图 2.5 Asymmetrical Circle Board

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。

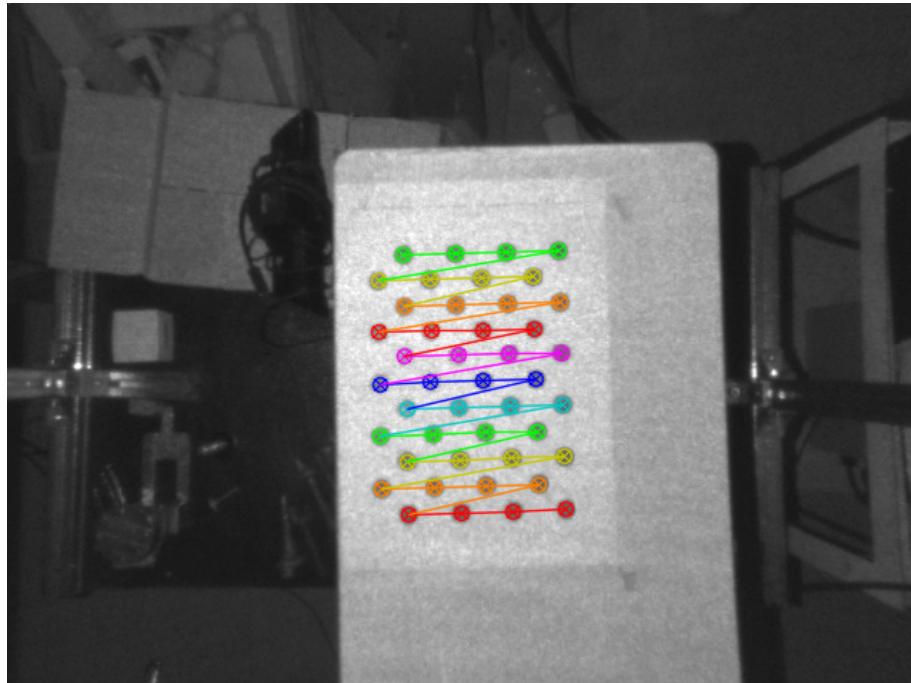


图 2.6 红外成像图中检测标定板

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 4×4 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 4×4 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

$${}^R_D \mathbf{H} := \begin{bmatrix} {}^R \mathbf{R} & {}^R \mathbf{t} \\ {}^D \mathbf{R} & {}^D \mathbf{t} \\ \mathbf{0}_{1 \times 3} & 1 \end{bmatrix} \quad (2.15)$$

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

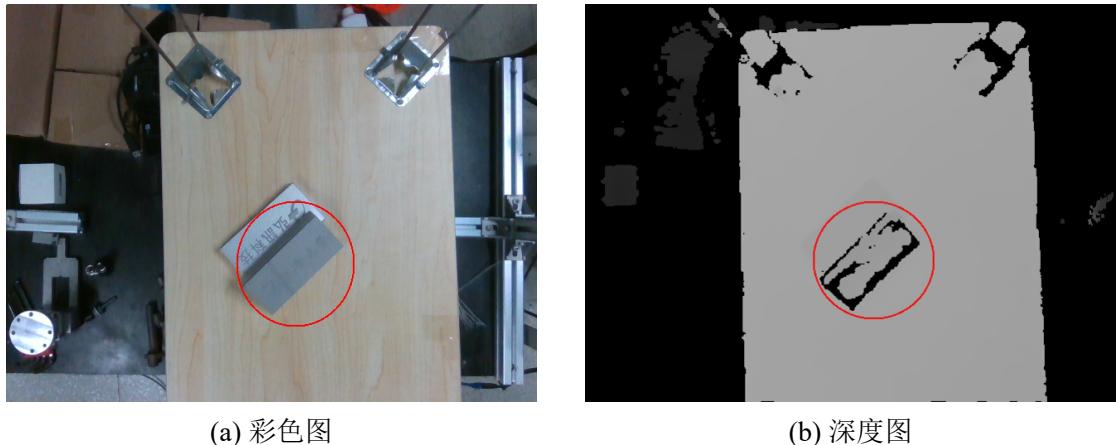


图 2.7 SR300 采集的物体深度信息部分缺失情况下的深度图

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。

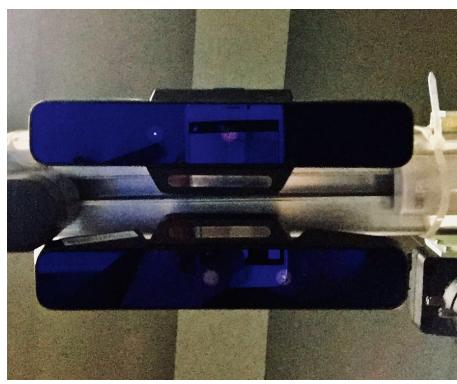


图 2.8 对偶 RGB-D 相机实际物理结构

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^②,有

^② 实际上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。

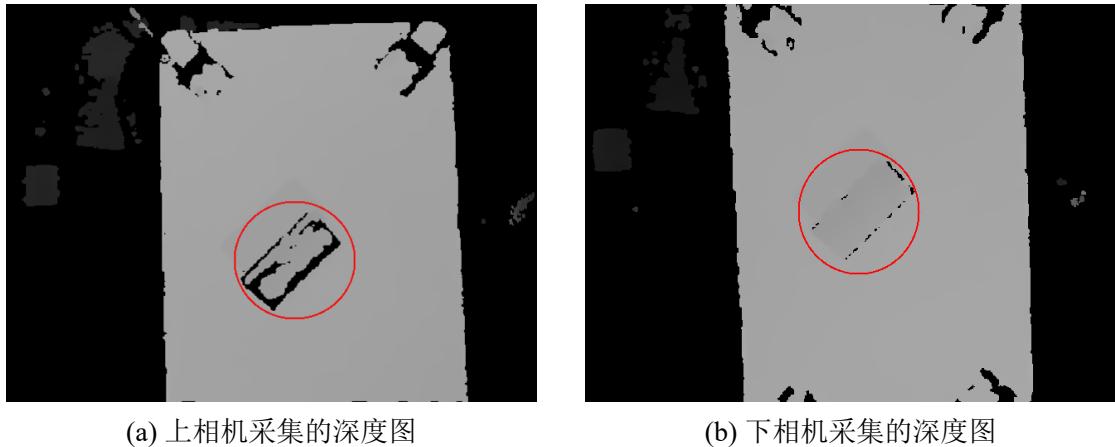


图 2.9 对偶 RGB-D 相机采集的左右两张深度图

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。

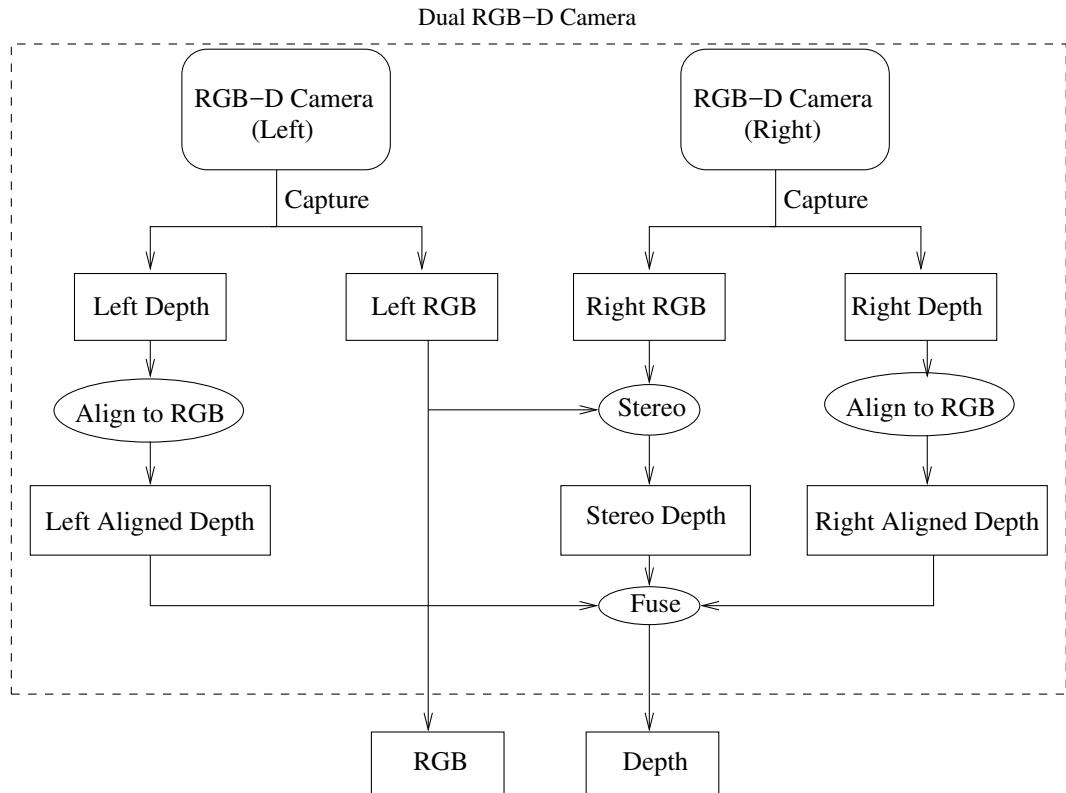


图 2.10 对偶 RGB-D 相机内部原理图

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 $Aligned_D_{ch \times cw}$

for p in $Aligned_D$ **do**

$p = 0$

for $dy = 1; dy <= dh; ++dy$ **do**

for $dx = 1; dx <= dw; ++dx$ **do**

 通过深度相机内参将点 (dx, dy) 反投影到三维空间一点 ${}^D\mathbf{X}$;

 坐标变换 ${}^R\mathbf{X} = {}_D^R\mathbf{R} {}^D\mathbf{X} + {}_D^R\mathbf{t}$;

 通过彩色相机内参将点 ${}^R\mathbf{X}$ 投影变换到彩色图像坐标系下一点

(cx, cy) ;

if cx in $(0, cw]$ and cy in $(0, ch]$ **then**

$Aligned_D(cx, cy) = Raw_D(dx, dy);$

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

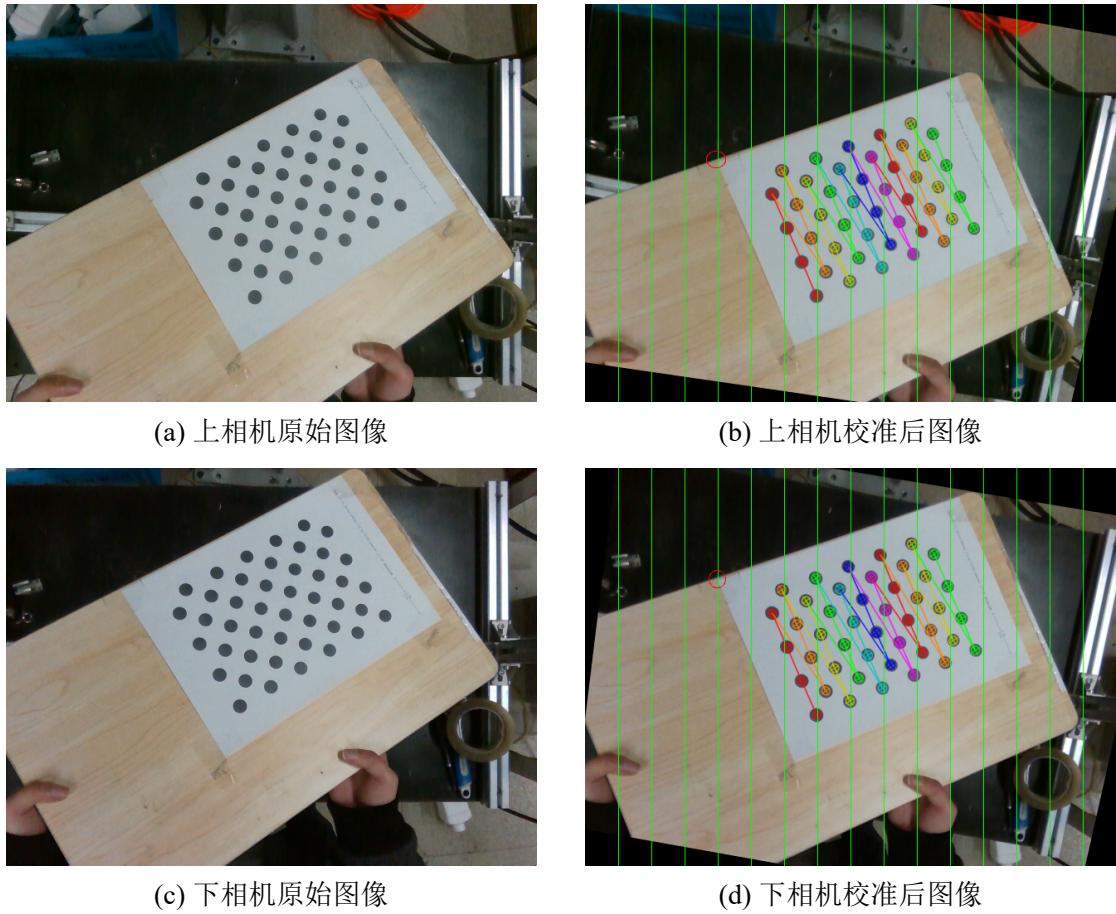


图 2.11 双目相机原始图像和校准后图像

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 $FD_{h \times w}$

```

for  $y = 1; y <= h; ++y$  do
    for  $x = 1; x <= w; ++x$  do
        if  $valid(D_{x,y})$  then
             $FD_{x,y} = D_{x,y};$ 
        else
             $FD_{x,y} = NAN;$ 
            bool leftTop =  $valid(D_{x-1,y-1})$  or  $valid(D_{x,y-1})$  or  $valid(D_{x-1,y})$ ;
            bool leftBottom =  $valid(D_{x-1,y+1})$  or  $valid(D_{x,y+1})$  or  $valid(D_{x-1,y})$ ;
            bool rightTop =  $valid(D_{x+1,y-1})$  or  $valid(D_{x,y-1})$  or  $valid(D_{x+1,y})$ ;
            bool rightBottom =  $valid(D_{x+1,y+1})$  or  $valid(D_{x,y+1})$  or  $valid(D_{x+1,y})$ ;
            if  $leftTop$  and  $leftBottom$  and  $rightTop$  and  $rightBottom$  then
                validPoints = {};
                for  $dy = -1; dy <= 1; ++dy$  do
                    for  $dx = -1; dx <= 1; ++dx$  do
                        if  $valid(D_{dx,dy})$  then
                            push back  $D_{x,y}$  to validPoints;
                if  $max(validPoints) - min(validPoints) < 0.05$  then
                     $FD_{x,y} = mean(validPoints);$ 
    
```

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

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 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 :

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

算法 3: Fuse Depth Frames

Input: leftDepth, rightDepth, stereoDepth

Output: fuseDepth

Initialize w1,w2,w3;

for (d_1, d_2, d_3, d_4) in ($leftDepth, rightDepth, stereoDepth, fuseDepth$) **do**

validDepth = [], validWeight = [];

for $i = 1$ to 3 **do**

if d_i is valid **then**

push back d_i to validDepth, w_i to validWeight;

if size of validDepth == 0 **then**

$d_4 = \text{NAN}$;

else if size of validDepth == 1 **then**

$d_4 = \text{validDepth}[1]$;

else if size of validDepth == 2 **then**

if extremum of validDepth < 0.03 **then**

$d_4 = \text{validDepth} \cdot \text{validWeight} / \text{sum of validWeight};$

else

$d_4 = \text{NAN}$;

else

mediumDepth = medium(validDepth);

$d_4 = 0$, sum = 0;

for (d, w) in (validDepth, validWeight) **do**

if $\text{abs}(d - \text{mediumDepth}) < 0.03$ **then**

$d_4 += d * w$;

sum += w;

if sum > 0 **then**

$d_4 = d_4 / \text{sum}$;

else

$d_4 = \text{NAN}$;

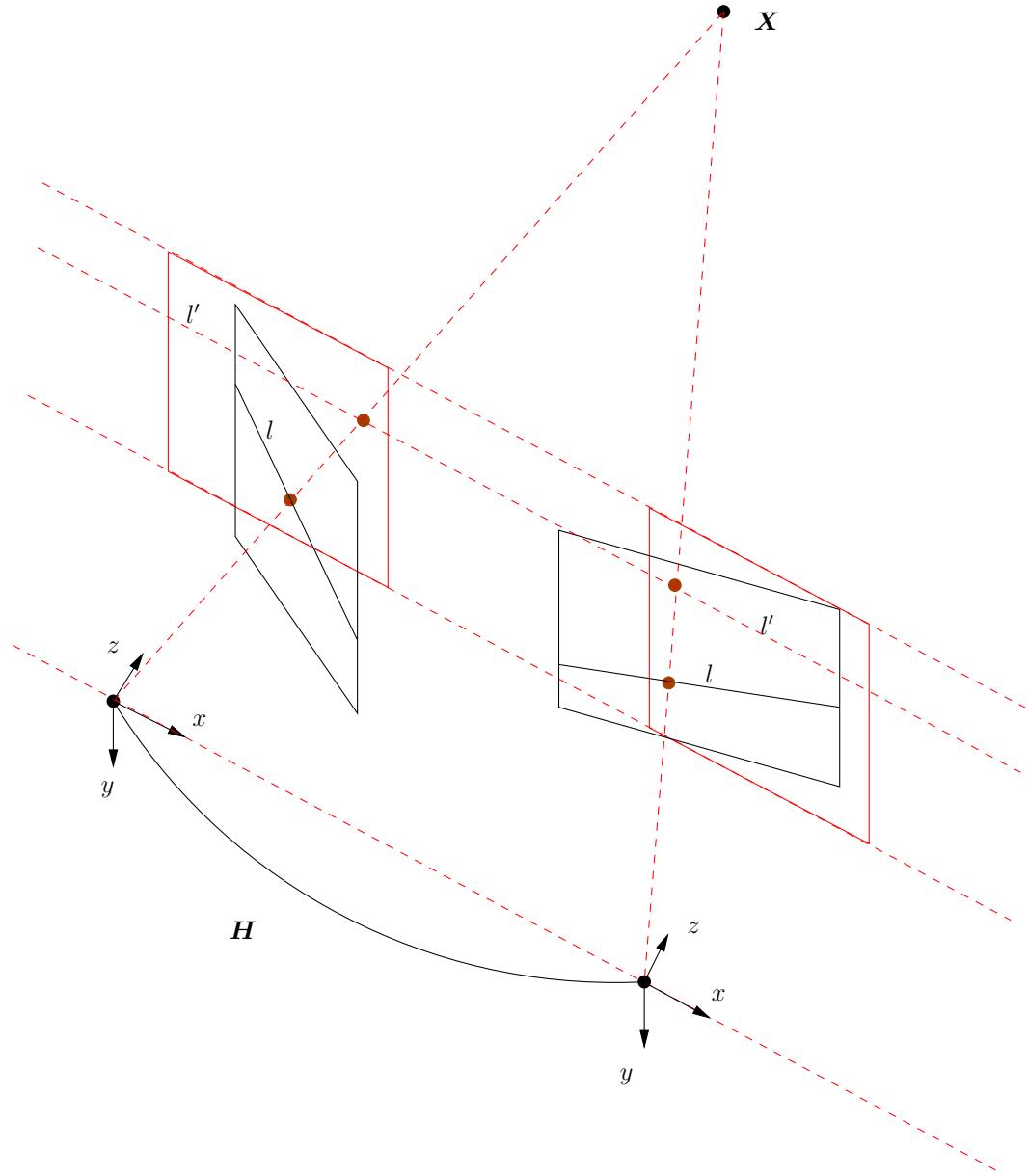


图 2.14 双目几何结构

其中 $R_z(\theta)$ 表示绕 Z 轴旋转 θ 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. 2001), 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

TODO 深度测试实验装置图

图 2.15 深度测试实验装置

- 主相机^③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

^③ 另外一个相机的投影变换矩阵也可以得到,但没有必要。

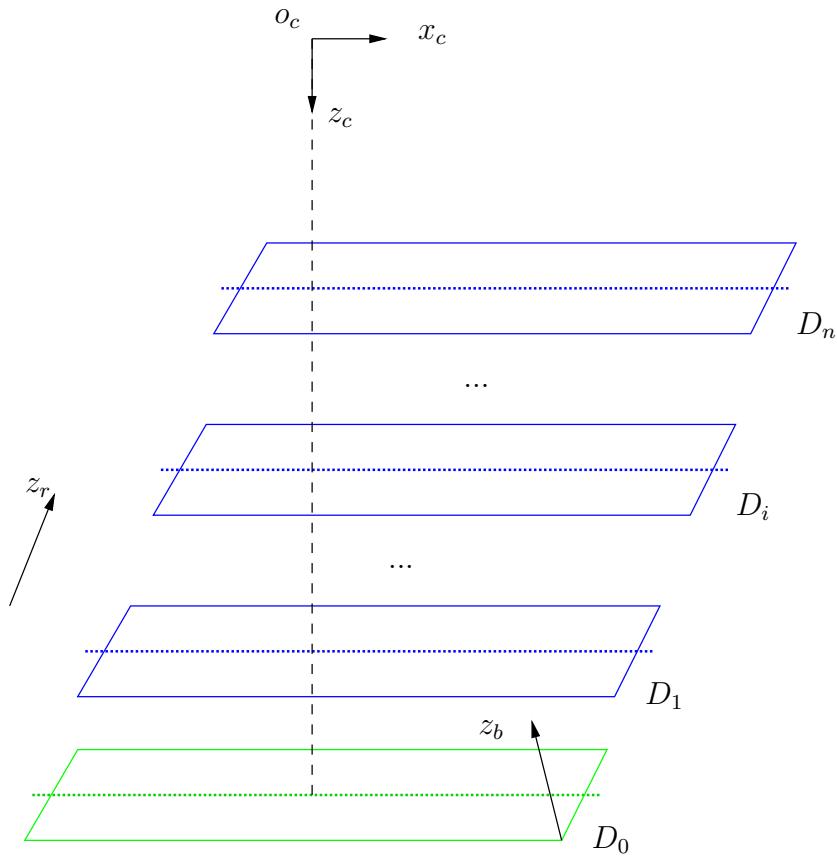


图 2.16 实验流程图

- 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

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 δ 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

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

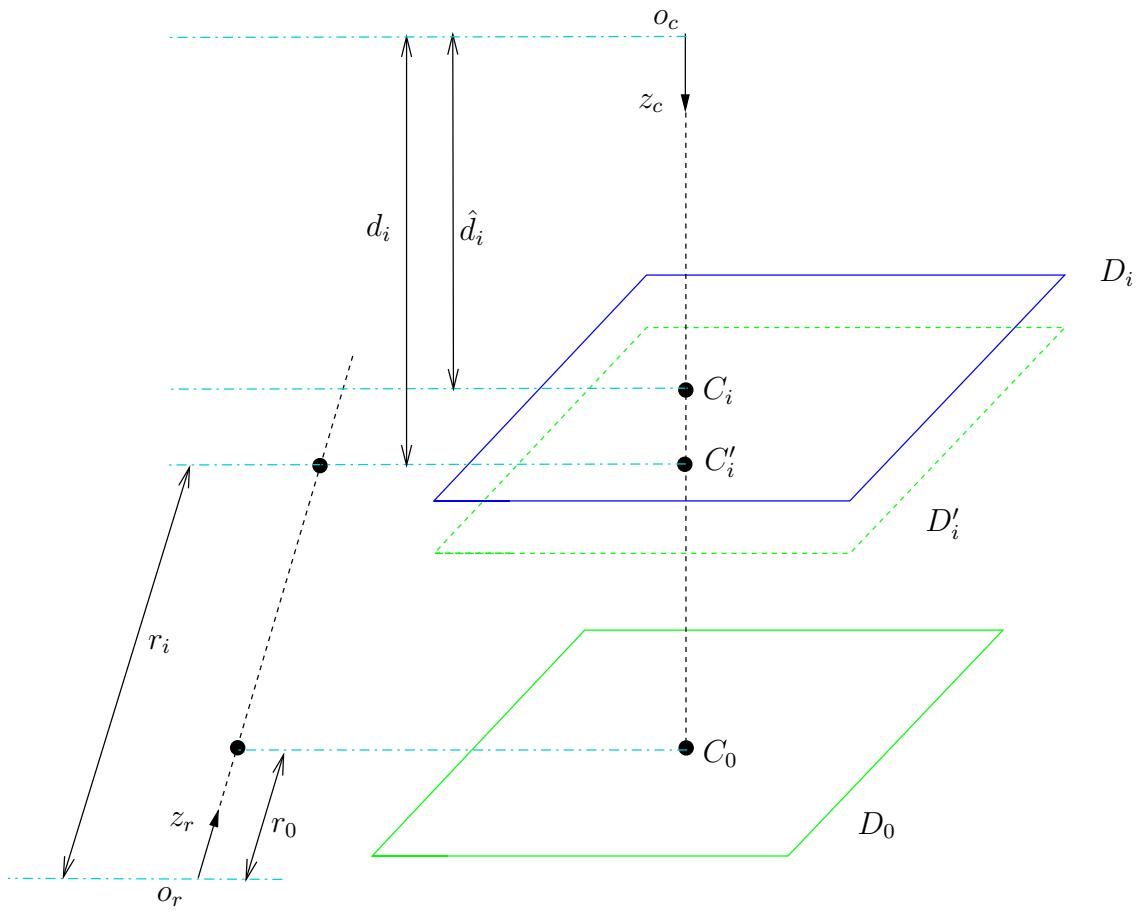


图 2.17 精度测量原理图

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 中拟合出相机坐标系下平板的平面方程的参数 Θ_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 θ'_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):

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

图 2.18 深度图质量测试实验结果

其中 δ_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 $1 + 19$ 组数据，测得每组数据深度信息的三个指标，如图2.18所示。从图2.18可以看出，随着距离的增加，深度相机的填充率、精度下降，噪声增加。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。

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. 2016)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

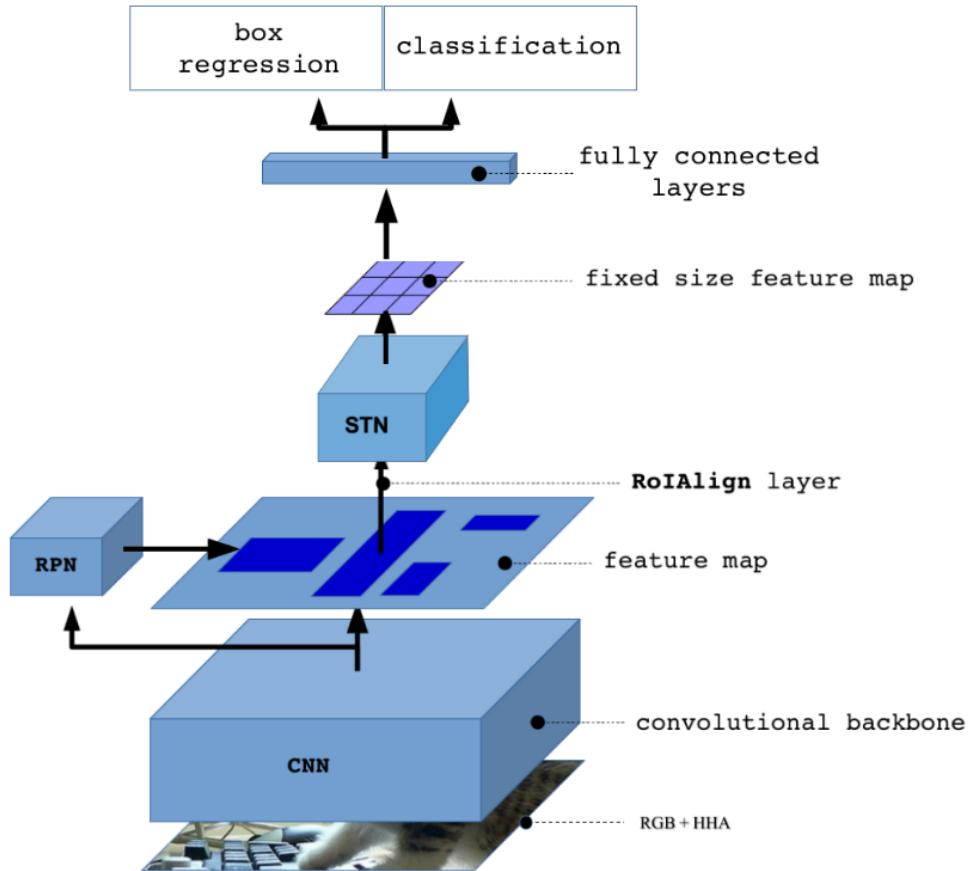


图 3.1 3D Faster R-CNN 结构

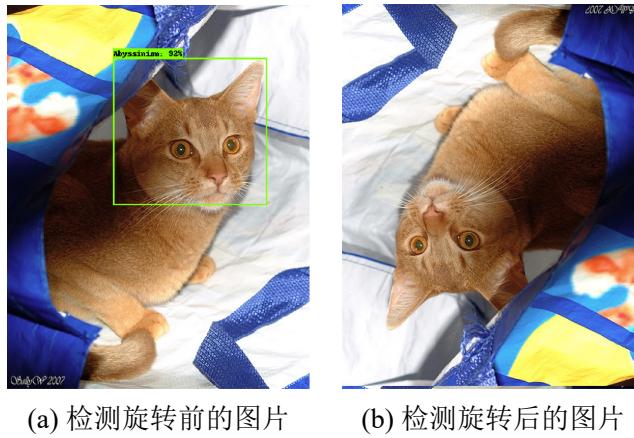


图 3.2 Faster R-CNN 检测识别宠物猫示例

- 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

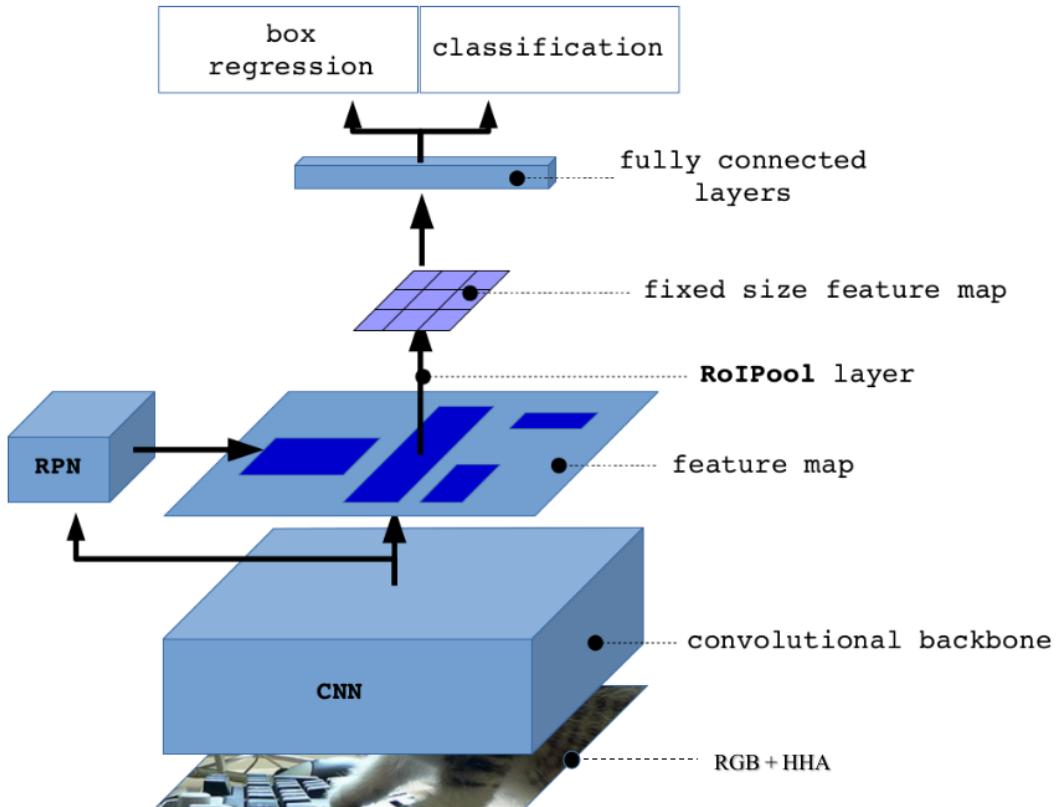


图 3.3 Faster R-CNN 结构

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 $H_{h \times w}$

$$h_{floor} = 1/d_{ceil}, h_{ceil} = 1/d_{floor};$$

for $y \leftarrow 1$ **to** h **do**

for $x \leftarrow 1$ **to** w **do**

$$H[y, x] = 1/D[y, x];$$

$$H[y, x] = (H[y, x] - h_{floor}) / (h_{ceil} - h_{floor});$$

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 $H_{h \times w}$

for $y \leftarrow 1$ **to** h **do**

for $x \leftarrow 1$ **to** w **do**

$$p = {}^W_C R P[y, x] + {}^W_C T;$$

$$H[y, x] = p.z;$$

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 $A_{h \times w}$

for $y \leftarrow 1$ **to** h **do**

for $x \leftarrow 1$ **to** w **do**

Calculate surface normal ${}^C\vec{n}$ at point $P[y, x]$;

${}^W\vec{n} = {}^W_C R {}^C\vec{n} + {}^W_C T$;

$A[y, x] = \arccos(-(\vec{n} \cdot \vec{o}_z) / (\|\vec{n}\| \|\vec{o}_z\|))$;

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}_θ 的参数 θ ,参数 θ 的个数由空间变换的类型决

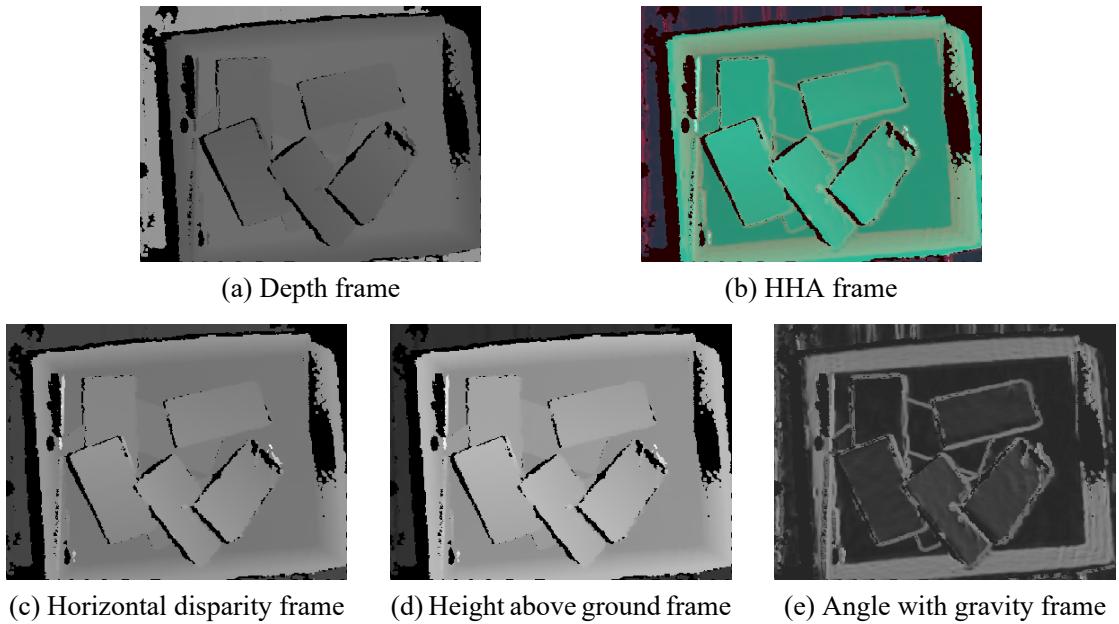


图 3.4 HHA 可视化效果图

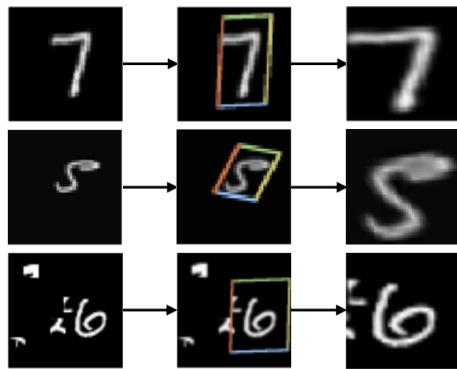


图 3.5 Spatial Transformer 效果图

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}_θ 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 θ 和 G 我们就可以在输

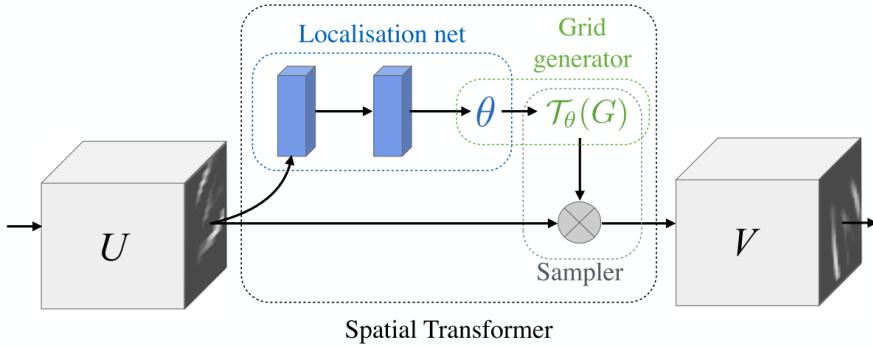


图 3.6 Spatial Transformer 结构图

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}_θ ,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 Φ_x 和 Φ_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

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

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

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

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。

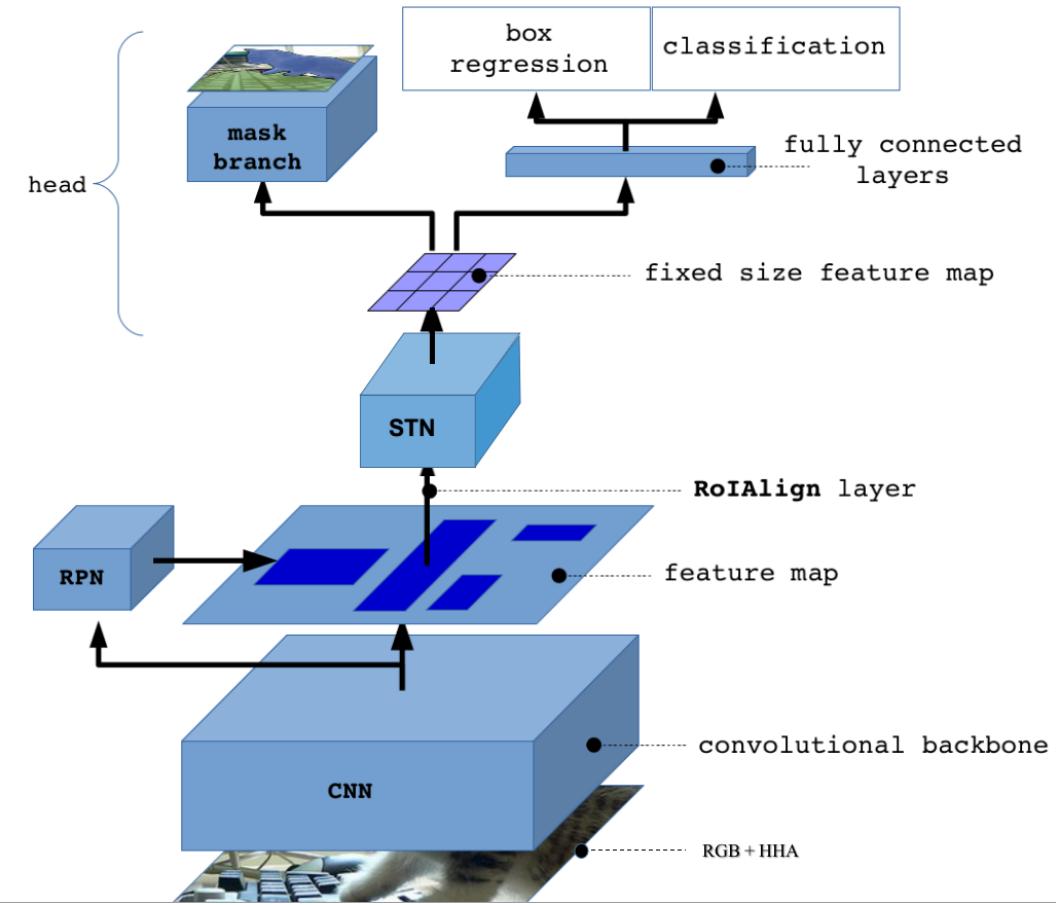


图 3.7 3D mask R-CNN 结构

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(?)，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 3×3 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 3×3 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 7×7 卷积核，其参数总量为 $49C^2$ 。而且 3×3 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(?) 和 FPN(?) 两部分构成。ResNeXt 是对残差网络 ResNet(?) 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题的背后逻辑在于此：想象

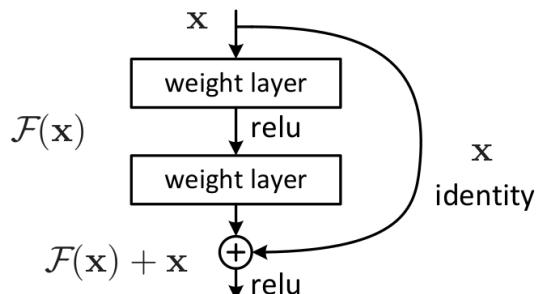


图 3.8 残差模块

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 1×1 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，和 200 层的 ResNet 准确度差不多，但是计算量只有后者的一半。

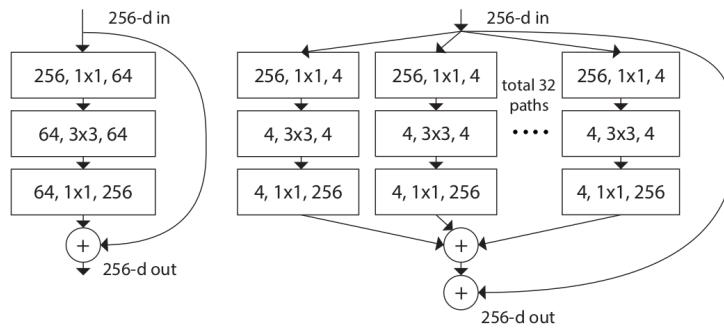


图 3.9 ResNeXt 对 ResNet 的改进

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 8×8 的特征图中提取提取 2×2 的输出, ROIPooling 的示意图如3.10所示, ROIAlign 的示意图如3.11所示。

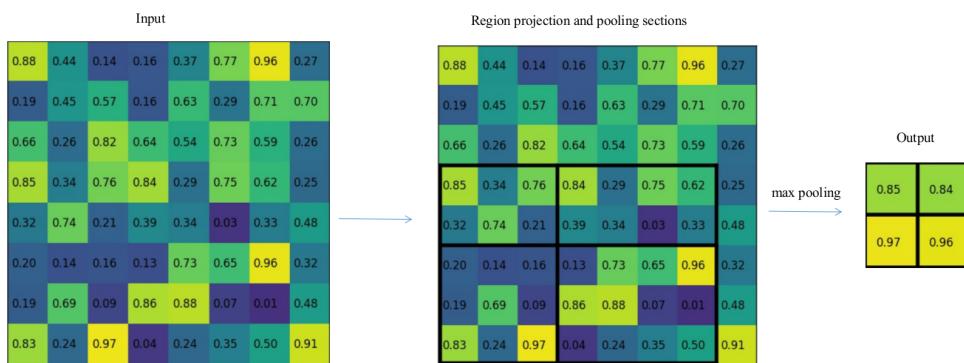


图 3.10 ROIPooling 示意图

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

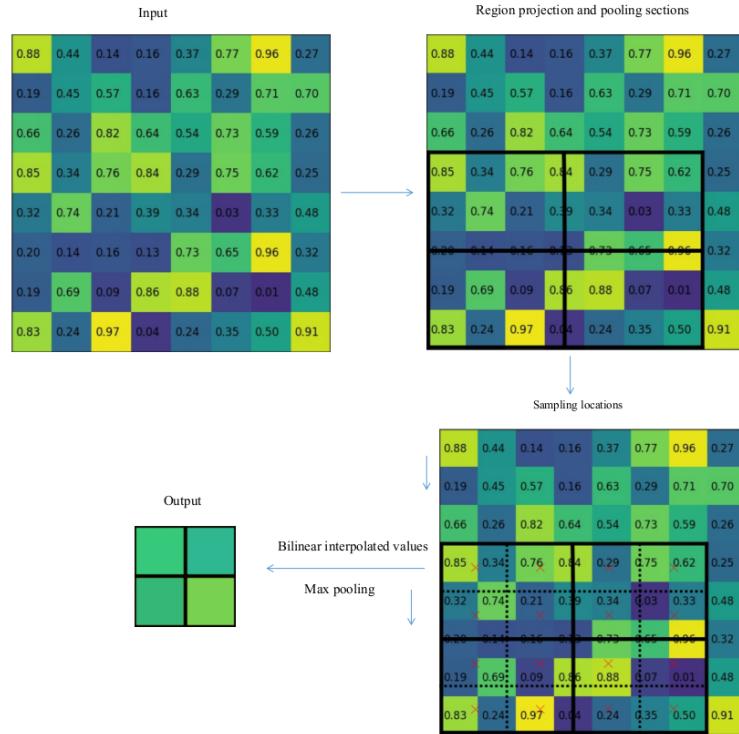


图 3.11 ROIAlign 示意图

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。

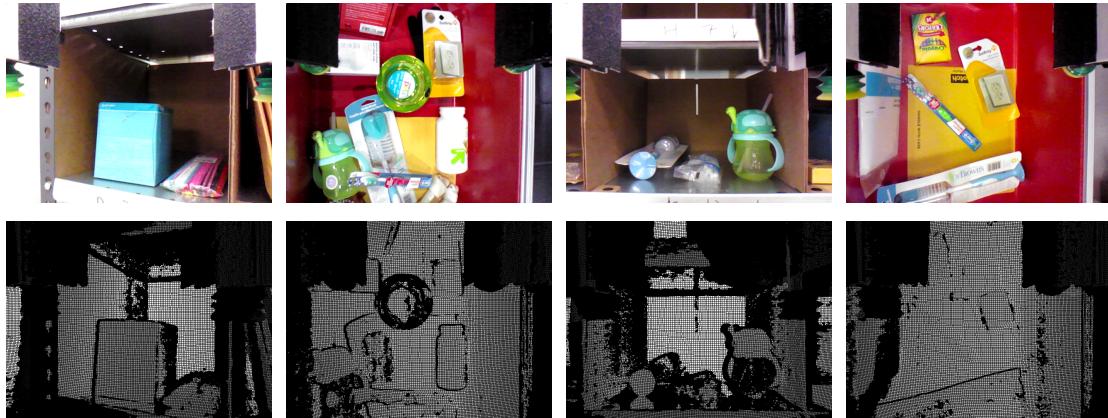


图 3.12 APC 数据集部分数据: 第一栏为彩色图像, 第二栏为与彩色图像相匹配的深度图

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

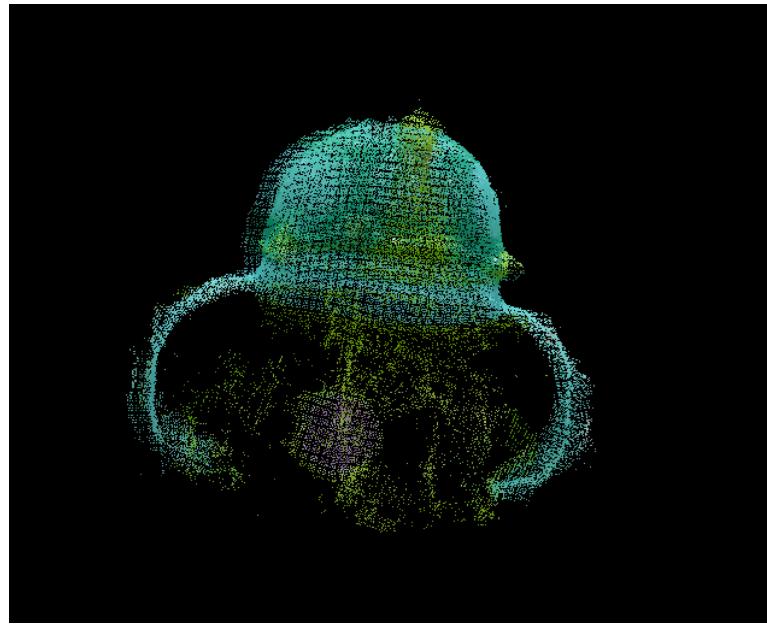


图 3.14 easter turtle sippy cup point cloud

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

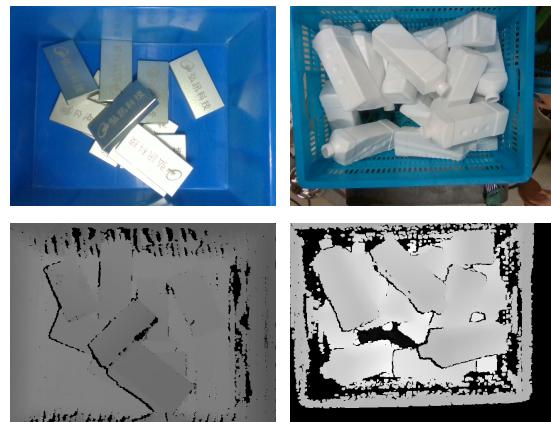


图 3.15 workpiece 数据集部分图片

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^①。

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:

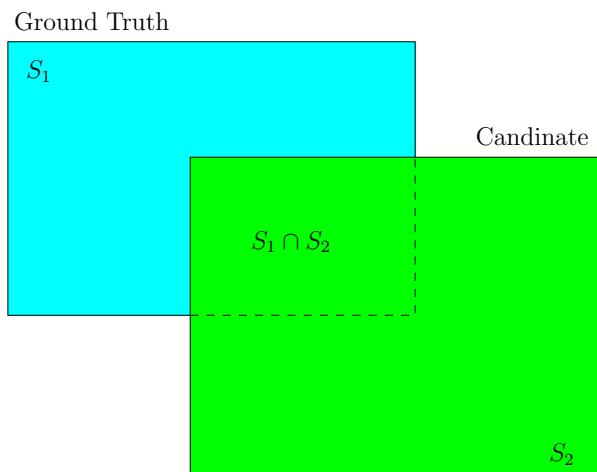


图 3.16 bounding box 交并比

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

$AP_{0.5}$ 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

^① 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
3D Faster R-CNN	RGB+HHA	bbox	34.55	57.99	34.69
Mask R-CNN	RGB	mask	32.34	55.78	33.12
3D Mask R-CNN	RGB+HHA	mask	33.94	56.45	33.99

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
3D Faster R-CNN	RGB+HHA	bbox	32.39	56.37	33.54
Mask R-CNN	RGB	mask	16.12	35.95	18.74
3D Mask R-CNN	RGB+HHA	mask	30.98	53.74	32.19

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

图3.18 算法workpiece数据集上部分检测结果

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	3D Faster R-CNN	Mask R-CNN	3D Mask R-CNN
FPS	5	3	4	2

表3.3 算法时间性能

3.4 本章小结

第 4 章 基于点云的位姿估计算法

4.1 ICP

4.2 Super4PCS

第5章 实验验证

第 6 章 结论与展望

6.1 结论

6.2 进一步工作的方向

致谢

逾尺的札记和研究纪录凝聚成这么薄薄的一本，高兴和欣慰之余，不禁感慨系之。记得鲁迅在一篇文章里写道：“人类的奋战前行的历史，正如煤的形成，当时用大量的木材，结果却只是一小块”。倘若这一小块有点意义的话，则是我读书生活的最好纪念，也令我对于即将迈入的新生活更加充满信心。回想读书生活，已经整整二十个年头，到同济求学将近五年，攻读博士学位也已三年了。进入同济大学以来，深深醉心于一流学府的大家风范。名师巨擘，各具特点；中西融合，文质相顾。处如此佳境以陶铸自我，实乃人生幸事。

2018 年 3 月

参考文献

- [1] BROWN D C. Decentering distortion of lenses[J/OL]. Photogrammetric Engineering and Remote Sensing, 1966. <https://ci.nii.ac.jp/naid/10022411406/en/>.
- [2] GEIGER A, ROSER M, URTASUN R. Efficient Large-Scale Stereo Matching[J]. Accv, 2010.
- [3] GOOGLE, 2012. Tango[EB/OL]. <https://developers.google.com/tango>.
- [4] GUPTA S, ARBELAEZ P, MALIK J, 2013. Perceptual Organization and Recognition of Indoor Scenes from RGB-D Images[C/OL]//2013 IEEE Conference on Computer Vision and Pattern Recognition. IEEE: 564–571. <http://ieeexplore.ieee.org/document/6618923/>. DOI: 10.1109/CVPR.2013.79.
- [5] GUPTA S, GIRSHICK R B, ARBELÁEZ P A, et al. Learning Rich Features from {RGB-D} Images for Object Detection and Segmentation[J/OL]. Computer Vision - {ECCV} 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part {VII}, 2014, 8695: 345–360. http://dx.doi.org/10.1007/978-3-319-10584-0_23. DOI: 10.1007/978-3-319-10584-0_23.
- [6] HE K, GKIOXARI G, DOLLÁR P, et al., 2017. Mask R-CNN[EB/OL]. <http://arxiv.org/abs/1703.06870>.
- [7] 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.
- [8] IMAGENET, 2011. Imagenet[EB/OL]. <http://www.image-net.org>.
- [9] KNUTH D E. The TeX book[M]. 15th ed. Reading, MA: Addison-Wesley Publishing Company, 1989.
- [10] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Alexnet[J]. Advances In Neural Information Processing Systems, 2012: 1–9. DOI: <http://dx.doi.org/10.1016/j.protcy.2014.09.007>.
- [11] LOOP C, Zhengyou Zhang. Computing rectifying homographies for stereo vision[J/OL]. Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), 2001, 1: 125–131. <http://ieeexplore.ieee.org/document/786928/>. DOI: 10.1109/CVPR.1999.786928.
- [12] MICROSOFT, 2012. Kinect[EB/OL]. <https://www.xbox.com/en-US/xbox-one/accessories/kinect>.
- [13] MIT-PRINCETON, 2016. "shelf & tote" benchmark dataset for 6d object pose estimation [EB/OL]. <http://apc.cs.princeton.edu/#shelf-and-tote-benchmark-dataset>.
- [14] REN S, HE K, GIRSHICK R, et al. Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks[M]. [S.l.: s.n.], 2016: 1–14.
- [15] 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.
- [16] 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

可能需要补充的内容……

个人简历、在学期间发表的学术论文与研究成果

个人简历

同济人，男/女，xxxx 年 xx 月生。

xxxx 年 xx 月毕业于 xxxx 大学 xxxx 专业获 xx 学位。

xxxx 年 xx 月入同济大学攻读 xx 学位。

已发表论文

[1] ...

[2] ...

[3] ...

已获得专利

[1] ...

[2] ...