中国科学院大学

University of Chinese Academy of Sciences

博士学位论文

基于傅里叶梅林变换的视觉机器人定位算法的研究

作者姓名：徐晴雯

指导教师：Sören Schwertfeger　研究员

上海科技大学

学位类别：工学博士

学科专业：通信与信息系统

培养单位：中国科学院上海微系统与信息技术研究所

2021 年 6 月

Fourier-Mellin Transform for Robot Visual Localization

A dissertation submitted to the

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in Communication and Information Systems

By

Xu Qingwen

Supervisor: Professor Sören Schwertfeger

Shanghai Institute of Microsystems and Information Technology

June, 2021

中国科学院大学
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的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对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或致谢。本人完全意识到本声明的法律结果由本人承担。

作者签名：

日　　期：

中国科学院大学
学位论文授权使用声明

本人完全了解并同意遵守中国科学院大学有关保存和使用学位论文的规定，即中国科学院大学有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：　　导师签名：

日　　期：　　日　　期：

摘 要

近年来随着视觉里程计 (Visual Odometry, VO) 技术的发展以及相关算法的开源，基于视觉的定位技术在机器人领域越来越流行。VO 在机器人中的应用范围很广，但是在一些特定场景下可能会遇到挑战，如浑浊的水下、能见度较低的大雾天气、特征匮乏的场景等。由于缺乏清晰的纹理，传统的 VO 算法可能在这些场景下定位失败。为了应对这些挑战，本文利用傅里叶梅林变换 (Fourier-Mellin Transform, FMT) 来进行图像的运动估计。FMT 是一种基于整体外观的频域图像配准算法，比基于特征或者光度一致性的算法更加鲁棒，但是 FMT 的一个缺点是：它要求图像上所有的像素点到成像平面的距离相同，即单一深度的图像。

考虑到 FMT 的局限性，本文将从两个方面改进 FMT 并将其应用到 VO 算法中，一是从原图像中划分子图，使子图可以满足 FMT 的要求；二是从 FMT 本身出发，重新考虑平移和缩放估计。具体来说，本文的主要贡献总结如下：

• 提出了基于 FMT 的全向图像匹配。首先将连续的两帧全向图像转换为全景图像，利用递归划分子图策略从全景图像上提取子图集合；然后利用 FMT 计算出对应子图之间的运动向量，从而得到全向图像之间的运动流场，即全向图像匹配；接着，本文利用这些匹配的一致点对进行了全向相机的位姿估计。实验表明该算法的性能比其他特征点法和光流法都更鲁棒和准确。

• 基于 FMT 估计得到的子图之间的运动向量，提出了一种新颖的全向相机姿态估计算法。不同于传统的几何法，如五点法，本文从全向相机本身的特点出发，将相机的旋转估计建模成正弦曲线拟合问题。相关实验表明该正弦曲线拟合算法比传统的几何法更加鲁棒，主要得益于两点：一是 FMT 在单一深度子图的匹配上提供了准确的运动估计，二是当 FMT 在多深度子图上不能给出准确配准结果时，正弦曲线拟合算法可以滤除这些离群点。

• 在应用 FMT 算法的过程中，本文发现了一个有趣的规律：在单一深度场景下，FMT 的相移图上只有单个峰；而在多深度场景下，当相机发生平移时，FMT 的相移图上将出现多个能量值较大的位置，且这些位置在同一条直线上。基于该发现，本文提出了扩展傅里叶梅林算法 (extended Fourier-Mellin Transform, eFMT)，将 FMT 的应用扩展到多深度场景。具体而言，本文把在相移图上的单

峰检测改成寻找能量值总和最大的线，即能量向量。最后本文实现了一个基于eFMT 的 VO 算法。由于单目 VO 需要保证连续运动估计之间的尺度一致性，本文利用能量向量的图样匹配实现了这一点。文中的相关实验表明，基于 eFMT 的 VO 算法比当下流行的 VO 框架更加鲁棒，主要是因为它保留了 FMT 算法良好的鲁棒性。

关键词：傅里叶梅林变换，视觉里程计，位姿估计，全向视觉，正弦曲线拟合

Abstract

With the development of visual odometry (VO) and related open-source algorithms, vision-based robot localization is becoming more and more popular. VO is applied in various scenarios, in which it may face challenges in certain environments, such as underwater turbidity, foggy weather with low visibility or feature-deprived settings. Traditional methods often fail here, due to the lack of clear textures. To overcome these challenges, we exploit the Fourier-Mellin Transform (FMT) to estimate the motion between images. FMT is a spectral method for image registration that is based on holistic descriptors and thus more robust than approaches using features or brightness consistency. But one of the drawbacks of FMT is that it requires all pixels in an image have the same distance to the imaging plane, i.e. single-depth images.

This thesis improves FMT in two major aspects and then applies it to visual odometry. One is to extract sub-images from original images, such that the sub-images meet the requirements of FMT. Another is to extend the application of FMT to multi-depth environments by rethinking its translation and zoom estimation. Concretely, the contributions of this work are summarized as follows:

• We propose omni-directional image matching based on FMT. For that, we first convert the omni-directional images from two consecutive frames to panorama images, from which we then extract sub-image sets. Afterwards, we apply FMT to calculate motion vectors from the corresponding sub-images of the two panorama images. Then we construct a motion flow field based on these motion vectors, i.e., omni-directional image matching. Additionally, this work utilizes these matched concordant points for omni-directional camera pose estimation. The experiments show the superior performance of our method compared to other feature-based approaches and optical flow by applying them to a pose estimation task.

• We propose a novel rotation estimation algorithm for omni-directional cameras based on the motion vectors calculated by FMT. Different from geometry methods like the five-point algorithm, this work models the rotation estimation of omni-directional

cameras as sinusoidal fitting based on the properties of omni-directional cameras. The experiments show that the proposed method is more robust than the traditional geometry-based algorithms. The main reason for the robustness of the sinusoidal fitting approach owes to two points: one is that FMT provides accurate motion estimations on single-depth sub-images and another is that sinusoid fitting is very robust to outliers.

- We made a very interesting observation: There is a single peak in the phase shift diagram of FMT in single-depth scenarios. But in multi-depth environments, when the camera is translating, there are multiple high energy values lying in one line in the phase shift diagram of FMT. Based on this observation, we propose the extended Fourier-Mellin Transform (eFMT), that extends FMT to multi-depth scenarios. Specifically, eFMT finds the line with the maximum sum of energy, instead of a single peak, in the phase shift diagram. Since monocular VO algorithms like this one are up to an unknown scale factor, we need to re-scale between consecutive motion estimates. eFMT does this via pattern matching on the extracted energy vectors. Our experiments show that eFMT-VO is more robust than current popular visual odometry frameworks because eFMT maintains the superior robustness of FMT.

Keywords: Fourier-Mellin Transform, Visual Odometry, Pose Estimation, Omni-directional Vision, Sinusoidal Fitting

目 录

图形列表

表格列表

第 1 章　绪论

近年来，机器人技术的快速发展使其在生产生活的各个方面得已广泛应用，从工业 4.0[6,7]，到农业自动化[8,9] 以及智能家居[10]。在多种应用中，机器人的自主性直接影响着其工作效率和应用范畴，而自主定位是其中的关键性任务之一。比如扫地机器人[11]、农用喷洒无人机[12,13]、无人驾驶汽车[14] 在运行时都需要先估计自身的位姿，然后才能进行路径规划与运动控制，从而完成上层任务。在目前的很多机器人产品中，其自主定位主要依赖于多传感器融合，而根据应用场景的不同，所选用的传感器也不同。其中水下机器人上通常配备多普勒测距仪（Doppler Velocity logs, DVL）、数字罗盘、声呐等传感器；室外无人机常用全球定位系统 (Global Positioning System, GPS) 和相机作为主要传感器；室内服务机器人上一般会装有 2D 激光测距仪（Laser Range Finder, LRF）、相机、惯性测量单元 (Inertial Measurement Units, IMU) 和里程计；无人车上会装有 GPS、3D LRF、相机等来保证其定位的准确性。

在上述的不同机器人应用中，这些传感器各有优缺点，如高精度的激光测距仪、IMU 一般比较昂贵；差分 GPS 的室外定位精度可以达到厘米级，但是其在室内环境相关的应用中受到了限制；里程计虽然可以方便地给出机器人的位姿，但其累计误差较大。与上述传感器相比，相机的价格相对低廉、体积较小、功耗也较小，使其在众多的传感器中脱颖而出，从而在多种场景中被应用于不同的机器人上。早期相机在机器人上多被用于监测、校准等任务，而近几年由于基于相机的定位技术的发展，越来越多的机器人将其用作主要的定位传感器。此外，为了获得更加鲁棒的定位效果，许多应用都采用了全向镜头或者鱼眼镜头来扩大相机的视野。

本文研究的主要问题是利用傅里叶梅林变换 (Fourier-Mellin Transform, FMT) 来估计相机的位姿。在一段连续的图像序列中，通过相机图像来确定相机位姿的方法被称为视觉里程计 (Visual Odometry, VO)。该技术在机器人领域已经被研究了几十年，其基础理论已经较为成熟。本章先对 VO 技术进行了综述：1.1节概述了视觉里程计的基本原理；1.2节回顾了已有的相关工作并介绍了当下比较流行的几类视觉里程计方法。接着，本章介绍了图像配准技术——FMT，该技术与当

下流行的 VO 算法的前端不同，不依赖于特征点或者像素的灰度不变性，而是考虑了整张图像的内容，从而使得配准结果更为准确和鲁棒，其中1.3节介绍了该图像配准算法的基本原理，1.4节回顾了该算法的相关工作，及其在 VO 中的初步应用。然后，1.5节总结了当下 VO 的热点与难点，以及将 FMT 应用于 VO 的难点。为了解决其中的部分难点问题，1.6节提出了本文的研究动机与研究内容。

1.1 视觉里程计的基本原理

VO 的主要研究问题是利用相机采集到的图像序列来估计相机的位姿。本小节将基于文献[15] 中的内容介绍视觉里程计的基本原理，包括相机模型、特征匹配、运动估计和后端优化。1.2小节将对过去几十年相关的研究工作进行回顾以及归纳总结。

1.1.1 相机模型

本文将会涉及到三种不同的相机模型，分别是：针孔模型、折反射模型和圆柱模型。其中，针孔模型主要用于普通的透射投影相机，折反射模型和圆柱模型多用于鱼眼相机和全向相机。

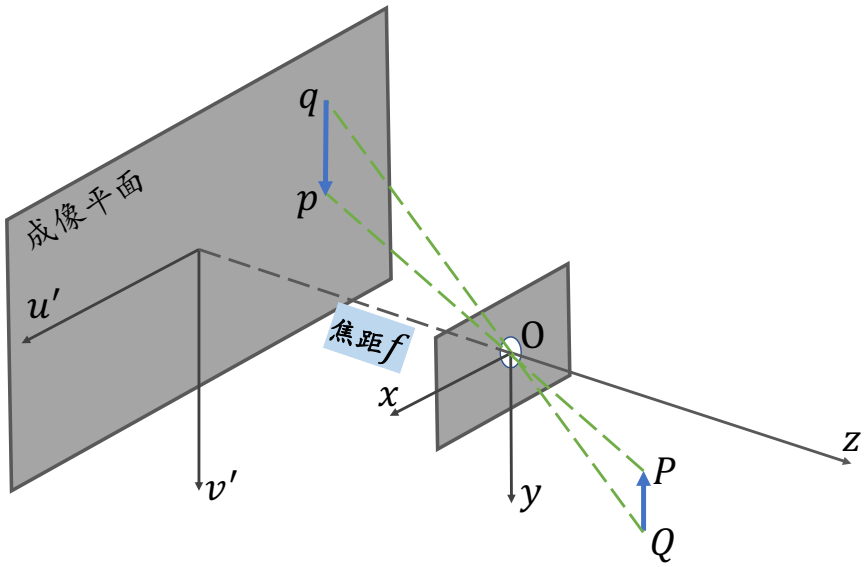


图 1.1 针孔相机模型

Figure 1.1 Pinhole camera model

针孔相机模型：又称为小孔相机模型。如图1.1所示，光线从小孔中穿过，落在成像平面上，形成了图像。换而言之，相机将三维空间的点映射成了二维的像

素坐标，比如图中的三维点 P 经过经过相机的光心 O，映射到了成像平面的像素 p。以相机坐标系为参考坐标系，三维点 $P=[x,y,z]^\top$ 和二维像素 $p=[u,v]^\top$ 之间的关系可以表示为

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \lambda \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{1.1}$$

其中，f_x、f_y 为相机的焦距，c_x、c_y 为像素平面的中心。需要说明的是图1.1展示的小孔模型只考虑了一个焦距 f，该焦距的单位一般为 mm，被称为毫米焦距，而公式(1.1)中的焦距 f_x、f_y 指的是像素焦距，其单位为像素。毫米焦距和像素焦距之间可以通过系数 dp（mm/pixel）来进行转换，由于制造工艺的影响，水平和竖直方向上的 dp 值不一定相同，因而公式(1.1)中会包含水平和竖直方向上的两个焦距 f_x、f_y。

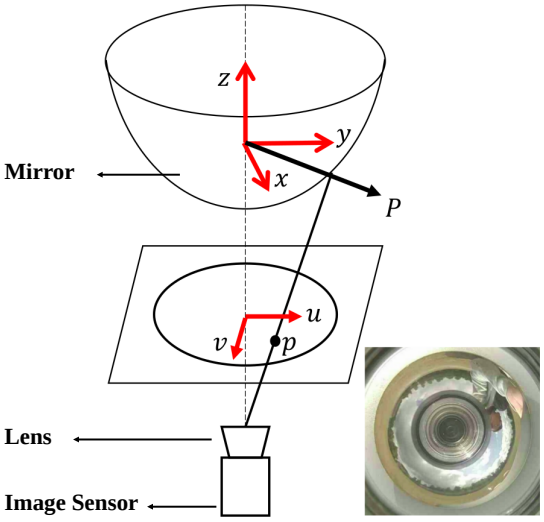


图 1.2 全向相机模型示意图，参考[1]

Figure 1.2 Omni-directional camera model according to[1]

折反射相机模型：本文用到的折反射相机模型主要基于文献[1] 中提出的全向相机模型。如图 1.2所示，文献[1]中的全向相机模型描绘了图像像素与相机光线之间的对应关系。该工作有两个基本假设：（a）相机中心与全向镜头中心对齐；（b）全向镜头对称旋转。假设全向图像上一点 p 的坐标为 (u,v)，以全向图像中心为坐标系原点，则该像素点 p 对应的归一化三维点 P ，即相机光线的方

向向量，可表示为

$$P=\begin{bmatrix}x\\y\\z\end{bmatrix}=\begin{bmatrix}\alpha u\\\alpha v\\f(u,v)\end{bmatrix}=>\begin{bmatrix}u\\v\\f(u,v)\end{bmatrix}=\pi^{-1}(p)\ . \tag{1.2}$$

由于全向镜头是对称旋转的，因此 $f(u,v)$ 只依赖于像素点 p 到全向图像中心的距离 $\rho=\sqrt{u^2+v^2}$ 。$f(\rho)$ 可以由一个高阶多项式表示成：

$$f(u,v)=f(\rho)=\alpha_0+\alpha_1\rho+\alpha_2\rho^2+\alpha_3\rho^3+\ldots \tag{1.3}$$

来描述不同类型的全向镜头。最后，该模型利用一个额外的仿射变换来补偿假设可能引入的误差，比如相机与图像中心不完全对齐引入的误差、相机成像过程中像素不是方的引入的误差等。该仿射变换表示如下：

$$\begin{bmatrix}u^*\\v^*\end{bmatrix}=\begin{bmatrix}c&d\\e&1\end{bmatrix}\begin{bmatrix}u\\v\end{bmatrix}+\begin{bmatrix}x_c'\\y_c'\end{bmatrix}, \tag{1.4}$$

其中，(x_c',y_c') 是全向图像 I_o 的中心，(u^*,v^*) 表示以全向图像 I_o 左上角为原点时 p 点的坐标。

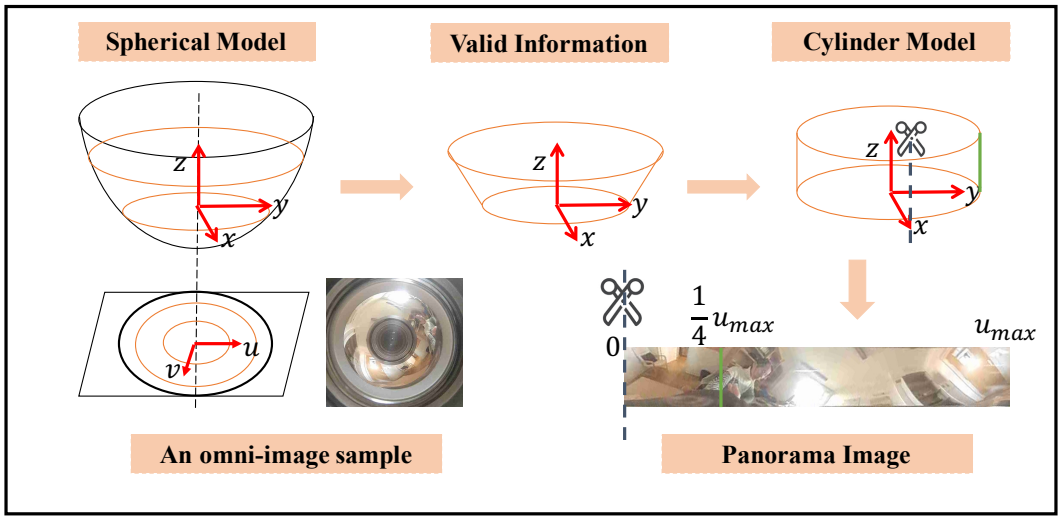


图 1.3 圆柱相机模型

Figure 1.3 Cylindric camera model

圆柱模型：如文献[16]中所述，折反射全向相机采集的全向图像通常被转化为圆柱全景图像。图1.3给出了从球形模型到圆柱模型的直观描述以及从全向图像到全景图像之间的变换，该变换主要通过映射和插值实现。图1.3中先将全向

图像上的有效信息转换为圆柱图像，圆柱图像和全景图像之间的转换可以描述为

$$u = r \cdot \theta = r \cdot \arctan(\frac{y}{x}) , \tag{1.5a}$$

$$v = \frac{H}{2} - z , \tag{1.5b}$$

和

$$x = r \cdot \cos\theta = r \cdot \cos\frac{u}{r} , \tag{1.6a}$$

$$y = r \cdot \sin\frac{u}{r} , \tag{1.6b}$$

$$z = \frac{H}{2} - v . \tag{1.6c}$$

当使用公式(1.5)和(1.6)将全向图展开为全景图时，无法确保每个像素都是正方形，即可能存在分辨率不一致的情况。换句话说，单位像素在 u 和 v 方向上的对应的入射角度可能会不相同。本文使用一种简单的校准方法来找到 u 和 v 方向上像素对应的入射角比值，即分辨率一致性因子。具体来说，基于像素均匀分布的假设以及 u 方向上的总和角为 360° 的事实，可以简单地计算出 u 方向上每个像素对应的角度；然后通过标定方格计算出图像 v 方向上对应的角度总和，从而可以推算出 v 轴方向上每个像素对应的角度，然后得出 u 和 v 轴方向上像素分辨率的一致性因子。

1.1.2　特征提取与匹配

1.2中将会介绍，特征匹配并不是 VO 算法中的必要步骤，比如基于直接法的 VO 会直接通过像素的光度误差来估计相机运动。而对于基于特征法的 VO 而言，特征的提取与匹配则是必不可少的一步。本小节将只介绍相关的特征提取与匹配方法，在下一小节的运动估计中再介绍利用像素的光度误差估计相机运动。

图像的特征一般可以分为两大类：角点和 blob。角点是两条或多条线的交点，对应着图像上的单个像素，常见的角点有 Harris[17]、Shi-Tomasi[18]、SUSAN[19] 和 FAST[20] 等；blob 不是图像上某个像素，而是图像上的某个区域，其强度和纹理都和邻居区域有着较大的区别，常见的 blob 有 SIFT [21,22]、SURF [23,24]、ORB [25]、KAZE 和 AKAZE [26,27] 等。相比之下，角点的优势在于它在图像上的位置精度高，对应着单一像素；blob 的优势是具有唯一性，因为它描述的是部分

区域的特征，比单一的像素拥有更多的信息。因此，当两帧图像之间的变化较小时，角点能够更快更准确地找到匹配点；当两帧图像变换较大时，采用 blob 来描述图像的特征更为鲁棒。

在提取完特征后，需要对两帧图像上的特征进行匹配。当两帧之间的变化不大时，可以利用光流法[28] 计算出第一帧图像上的特征 x_1 在第二帧图像上的对应点 x_2，这一方法主要基于光度不变假设，即很短时间内对应像素的强度不变：

$$I(u, v, t) = I(u + \Delta u, v + \Delta v, t + \Delta t) \,. \tag{1.7}$$

当两帧之间的运动较大时，光流法的基本假设不一定成立，因而容易失败，此时多采用 blob 提取特征，如 SIFT、SURF 和 ORB 等。这类方法在提取完特征后，往往为每个特征计算一个描述子，通过计算描述子之间的距离可以度量两个特征之间的相似度。基于此，最简单的方法就是采用暴力匹配法计算出两帧图像之间的特征匹配结果。为了减少计算量，文献[29]提出通过最近邻的方法来加快特征匹配。

1.1.3 运动估计

根据匹配点对的维度，可以将运动估计分为三类：2D-2D、3D-2D 和 3D-3D。单目 VO 通常使用 2D-2D、3D-2D 的估计，3D-3D 的运动估计多出现于双目 VO 或点云配准中。考虑到本文的研究对象是单目 VO，因此本小节中主要对前者进行介绍。基于 2D-2D 匹配点对的运动估计主要是通过对应的像素点来估计图像之间的位姿变换，3D-2D 的估计则是在进行 2D 像素点和 3D 地图点之间的配准。本节首先分别介绍特征法和直接法使用的 2D-2D 运动估计方法，然后对常用的 3D-2D 运动估计方法进行概述。

在特征匹配后，特征法通常通过最小化重投影误差来估计两帧之间的位姿变换：

$$_{k-1}^{k}T = \arg\min_{T} \sum_{i} \|u_i' - u_i\|_{\Sigma}^{2} \,, \tag{1.8}$$

其中，u_i 与 u_i' 是三维空间中的一点 P_i 分别在第 $k-1$ 和 k 帧图像上的投影点，该点对可以由特征匹配得到。为了计算问题 1.8，可以先通过基于对极几何的方法求出本质矩阵，然后从本质矩阵中分解出旋转和平移，组成 $_{k-1}^{k}T$。常用的求解本质矩阵的方法有五点法[30] 和八点法[31] 等。在特征点对比较多时，可以采用 RANSAC 的方法来加快运算速度和提高鲁棒性。

直接法不提取特征点，直接通过最小化像素的光度误差来估计两帧之间的位姿：

$$ {}^{k}_{k-1}T = \arg\min_{T} \sum_{i} \| {}^{k}I(u'_i) - {}^{k-1}I(u_i) \|^2_{\sigma} , \tag{1.9} $$

其中 $I(\cdot)$ 代表像素的强度。由于直接法不提取特征点，无法直接得到对应的 u'_i 和 u_i，因此需要利用相机模型和概率深度模型构建 u'_i 和 u_i 之间关于变换 T 的关系方程：

$$ u'_i = \pi(T \cdot \pi^{-1}(u_i)d) , \tag{1.10} $$

其中 $\pi(\cdot)$ 为相机模型给出的 3D 点到 2D 像素的投影方程，d 为像素的深度，可由概率深度模型估计得到。

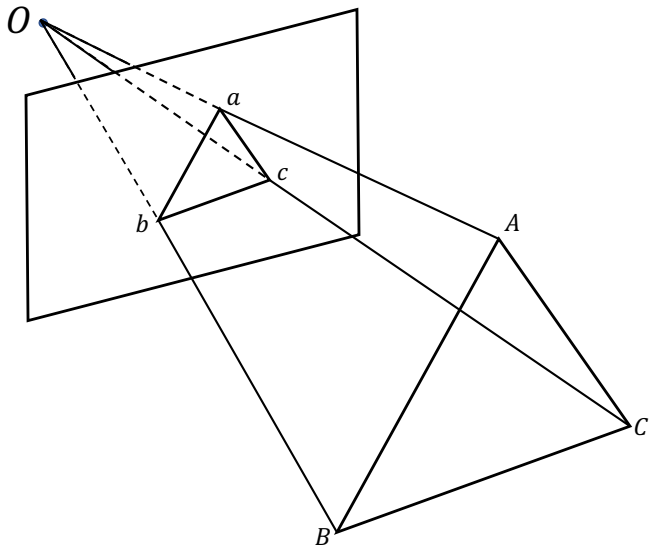


图 1.4 P3P 问题示意图[2]

Figure 1.4 Example of a P3P problem[2]

如果在估计相机位姿的过程中，同时生成和维护环境地图，那么可以利用 3D 地图点与 2D 像素点的匹配来估计相机坐标系和地图参考系之间的变换。该变换也可以通过最小化重投影误差来实现：

$$ {}^{k}_{k-1}T = \arg\min_{T} \sum_{i} \| u'_i - \pi(P_i) \|^2_{\Sigma} . \tag{1.11} $$

求解式1.11的最少需要 3 组匹配点对，常用方法有 PnP[32]、EPnP[33] 和 P3P[34] 等。本小节以 P3P 为例，参考文献[2]，解释如何从 3D-2D 匹配点中计算得出位姿变换。P3P 即使用三对 3D-2D 匹配点计算两帧之间的位姿变换。如图1.4所示，基

于余弦定理可得，

$$\begin{aligned} OA^2 + OB^2 - 2OA \cdot OB \cdot \cos < a, b >= AB^2 , \\ OB^2 + OC^2 - 2OB \cdot OC \cdot \cos < b, c >= BC^2 , \\ OA^2 + OC^2 - 2OA \cdot OC \cdot \cos < a, c >= AC^2 . \end{aligned} \tag{1.12}$$

令 $x = \frac{OA}{OC}$、$y = \frac{OB}{OC}$、$g = \frac{AB^2}{OC^2}$，则可以用 $m \cdot g$ 和 $n \cdot g$ 分别表示 $\frac{BC^2}{OC^2}$ 和 $\frac{AC^2}{OC^2}$。此时，式1.12变为

$$\begin{aligned} x^2 + y^2 - 2xy \cos < a, b > -g = 0 , \\ y^2 + 1^2 - 2y \cos < b, c > -mg = 0 , \\ x^2 + 1^2 - 2x \cos < a, c > -ng = 0 , \end{aligned} \tag{1.13}$$

其中，只有 x 和 y 是未知参数。以上方程可以通过吴消元法求解，求解的具体细节参见[2]。

1.1.4 后端优化

后端优化的第一步是闭环检测，只有在形成闭环的前提下，进行后端优化才具有意义。闭环检测本质上也是计算特征的相似程度，与特征匹配的原理类似，在实际应用中，通常使用词袋技术[35]来实现闭环检测，从而提高检测效率和鲁棒性。在检测到闭环后，后端优化的方式通常有两种：集束调整法和位姿图优化。前者同时优化相机位姿和路标点位置，后者只优化相机位姿，对应的目标函数分别为

$$\arg\min_{{}^iC} \sum_i \sum_j \| {}^iC - {}^i_jT \, {}^jC \|^2 \tag{1.14}$$

和

$$\arg\min_{P_i, {}^kC} \sum_i \| {}^ku_i - \pi(P_i, {}^kC) \|^2 , \tag{1.15}$$

其中，${}^kC$ 为第 k 帧图像对应的相机位姿。在实际求解时，可以通过 g2o[36]、ceres[37]等开源优化工具库进行计算，从而得到优化后的位姿点。

1.2 视觉里程计的相关工作回顾

本节对 VO 的相关工作进行了回顾。文献[38]中对各种相机模型进行了详细阐述，考虑到针孔相机模型及其标定已经非常成熟，且被广泛应用到各项视觉任务中，本节不再对其进行讨论，但是将对全向相机的模型与标定工作进行回顾与

总结；然后，本节将讨论以特征法和直接法为代表的 VO 算法的相关工作；最后，本节将回顾将这些基于针孔相机模型的 VO 方法拓展到全向相机的相关工作。

1.2.1　全向相机模型与标定

文献[39]将常见的全向相机总结为三类：反射相机、折反射相机和多视角相机（示例如图1.5，从左至右）。反射相机即鱼眼相机，它的视角可以达到 180° 及以上；折反射相机的构造通常包括一个普通的针孔相机和一个反射镜，反射镜的镜面形状可以是抛物线、双曲线以及椭圆形的，光线通过镜面反射后再经过相机光心、到达成像平面，折反射相机的水平视角一般为 360°，垂直视角主要取决于相机的设计，也可达到 100° 以上；多视角相机可以看做是由多个视野较小的相机组成的多相机系统，利用多个相机的视野覆盖整个环境，达到真正的全向：水平方向和垂直方向均为 360°。近几年比较流行的商业全向相机多为利用两个鱼眼相机组合成 360° 的视野，比如 Ricoh Theta V、Insta360 One X 等。

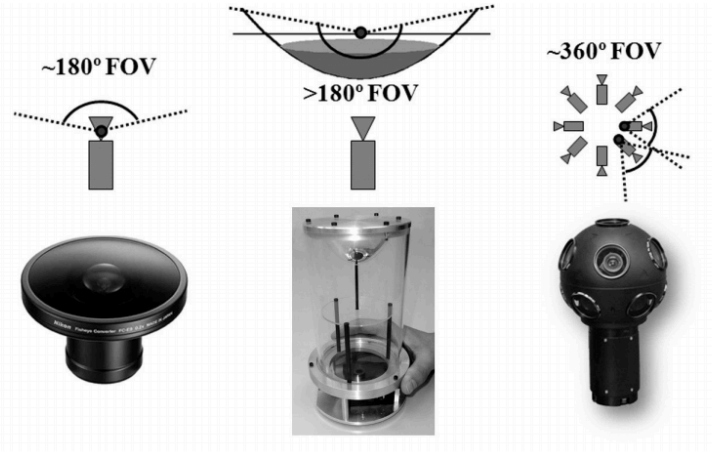


图 1.5 三类全向相机示例图

Figure 1.5 Three kinds of omni-directional cameras

文献[40]指出，传统的针孔相机模型无法描述这些视野较大的相机，因此出现了专门用于描述全向相机的模型。一类是用于各类折反射全向相机的通用模型，由[41]率先提出，之后[42]在其基础上进行了改进，在模型中考虑了径向畸变；[43] 中讨论了如何将该模型应用于鱼眼相机，但是结果显示标定精度较低，主要原因是折反射相机可以用精确的参数模型来表示，而鱼眼相机不可以。为了实现鱼眼相机的标定，[44]通过推广分割模型和多项式模型成功实现了其标定。另一类是由 Scaramuzza 等人提出的基于泰勒多项式的全向相机模型[1]，既可以用于折

反射相机的标定又可以用于鱼眼相机的标定，因此该方法的优势在于更加通用。这两类全向相机模型也衍生出了不同的标定方法与工具库。例如除了[41,42]中提出的标定方法，[45]基于单个投影中心的几何约束，提高了标定精度；Mei 等人通过增加参数来弥补现实世界与理论情形之间的误差，从而形成了完整的标定工具[46]。对于第二类模型，Scaramuzza 等人在同年开源了相应的 MATLAB 工具包[47]；之后 Urban 等人通过替换残差函数和联合优化所有参数，提高了标定的鲁棒性和准确性[48]。

以上方法在对全向相机建模时，均假设相机系统为单一有效的视点系统，即所有来自相机且被镜面反（折）射的光线相交于同一点。但是由于制造工艺的影响，有的全向相机不一定能满足该假设。[49,50]将全向相机模型描述为像素点和三维点的映射关系，该映射关系可以通过已知几何关系的标定网格来计算，这种映射关系的表述既可以用于单一有效视点相机也可以用于非单一有效视点相机。

1.2.2 单目视觉里程计

VO、从运动中恢复结构 (Structure from motion, SFM)、基于视觉的即时定位与建图 (Visual simultaneous localization and mapping, VSLAM) 三者常被认为是同一类任务，都是利用图像序列估计相机位姿并重建环境结构。其中 VO 主要关注相机位姿估计的准确性以及计算的实时性，其输入一般都是有序的图像；而 SFM 解决的则是更加一般的问题，从有序或者无序的图像集合中恢复相机位姿、重建三维环境地图，由于此类问题求解成本高、耗时长，通常都将其进行离线优化提高位姿估计和三维重建的精度；VSLAM 和 SFM 一样既需要估计相机位姿又需要重建三维环境，但输入和 VO 一样是有序图像集合，且要求较高的实时性。考虑到本文内容主要是 VO，本节接下来主要讨论 VO 的研究现状。此外，由于 VO 和 VSLAM 的相关工作重合度较高，在归纳总结时也会涉及到部分 VSLAM 工作。

按照描述环境的方式，VO 可以分为基于特征的、基于外观的和两者结合的三类方法[15]。基于特征的 VO 往往需要提取图像上显著且重复的区域，并为其构建相应的描述子；基于外观的方法不需要提取特征，主要依赖图像上的全部或者部分像素强度进行位姿估计；而两者混合的方法就是在估计位姿时既考虑帧间像素的一致性，也提取特征。基于特征的 VO 方法有很多，如[51–55]，其中[51]首次实现了大规模、实时的单目 VO，它采用了五点法来计算可能的相对位姿，并使

用 RANSAC[56] 技术从大量的匹配进行采样，从而加快了运算效率。之后，不少方法都开始采用五点法来进行位姿估计，比如[57,58]。五点法估算的是三维空间中的无约束 6 自由度 (Degrees of Freedom, DoF) 运动，而很多时候传感器安装的移动平台是受约束的，例如无人车、扫地机器人等，基于这些约束，文献[52,59]在进行位姿估计时只需要更少的点来进行求解。基于外观的 VO 方法包括[60–62]，其中[60,62]都利用了 FMT 来计算图像之间的运动向量，前者用其估计了车辆的位姿，后者基于估计的运动向量进行了水下图像拼接。文献[63]则采用了特征与外观结合的方式进行了相机的位姿估计，利用局部外观的相关性估算了相机的二维旋转，通过地面的特征估计了相机的平移。

视觉里程计在计算了两帧之间的相对位姿之后，只需要通过链式法则即可计算出相机在参考坐标系下的绝对位姿，但是在相机长时间运动后可能会产生较大的累计误差。为此，一些研究者对一段时间内的相机位姿通过集束调整法进行优化，从而减小了累计误差[64–66]。值得一提的是，近些年所提出的 VO/VSLAM 框架几乎都包含了闭环检测和位姿优化。

从计算相对位姿的角度，近几年比较流行的 VO/VSLAM 框架可以分为基于滤波、基于关键帧和基于直接法三类。Davision 等人提出了第一个基于单目相机的 VSLAM 系统——MonoSLAM [58]，该系统通过提取的特征点作为观测值，然后利用扩展卡尔曼滤波进行位姿的迭代更新，基于扩展卡尔曼滤波器的方法每次更新只更新当前时刻的位姿，不对之前时刻的位姿进行优化，因此所需的运算资源较少，但同时也会带来相应的累积误差。相比之下，基于关键帧的方法[67–69]一般都会进行集束调整，从而在大规模长时间的场景下更能提供较为准确的位姿估计。其中 PTAM (Parallel Tracking and Mapping) [67] 首次在 VO 后端中使用了优化的方法，并提出了定位、建图双线程的 SLAM 框架；其后，ORB-SLAM [69] 在此基础上改进了特征提取的方法、加入了闭环检测来提高其鲁棒性和定位精度；RDSLAM [70] 则是在 PTAM 的基础上对 RANSAC 方法进行了改进，从而实现了动态场景下的鲁棒定位。基于直接法的 VO 不提取特征，主要利用图像之间的光度不变性来估计位姿，因而其在特征缺失的情况下比需要提取特征的方法更鲁棒。最早提出的基于直接法的 VSLAM 是 DTAM (Dense Tracking and Mapping) [71]，其构建了稠密地图，但是由于运算开销较大，需要使用 GPU；近几年提出的 LSD-SLAM [72] 仅恢复半稠密地图，且每个像素点的深度可以独立

进行计算，从而提高了运算效率；而且，LSD-SLAM 的研究团队还推出了基于直接法的稀疏 VO：DSO (Direct Sparse Odometry)[73]，通过光度标定使得直接法更加鲁棒。由于 LSD-SLAM 和 DSO 所需的运算资源较小，它们都可以直接运行在 CPU 上。还有一类方法被称为半直接法，主要代表为 SVO (Semi-direct Visual Odomety)[74]，它在前端图像配准的部分使用了直接法，在位姿估计、集束调整时采用了最小化重投影误差的方法，从而实现了鲁棒且高效的 VO 算法。

1.2.3 基于全向相机的视觉里程计

全向相机因其视角较大、获得信息更多，被广泛应用于移动机器人中。文献[75]通过从全景图像中获得视觉线索，帮助机器人实现了自动归航；[76]提出基于全向相机的局部窗口集束调整法，用来恢复相机轨迹和环境地图；此外，文献[77]还总结了不少基于全向相机的应用。本节主要回顾基于全向相机的 VO 的相关工作。

和基于针孔相机的 VO 算法类似，基于全向相机的 VO 算法也可以分为基于像素、基于特征和基于外观的。基于像素的 VO 方法主要依赖于光度一致性。例如，[78]中在行星探测车上安装了一个折反射式的全向相机，在采集到的全向图像上通过光流的计算估算出探测车的 2D 运动。大部分基于全向相机的 VO 方法都是依赖于特征，如[63,79–81]。其中，[82]中通过全景图像提供的丰富的特征点实现了一个纯方位 SLAM 系统；[80]中比较了 FAST 角点和 SIFT 特征点，发现 SIFT 进行路标提取效果更好，且该工作在不使用集束调整的情况下，利用对极几何约束和三维地图信息，实现了基于全向相机的长时间的鲁棒定位；[79]中揭示了同一场景下采用全向相机的定位法相比于传统针孔相机更加精确。还有一些工作专门为全向相机的大畸变设计了改进的特征匹配方法[83,84]。也有部分基于全向相机的 VO 算法是依赖于外观的。[85]中比较了几种不同整体外观描述子对于全向相机定位效果的影响；[63]中计算全向相机的朝向时采用的就是基于外观的相位相关法。

随着制造工艺的提升，工业级甚至消费级的全向相机变得越来越流行，近几年的主流 VO/VSLAM 算法也都提供了全向相机接口。例如，ORB-SLAM 在最新的版本中加入了鱼眼相机模型[86]，但其使用的还是参数化的相机模型，因此不能用于视野过大的鱼眼相机；MultiCol SLAM[87] 也是 ORB-SLAM 关于全向相机的一个变种，其中采用的全向相机模型是文献[47] 中提出的全向相机模型，可

以适用于多种全向相机；SVO 在最新的版本中[88] 也支持该全向相机模型，可以应用在各类相机上。DSO 的团队也在[89] 中利用[41] 中所提出的相机模型将其算法扩展到全向相机上，不过该基于全向相机的 DSO 版本尚未开源。此外，还有一类基于全向相机的算法主要考虑的不是单个全向相机，而是由多个相机组成的全向采集系统，如上文中提到的 MultiCol SLAM[87] 也是为多相机系统设计的；[90]中也为由四个鱼眼相机组成的全向相机系统设计了鲁棒的 VO 算法，与标准 VO 系统相比，该工作的主要创新点是利用混合投影模型改进了特征匹配、多视角 P3P 以及在线外参更新。总得来说，这些算法大部分都是将 VO 中的针孔相机模型更换为全向相机模型，从而实现了基于全向相机的 VO 系统。

1.3　傅里叶梅林变换的基本原理

本节回顾了经典 FMT[91] 的主要思想。给定两帧图像信号 ${}^1I$、${}^2I$，其关系满足

$$\begin{aligned} {}^2I(x,y) = {}^1I(&zx\cos\theta_0 + zy\sin\theta_0 - x_0, \\ &-zx\sin\theta_0 + zy\cos\theta_0 - y_0) \end{aligned} \tag{1.16}$$

其中，z 和 θ_0 为常量，分别表示缩放和旋转；(x_0, y_0) 是 ${}^1I$ 和 ${}^2I$ 之间的平移。这些描述两帧图像之间的运动参数 (z, θ, x_0, y_0) 可由 FMT 估算得到，步骤如下：

(1) 对公式(1.16)两边的信号都进行傅里叶变换：

$$\begin{aligned} {}^2\mathcal{F}(\xi,\eta) = \ & e^{-j2\pi(\xi x_0+\eta y_0)}z^{-2} \\ & {}^1\mathcal{F}(z^{-1}\xi\cos\theta_0 + z^{-1}\eta\sin\theta_0, \ , \\ & -z^{-1}\xi\sin\theta_0 + z^{-1}\eta\cos\theta_0) \end{aligned} \tag{1.17}$$

(2) 将公式(1.17)两边信号的模 $\mathcal{M}$ 转换到极坐标系下，忽略系数的影响：

$${}^2\mathcal{M}(\rho,\theta) = {}^1\mathcal{M}(z^{-1}\rho, \theta-\theta_0)\ . \tag{1.18}$$

(3) 对公式(1.18)两边取对数：

$${}^2\mathcal{M}(\xi,\theta) = {}^1\mathcal{M}(\xi-d, \theta-\theta_0)\ , \tag{1.19}$$

其中，$\xi = \log\rho$、$d = \log z$。

(4) 通过傅里叶变换的平移性质，从公式(1.19)中估算出 z 和 θ_0 ，然后利用估算得到的参数对 ${}^2I$ 进行逆旋转、逆缩放，得到 ${}^2I'$ 使得

$$ {}^2I'(x,y) = {}^1I(x - x_0, y - y_0) \, . \tag{1.20} $$

相应地，

$$ {}^2\mathcal{F}'(\xi,\eta) = e^{-j2\pi(\xi x_0 + \eta y_0)} \, {}^1\mathcal{F}(\xi,\eta) \, . \tag{1.21} $$

那么 (x_0, y_0) 也可以利用傅里叶变换的平移性质估算出。

综上，所有的运动参数 (z, θ_0, x_0, y_0) 都可以通过相位相关法从公式(1.19)和(1.20)中得出。以公式(1.20)为例，首先计算公式左右两边信号的交叉功率谱，方法如下：

$$ Q = \frac{{}^1\mathcal{F}(\xi,\eta) \circ {}^2\mathcal{F}'^{*}(\xi,\eta)}{|{}^1\mathcal{F}(\xi,\eta) \circ {}^2\mathcal{F}'^{*}(\xi,\eta)|} \, , \tag{1.22} $$

其中，$\circ$ 是对应位置元素的乘积，* 表示复数的共轭。通过傅里叶逆变换，可以得到归一化的互相关结果：

$$ q = \mathcal{F}^{-1}\{Q\} \, , \tag{1.23} $$

在本文中又称作**相移图（*PSD*）**。该相移图 q 上能量峰值最大的位置与图中心的偏移就对应着两帧图像帧之间的偏移 (x_0, y_0)：

$$ (x_0, y_0) = \arg\max_{(x,y)} \{q\} \, . \tag{1.24} $$

在本文的实现过程中，相移图被离散化成单元格，最高峰的位置即在其中一个单元格内。需要注意的一点是由于图像的运动，两帧图像之间会有非重叠区域。非重叠区域没有对应关系，不会对最高峰的能量有贡献，而是会在相移图上生成噪声，这些噪声会分布在整张相移图上。因此当重叠区域足够大的时候，这些噪声不会影响最高峰的位置及其检测。

综上，经典的 FMT 描述了两帧图像之间的运动，对应采集图像时的 4DoF 相机运动，包括绕 z 轴的旋转（假设 z 轴垂直于成像平面）和 3 自由度的平移（缩放是沿 z 轴的平移造成的）。而当相机发生横滚或者俯仰运动时，FMT 无法进行图像之间的配准以及相机的运动估计。此外，由于该算法假设缩放参数 z 和平移 (x_0, y_0) 是常量且唯一，导致经典傅 FMT 的应用场景限制在单一深度下，即要求图像上所有的点到成像平面的距离相同。在多深度场景下由于透射投影的作用，两帧图像之间的运动通常包含多种缩放和平移。

1.4　傅里叶梅林变换的相关工作回顾

1.2.2节中提到了有一类的 VO 基于图像外观进行运动估计，用于图像配准的频域法就是这样的一类方法，它以整张图像作为输入，将其转换到频域进行处理。频域法的核心就是傅里叶变换，更准确来说是快速傅里叶变换 (Fast Fourier Transform, FFT)[92]。早期，计算机视觉领域用其来估计图像之间的平移 [93]；之后，通过加入梅林变换，频域法还可以用来估计图像的缩放和旋转，该方法被称为 FMT[91,94]。其后，一直有相关工作致力于进一步提高 FMT 的性能，例如[95,96]提出了改进的 FMT，使其计算速度更快、鲁棒性更高。[97]中详细总结了这些相关工作，并指出即使在一些比较有挑战的场景下 FMT 的效果仍然很好，包括特征较少的环境、能见度较低以及有动态物体的场景。无独有偶，文献[95]中通过对比 FMT 和 SIFT，发现在特定场景下 FMT 比 SIFT 计算更快、更准确；[98]也表明基于 FMT 的 VO 在特征不明显的环境下比其他基于特征（如 ORB、AKAZE）的 VO 方法表现得更加准确和鲁棒。FMT 的鲁棒性较高主要归功于它考虑了图像中从小的局部到整张图像的结构。

正因为 FMT 的鲁棒性和准确性很高，它已经被成功应用于多种不同的任务中，如图像配准[99–101]、指纹图像哈希[102]、视觉归航[103]、点云配准[104]、3D 建模[105]、遥感[106,107] 以及定位与建图[108,109]。但是，FMT 也有一些缺点：

- 它要求采集设备不能倾斜；
- 环境需要是平面的且与成像平面平行。

目前已有一些相关工作致力于解决第一个弊端，例如 Lucchese 基于仿射 FMT 分析估算出了两帧图像之间的仿射变换[110]；文献[111]利用过采样技术和基于 Dirichlet 的相位滤波器，使 FMT 对由采集设备造成的图像倾斜更加鲁棒；为了估算相机的倾斜角度，[98,112–114] 都采用了子图提取策略。但是对于第二点弊端，目前的相关研究还比较少，这也限制了 FMT 在 VO 中的应用。

从 VO 的角度来看，FMT 有很多劣势，因为它只能估计 4DoF 的运动：3DoF 的 2D 刚体运动加上 1DoF 的缩放。不过，在某些特定的应用中还是可以使用 FMT 来实现 VO 的[62,99,115–119]，比如装有向下相机的无人机、水下机器人等。[112]将图像划分成多个区域块，利用 FMT 计算两帧图像之间对应区域块的运动，从而实现了更加鲁棒的稠密光流法；[113]也利用类似的子图思想，通过 FMT 估计了相机的倾斜角；[5]利用 FMT 中相位相关的思想提出了基于全向相机的视觉

罗盘。不过这些方法都无法进行三维空间中的无约束运动估计。关于 FMT 中的一个完全三维扩展就是将其应用到点云上，[104]中介绍了基于 FMT 的一个新扩展 Fourier-Mellin-SOFT (FMS)，它可以用来计算两个三维点云之间的 7DoF 变换：6DoF 刚体变换加上 1DoF 的缩放。尽管该方法极大地扩展了 FMT 的应用范围，但是它还是无法基于单目相机估计其在三维空间中的运动。

1.5 近年研究热点与难点

随着近几年 VO/VSLAM 技术的进一步发展以及相关工作的开源，学术界和工业界对该技术的关注与日俱增，尤其是利用这些开源框架解决不同场景下的定位任务，例如无人机的室内外定位、服务机器人的定位与导航等。但是，由于实际应用场景与公开数据集有一定的区别，使得这些算法在实际应用中受到了不小的挑战。主要的研究热点与难点包括以下几个方面：

1) 随着制造工艺的发展，全向相机的成本进一步下降，在学术界和工业界受到持续关注。虽然不少算法都利用全向相机视野较大的优势来获得更加鲁棒的定位，但也正由于全向相机的视野很大，使得图像分辨率相同情况下其像素精度不高。另一方面，廉价的全向相机由于其制造工艺粗糙，无法进行精确的标定，现有的开源算法无法利用其采集到的图像获得较为准确的定位。

2) 在一些比较具备挑战性的场景中，如运动模糊、浑浊的水下场景、特征较少或重复特征较多的场景等，现有的特征法无法提取出有效的特征或进行较好的特征匹配，使得基于特征的 VO 无法正常工作。尽管直接法在特征较少的环境下比特征法鲁棒，但是在一些结构不明显的场景下也无法工作；尤其是直接法对于相机标定的要求较高，对于廉价相机的支持度不够好。

3) 虽然 FMT 在比较具备挑战性的场景中可以获得不错的结果，但是它只能估计 4DoF 的相机运动且要求整张图像上像素点的深度相同，极大限制了其在 VO 中的应用。

接下来，本文将针对这些难点，提出相应的解决方案。

1.6 本文内容与结构

考虑到 FMT 的鲁棒性及其在 VO 应用中的局限性，本文主要着重于扩展 FMT 在 VO 中的应用。全文总共有五章，第一章是绪论，介绍了本文所需的基

础知识并回顾了相关工作。接下来的四章安排如下：

• **第 2 章** 提出了利用 FMT 来实现全向图像的特征匹配。为了满足 FMT 的应用条件，本章提出递归划分子图策略从全向图像中提取子图，再利用 FMT 计算两帧图像之间对应子图对的运动，从而形成匹配的一致性点对。为了评估特征匹配的效果，本章利用全向相机模型和经典的五点法实现了一个简易 VO，比较了利用 FMT 的特征匹配和其他特征以及光流法进行位姿估计的效果。消融研究和与其他方法的对比试验表明，本章提出的利用递归子图策略和 FMT 的方法提供了较为鲁棒的特征匹配结果，尤其是在比较具有挑战性的场景下。

• **第 3 章** 利用第 2 章中基于 FMT 的全向图像特征匹配，提出了一种新颖的全向相机旋转估计方法。和以往研究中的基于几何法或直接法的 VO 不同，本章没有直接替换相机模型使其适应 VO/VSLAM 框架，而是从全向相机本身的特点出发，将其在三维空间中的运动建模成了正弦曲线，从而把全向相机的姿态估计问题转化为正弦曲线拟合问题。尽管本章并没有估计相机的平移，但是仍然在正弦曲线模型中包含了相机的平移项，并分析了其对于旋转估计的影响，同时为未来位姿估计的扩展提供了接口。最后，基于 FMT 和光流法估计得到的子图运动，实现了该全向相机旋转估计算法。通过与经典的几何法的对比，发现该正弦曲线拟合方法可以为全向相机的姿态估计提供更好的性能。

• **第 4 章** 扩展了经典的 FMT，使其可以应用于多深度场景。本章主要针对经典 FMT 需要单一深度的场景，提出了扩展傅里叶梅林变换 (extended Fourier-Mellin transform, eFMT)，既保留了 FMT 在具有挑战场景下比较鲁棒的优点，又打破了其单一深度的限制。该算法主要基于对 FMT 中相移图的观察，发现当场景由单一深度变为多深度时，相移图上的单个能量峰将变成多个高能量值，于是 eFMT 将原来的能量峰值检测扩展为检测能量总和最大的线，从而实现了多深度场景下的扩展。最终，在无人机采集的校园数据集上，该算法和主流 VO/VSLAM 框架的对比实验表明，该算法的准确度和 ORB-SLAM3 接近，但是在特征较少时 eFMT 更为鲁棒。

• **第 5 章** 总结了本文的主要工作，并在此基础上对未来研究方向进行了展望。

第 2 章　基于傅里叶梅林变换的全向特征匹配

本章主要介绍了利用 FMT 匹配两帧全向图像。不同于传统的特征匹配，本章首先将全向图像分成了若干子图，然后通过 FMT 计算出两帧子图之间的运动。为了评估该匹配算法，本章基于两帧之间的匹配结果，利用经典的五点法估计了相机的位姿，并与相应的真值进行了对比。同时，本章对传统的特征匹配也进行了相应的位姿估计操作，最终通过对比不同算法估计的相机位姿来评估本章提出的算法的鲁棒性和准确性。本章最后一部分的实验表明，基于 FMT 的位姿估计算法比传统的基于特征的位姿估计表现更为鲁棒，尤其是在能见度较差的情况下，如运动模糊、特征较少、大雾浓烟等环境中。

该基于 FMT 的位姿估计算法之所以能获得较好的表现，主要归功于本章所提出的递归划分子图策略。该策略以 FMT 中纯相位匹配滤波器的信噪比作为主要指标，来评判配准是否成功。一般而言，如果图像中只包含一种深度，则纯相位匹配滤波器得到的相移图上只有一个峰，表示两帧之间的位移仅有一种。如果相移图上出现了多个峰，则表示两帧图像之间的位移有多种可能性，即图像上存在不同的深度，该情况超出了 FMT 的应用范围，会导致两帧子图之间的配准结果不可信。此时，子图将会被进一步细分成更小的子图。该递归划分子图策略是本章所提方法中的重要一环。相比于文献[112] 中对傅里叶梅林变换的大量使用，本章中通过递归划分策略实现了极具竞争力的计算时间和较为鲁棒的算法性能。因此，该策略对有效实现基于子图的频域配准有着十分重要的意义，未来可以应用到 2.5D 的非平面地形拼接、非标准投影下的运动恢复结构（如成像声呐）等情况中。

本章将 FMT 应用于全向图像的位姿估计，主要贡献总结如下：

- 实现了 FMT 的 2D 配准到 3D 配准；
- 提出了基于子图的运动模型，来补偿全向图像中的非线性畸变；
- 进行了关于运动模糊、图像噪声和计算时间的鲁棒性消融实验；
- 提供了基于 FMT 的定位算法与基于特征的算法的基准比较。

2.1 算法框架

图 2.1简要地展示了基于 FMT 的全向图像位姿估计算法的流程图。首先，该算法将全向图像变换成全景图像，如图 1.3所示；其次，全景图被划分成了若干对共置的子图，即每对子图取自两个连续图像帧的同一窗口；然后，该算法利用 FMT 计算出每对子图之间的 2.5D 变换，即一致点对，从而获得两帧全景图像之间的稀疏运动流场；接着，该算法通过标定好的全向相机模型恢复出一致点对的归一化光线向量；最后，将这些一致的光线向量作为五点法的输入，估算出两帧之间的相对位姿。

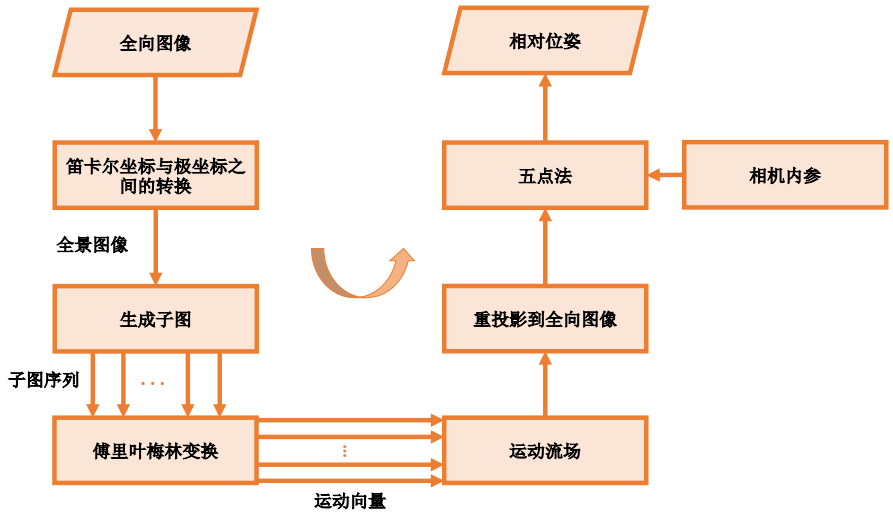


图 2.1 基于 FMT 的位姿估计流程图

Figure 2.1 Pipeline of the pose estimation based on FMT

需要指出的是，本章中我们使用“一致点对”一词来强调基于外观的一致性，用以区分视觉特征的匹配，即：

- 一致点对包含两个点 ${}^1p$ 和 ${}^2p$，分别代表从同一个窗口 w_i 得到的两个子图 ${}^1a_i$ 和 ${}^2a_i$，也就是说在连续的两帧中，一致点对的坐标集合相同；
- 一致点对 ${}^1p$ 和 ${}^2p$ 的相对位置独立于它们在子图 ${}^1a_i$ 和 ${}^2a_i$ 上的位置，但是与 FMT 计算出的子图 ${}^1a_i$ 和 ${}^2a_i$ 之间的 2.5D 变换有关；
- 一致点对 ${}^1p$ 和 ${}^2p$ 的相对位置还与相机模型有关，即与利用相机模型对视在运动的投影有关。

算法 1是该 3D 频域 VO 算法的伪代码。接下来的各小节将更加详细地描述各个关键步骤。关于全向图像与全景图像之间的相互转换已在1.1.1节中进行了描述，本章将不再赘述。

算法 1 基于全向图像的 3D 频域 VO

输入： 全向图像 ${}^1I_o, {}^2I_o$；噪声滤波器阈值 th_{pr}, th_{pnr}；子图大小阈值 th_e
利用笛卡尔坐标系到极坐标系的变换将全向图像 ${}^1I_o, {}^2I_o$ 转换为全景图像 ${}^1I_p$, ${}^2I_p$
从全景图像 ${}^1I_p$ 和 ${}^2I_p$ 提取子图集合 ${}^1\mathbb{A}, {}^2\mathbb{A}$
for all 子图 ${}^1a_i \in {}^1\mathbb{A}, {}^2a_i \in {}^2\mathbb{A}$ **do**
    计算相对视在运动 m_i = FMT$({}^1a_i, {}^2a_i, th_{pr}, th_{pnr}) = [s\ \theta\ t_x\ t_y]^\top$
    if $PR(m_i) < th_{PR}$ and $PNR(m_i) < th_{PNR}$ **then**
        选择点 ${}^1p_{a_i} = (c_x + \delta, c_y + \delta)$, where $\delta > 0$
        计算一致点对 $F_i = ({}^1p_{a_i}, {}^2p_{a_i})$ (Eq. (2.1))
        通过极坐标系到笛卡尔坐标系的变换将一致点对 F_i 变换到全向图片坐标系 *polar-to-Cartesian*(F_i)
        得到相机光线对 $({}^1P_i, {}^2P_i) = \pi^{-1}(F_i)$ (Eq. (1.2))
        将相机光线对 $({}^1P_i, {}^2P_i)$ 加入到一致点对集合 $\mathbb{S}$ 中
    else
        while ${}^1a_i$ 的尺寸小于阈值 th_e **do**
            将 ${}^1a_i, {}^2a_i$ 分成四个更小的正方形子图 ${}^1a_{i.j}, {}^2a_{i.j}$，更小的子图尺寸为原来子图大小的 $1/4$
            对更小的子图 ${}^1a_{i.j}, {}^2a_{i.j}$ 重复第 5 行到第 17 行的操作
        end while
    end if
end for
相机变换 T = 随机采样一致性 (五点法 ($\mathbb{S}$))
输出： T

2.2 利用傅里叶梅林变换估计一致点对

假设一全向相机在位姿 ${}^1C$ 和 ${}^2C$ 分别采集两张全向图像 ${}^1I_o$ 和 ${}^2I_o$，其对应的全景图像分别为 ${}^1I_p$ 和 ${}^2I_p$，则可以利用这两张全景图像估算出位姿 ${}^1C$ 和 ${}^2C$ 的相对变换 ${}^1_2T$。主要包含以下两个步骤：

1. 利用 FMT 配准（2.2.2节）生成由一致点对组成的运动流场（2.2.3节），并将其转换到标定后的视角下，即相机光线的方向向量（2.2.4节）；

2. 通过这些一致点对集合（归一化方向向量集合）估算出位姿之间的变换 ${}^1_2T$（2.4节）。

本节主要考虑第一步：利用 FMT 计算两帧全向图像之间的一致点对，其包含以下四个步骤：选取子图、估计子图间的视在运动、计算运动流场和归一化一致点对。

2.2.1 子图的选取

本节考虑的问题是如何从全景图像 ${}^1I_p$ 和 ${}^2I_p$ 中选取每一对子图 ${}^1a_i$ 和 ${}^2a_i$。一种比较详尽的方法就是文献[112] 的做法，利用 FMT 计算稠密的运动场，即每张子图都是原图中非常小的一块。该方法的缺点有二：一是在计算上非常占资源，二是子图的尺寸只有 32×32 像素，已经到达了 FMT 算法的临界值，文献[120] 中指出如果图像更小的话，FMT 配准的准确度和鲁棒性将无法保障。

关于子图的选取主要考虑两个方面。第一点是子图的大小 $N_a \times N_a$，它将直接影响到在能见度条件较差情况下算法的鲁棒性。本章将在2.5.2.2节通过实验来比较不同尺寸的子图对结果的影响，从而确定后续实验中合适的子图尺寸。结合该实验的结果，本章主要考虑尺寸较大的子图，以提高 FMT 的鲁棒性。从应用的角度出发，本章用全景图像 I_p 的高 ${}^yN_p$ 作为子图的边长，即 $N_a = {}^yN_p$。第二点是除了子图的大小以外，子图的数量也将影响算法的性能。如果子图的选取较为稠密，比如每隔一列都选取一张子图，那么算法的计算时间将会很长。因此，本章选择了一种稀疏的选取方法，即只用很少的子图，但是又足够保证位姿估计的结果准确。为了实现这一点，本章利用了递归的子图选取策略，策略的详细内容将在2.3中介绍。

利用上述子图的选取策略，两张全景图像将生成对应的两个子图集合 ${}^1\mathbb{A} = \{{}^1a_1, {}^1a_2, {}^1a_3, ..., {}^1a_m\}$ 和 ${}^2\mathbb{A} = \{{}^2a_1, {}^2a_2, {}^2a_3, ..., {}^2a_n\}$，然后利用 FMT 将每个 ${}^1a_i$ 和

其对应的 ${}^2a_i$ ($i \in \{1,...,m\}$) 配准。每一对子图 ${}^1a_i$、${}^2a_i$ 都是从全景图片中用同一窗口 w_i 选取的，也就是说子图在两帧连续的全景图像 ${}^1I_p$ 和 ${}^2I_p$ 上的坐标相同。下一小节将详细介绍这些窗口 w_i 是以何种固定的规律放置在全景图像 I_p 上的，即子图在全景图像上的位置。

2.2.2　子图间的视在运动估计

两张子图 ${}^1a_i$ 和 ${}^2a_i$ 之间的 2.5D 视在运动可以利用基于 FMT 的图像配准来估算。FMT 要求输入的图像 ${}^1a_i$ 和 ${}^2a_i$ 是方的，其大小记为 $N_a \times N_a$。但是一般用到的图像很可能是长方形的，此时有两种方法可以将其变成正方形的：从该图像上截取正方形的部分或者利用补零的方法来用到整张图片的内容。这两种方法各有缺点，前者可能会丢失掉部分信息，后者在补零过多时可能会导致图像包含的大部分信息为零，因此一般在将图像变为正方形的过程中需要尽量包含原来的信息。本章采取的是第一种策略：从全向图像上截取正方形。经典 FMT 的原理已在1.3节中进行了详细介绍，在此不再赘述。本章使用的 FMT 是由文献[62]中提出的改进 FMT，是经典 FMT 的一种变体，它通过有限冲击响应滤波器检测峰值，使运动估计结果更加准确。

2.2.3　运动流场估算

为了估计运动流场 $\mathbb{M}$，需要找到全景图像 ${}^1I_p$、${}^2I_p$ 之间的一致点对，也就是子图集合 ${}^1\mathbb{A}$、${}^2\mathbb{A}$ 之间的一致点对。为此，本章利用 2.5D 傅里叶梅林配准来计算对应的子图 ${}^1a_i$、${}^2a_i$ 之间的视在运动 m_i，具体方法已在2.2.2节中阐述。

假设四维运动向量 m_i 表示子图 ${}^1a_i$ 和 ${}^2a_i$ 经过 FMT 配准的结果，其包含子图 ${}^1a_i$ 和 ${}^2a_i$ 之间的缩放 s，旋转 θ 和平移 t_x, t_y，即 $m_i = [s\ \theta\ t_x\ t_y]^\top$，则可利用运动向量 m_i 来表示子图 ${}^1a_i$ 和 ${}^2a_i$ 之间的一致点对：${}^1a_i$ 上的 ${}^1p_{a_i} = (u'_1, v'_1)$ 和 ${}^2a_i$ 上的 ${}^2p_{a_j} = (u'_2, v'_2)$，公式如下：

$$\begin{bmatrix} u'_1 \\ v'_1 \end{bmatrix} = \begin{bmatrix} u'_2\alpha - v'_2\beta + c_x(1-\alpha) + c_y\beta + t_x \\ u'_2\beta + v'_2\alpha - c_x\beta + c_y(1-\alpha) + t_y \end{bmatrix}, \tag{2.1}$$

其中，$\alpha = s\cos\theta$；$\beta = s\sin\theta$；(c_x, c_y) 是子图窗口 w_i 的中心坐标，用来决定子图 ${}^1a_i$、${}^2a_i$ 的位置。基于公式(2.1)，子图 ${}^1a_j$、${}^2a_j$ 之间的所有一致点对都可以找到。理论上而言，子图 ${}^1a_j$ 上所有的点都可以在 ${}^2a_j$ 上找到其一致点，但是只要一对一致点对来表示子图的运动，因为子图 ${}^1a_j$、${}^2a_j$ 之间的所有一致点对对应的

视在运动 m_i 相同。因此，可任选子图 ${}^k a_i$ 上的一点 ${}^k p_{a_i} = (c_x + \delta, c_y + \delta), \delta > 0$ 来表示运动流。需要注意的是 δ 不能为 0，因为 δ 为 0 时，该一致点对不能表示旋转与缩放信息。图 2.2c给出了运动流场的一个示例，所有的一致性点对均由子图间的 FMT 配准估计得到。

(a) 第一帧全景图像

(b) 第二帧全景图

(c) 运动流场

图 2.2 两帧全景图像之间的运动流场示例

Figure 2.2 An example of a motion flow field between two panorama images

2.2.4 一致点对归一化

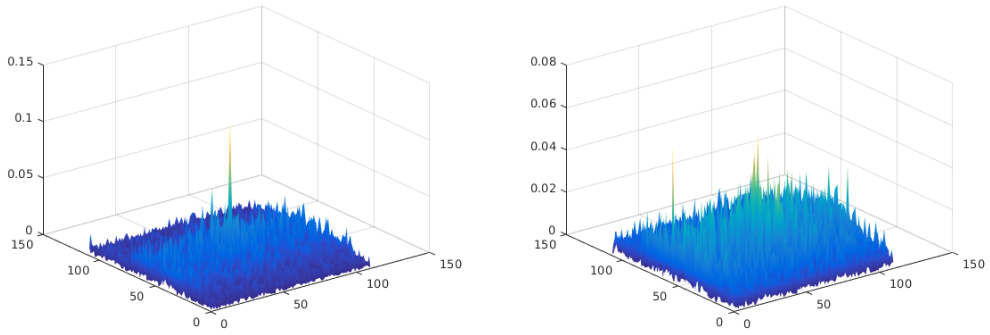
通过公式(2.1)计算得一致点对后，可利用五点法来计算相机位姿 ${}^1C$ 和 ${}^2C$ 之间的变换。但是五点法通常都要求归一化之后的一致点对，也就是需要找到一致点对对应的相机方向向量。1.1.1节中介绍的相机标定可以得到全向相机的模型，利用该模型可以得到全向图像 I_o 上每个像素 (u, v) 对应的去畸变、归一化之后的相机光线向量 $P = [u, v, f(u, v)]^T$，该向量方向以相机位姿 C 为坐标系。基于此，可计算出一致点对 ${}^1p_{a_i}$、${}^2p_{a_i}$ 相应的光线向量 ${}^1P$、${}^2P$。值得注意的是，一致点对 ${}^1p_{a_i}$、${}^2p_{a_i}$ 的坐标是其在全景图像上的坐标，需要先将其转换到全向图像上，才能计算出相应的相机光线向量。

2.3 递归划分子图策略

本节主要介绍如何判断当前子图是否符合 FMT 配准要求，如果不符合的话，应该如何进一步划分子图。

2.3.1　傅里叶梅林变换中的信噪比

利用 FMT 估计两帧图像之间的运动流的最后一步是通过相位相关法估计两帧之间的平移，通过相位相关法可以生成相移图。理想情况下，该相移图上只有一个狄拉克脉冲，即单个峰值，其与相移图中心的相对位置表示 2D 的平移参数。但是在实际应用中，由于传感器噪声、环境不是完美静止和平坦的以及没有完美的 2.5D 运动等原因，该狄拉克脉冲往往变得更平坦、更宽，并且参数空间中存在大量噪声。



(a) 高信噪比　　**(b) 低信噪比**

图 2.3 两种不同的傅里叶梅林配准结果下的相移示例图

Figure 2.3 Examples of phase shift diagrams in two different FMT registrations

图2.3展示了利用 FMT 计算平移过程中的两个相移图示例。图2.3b中的峰没有图2.3a中的清晰，且图2.3b中的峰噪声更大，即示例中两个相移图的信噪比大小不同。对 FMT 而言，信噪比的意义很大，可以用来判断配准成功与否。

本节考虑用两种不同的方法计算图像的信噪比。第一种方法是计算最高的峰值和第二高的峰值之间的比值 PR，即：

$$PR = \frac{E_{1stpeak}}{E_{2ndpeak}} \tag{2.2}$$

第二种方法是计算最高的峰值和它周围 $\pm 10 \times \pm 10$ 范围内的噪声能量总和的比值 PNR，即

$$PNR = \frac{E_{peak}}{\sum_{i,j} E_{noise}(i,j)} \tag{2.3}$$

其中 $i \in \{P_x - 10, P_x + 10\}$, $j \in \{P_y - 10, P_y + 10\}$，$peak = (P_x, P_y)$。当信噪比的值越大的时候，说明估计出来的运动向量结果越可靠。

配准成功或失败对应的信噪比完全不同，因此信噪比可以用来作为判断配准成功与否的标准。具体而言，信噪比较大时，配准成功；信噪比较小时，配准结果和正确结果相差甚远。最简单的方式就是用一个阈值来划分配准是否成功，上述两种方法可各对应一个阈值：PR 对应的阈值 th_{PR} 和 PNR 对应的阈值 th_{PNR}。2.5.2.1节中将通过实验来比较两种方法的效果以及阈值的选取。

2.3.2 递归划分子图

假设信噪比的结果显示子图 ${}^1a_i$ 和 ${}^2a_i$ 未配准成功，此时最简单的处理方式就是丢弃到该子图对的匹配结果。在后续的计算只使用由配准成功提供的一致点对集合 $\mathbb{S}$。但是低信噪比不一定表示完全没有办法配准该子图对，低信噪比有时是由于子图 ${}^1a_i$ 和 ${}^2a_i$ 之间有多种视在运动造成的，如图2.3b所示，多个峰表示配准的两帧子图之间有不同的运动向量，这与 FMT 配准的基本假设相违背。

因此，本章中提出一种可能的策略来提取这些不同的运动向量，而不是简单地丢弃掉子图 ${}^1a_i$ 和 ${}^2a_i$ 的信息。具体而言就是将用来划分原来子图的窗口 w_i 进一步细分，如算法1的第 6 行和第 13 到 17 行所述，划分成四等分。更准确来说，如果原来的子图窗口 w_i 大小为 $N \times N$，偏移量为 (x_o, y_o)，即：

$$\begin{aligned} w_i \equiv \{(x,y)\} \text{ with} \\ x_o \le x < x_o + N \wedge y_o \le y < y_o + N\,, \end{aligned} \tag{2.4}$$

则进一步细分的四个更小的窗口 $w_{i.j}$ 为

$$\begin{aligned} j \in \{1,...,4\} : w_{i.j} \equiv \{(x,y)\} \text{ with} \\ x_o + (j-1)\cdot\frac{N}{4} \le x < x_o + j\cdot\frac{N}{4} \\ y_o + (j-1)\cdot\frac{N}{4} \le y < y_o + j\cdot\frac{N}{4} \end{aligned} \tag{2.5}$$

那么从两张子图 ${}^1a_i$ 和 ${}^2a_i$ 上可以得到 2×4 张更小的子图，分别记为 ${}^1a_{i.j}$ 和 ${}^2a_{i.j}$，$j \in \{1,...,4\}$.

如果进一步划分之后，${}^1a_{i.j}$ 和 ${}^2a_{i.j}$ 的配准仍然没有成功，那么该策略很显然可以递归地应用到每一个细分窗口 $w_{i.j}$ 上，即变成更小的四个子窗口 $w_{i.j.k}$ ($k \in \{1,...,4\}$)，然后继续尝试配准。

该递归算法的限制因素主要有两点。一是在利用子图的 2.5D 配准结果计算两帧相机位姿的三维变换 T 时，算法（五点法）本身只需要较少的点即能估算

出变换 T。虽然更多的点对会使 RANSAC 算法得到更多的内点（2.4节），但同时也会增加五点法的计算量。因此，在非必要的情况下，不执行该递归算法。第二点是为了保证 FMT 的鲁棒性，输入图像的尺寸不能太小。文献[120]指出，为了兼顾算法的准确度，傅里叶梅林变换的输入图像尺寸最好不小于 64×64 像素，这样才可以保证其在低能见度以及动态环境下的鲁棒性。因此，在递归划分子图时，如果子图的尺寸小于 64×64 像素，则应停止递归，若此时还是配准失败，则舍弃当前子图对。在本章的算法实现中，这两点可以由划分子图的窗口大小阈值 th_e 和噪声滤波器的阈值 th_{pnr}、th_{pr} 来确定。

2.4　从一致点对估计三维运动

至此，假设全向相机在位姿 ${}^1C$ 和 ${}^2C$ 处分别采集一帧全向图像 ${}^1I_o$ 和 ${}^2I_o$，则可以计算出一致点对集合及其对应的一致相机光线方向集合 $\mathbb{S}$。接着，本章提出的 3D 频域 VO 可以进行最后一步的估算（算法1的第 19 行），即利用五点法[30,121] 估计相机位姿 ${}^1C$ 和 ${}^2C$ 之间的六维刚体变换 T。具体来说，一致点对对应的相机光线向量 $({}^1P_i, {}^2P_i)$ 之间的关系可以表示为

$$ {}^1P_i^{\,T}\, E\, {}^2P_i = 0\,, \tag{2.6} $$

其中，矩阵 E 被称作本征矩阵，其主要基于文献[122] 中介绍的对极几何。两帧之间的旋转和平移可以从该本征矩阵 E 中提取。在本工作的实现中，该 3D 频域 VO 算法利用了 OpenGV [123] 工具包中提供的 Stewenius 五点法，其实现主要基于文献[124,125]。

为了更进一步提高算法的鲁棒性，该实现将五点法和 RANSAC[56] 框架结合起来，即从一致相机光线方向集合 $\mathbb{S}$ 中随机选择点对，然后用五点法估算出相应的变换，最终选择内点最多的变换 T 即为位姿 ${}^1C$ 和 ${}^2C$ 之间的变换（算法1的第 19 行）。

2.5　实验与结果分析

本节将评估 3D 频域 VO 算法的主要参数对于算法性能的影响，包括信噪比阈值、子图的选取策略以及递归细分策略。本节得出合适的参数，将用于2.5.3节中与其他算法的比较，2.5.3节中的主要评估指标包括图像匹配的几何误差、旋转

误差和平移误差。同时，算法在低能见度情况下的表现也是2.5.3节的评估内容之一。

2.5.1 实验中用到的数据集

在本章接下来的实验中，相关评估主要在四个数据集上进行。图2.5展示了每个数据集的一帧图像示例。前两个数据集是由本文作者采集的，后两个数据集取自现有的公开数据集。

图2.4a展示了采集前两个数据集所用到的搜救机器人，它是一个低成本的搜救系统，其主要传感器是智能手机上的相机，为了扩大其视野，在该相机上覆盖了一个低成本的 360° 全向镜头，用来采集全向图像。采集场景主要是室内环境和室外草坪环境，对应的数据集分别是 *office* 和 *lawn* 数据集，每个数据集中的图像大约是 500 帧。

第三个数据集是文献[3,4] 中提供的 *CVLIBS* 数据集，其包含 12607 张全景图像，主要场景为城市场景，由一个汽车平台采集。本实验中用到其中的 200 帧图像作为测试，用来评估算法的旋转准确度。实验中用到的第四个数据集是文献[5] 提供的 *OVMIS* 数据集，其采集平台为一移动机器人，采集传感器为全向相机，采集场景是室外草坪。和作者采集的室外草坪数据集不同的是，*OVMIS* 数据集用到的全向相机设备更为先进，成本更高，采集到的图像质量更好。

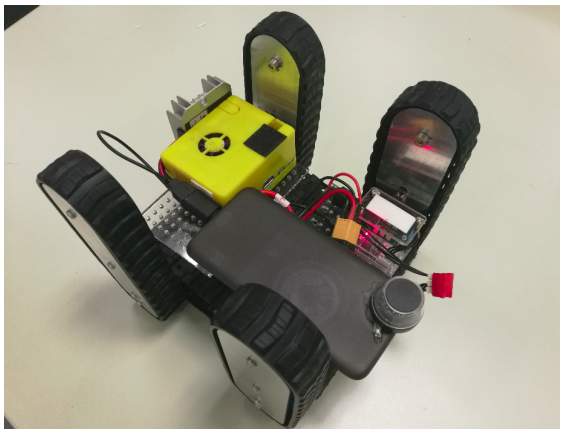
对所有的数据集而言，其真值由相应采集平台的惯性导航系统的测量值提供。

2.5.2 消融研究

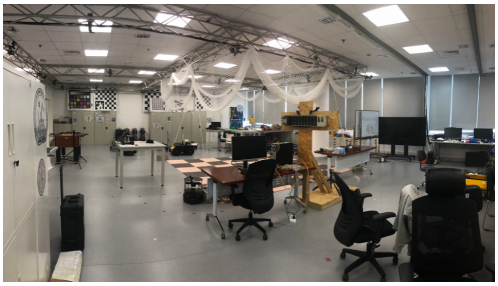
2.5.2.1 信噪比阈值对于算法性能的影响

正如2.3.1节中所讨论的，FMT 配准最后一步相位相关中的信噪比可用来指示配准是否成功。更进一步地，有两种不同的方法来表征信噪比：最高峰值与第二高峰值的比值 PR、最高峰值与其周围 $\pm 10 \times \pm 10$ 像素范围内噪声能量总和的比值 PNR。这两种方法对应的阈值分别设为 th_{pr} 和 th_{pnr}，即高于该阈值可认为配准成功，反之则为配准失败。

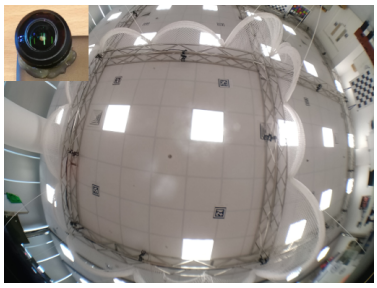
为了更好地解释这两种方法的影响，本节接下来主要评估其对于位姿估计最后一步的影响，即在 RANSAC 的框架下实现五点法时，该阈值对 RANSAC 内点以及最终位姿估计误差的影响（2.4节）。具体来说，当 PR 或 PNR 的结果显



(a) 低成本搜救机器人



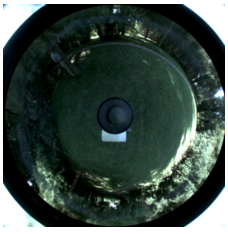
(b) 示例环境



(c) 鱼眼镜头采集的图像　　**(d) 360° 镜头采集的图像**

图 2.4 实验采集装置及图像示例

Figure 2.4 Image capture device and example images

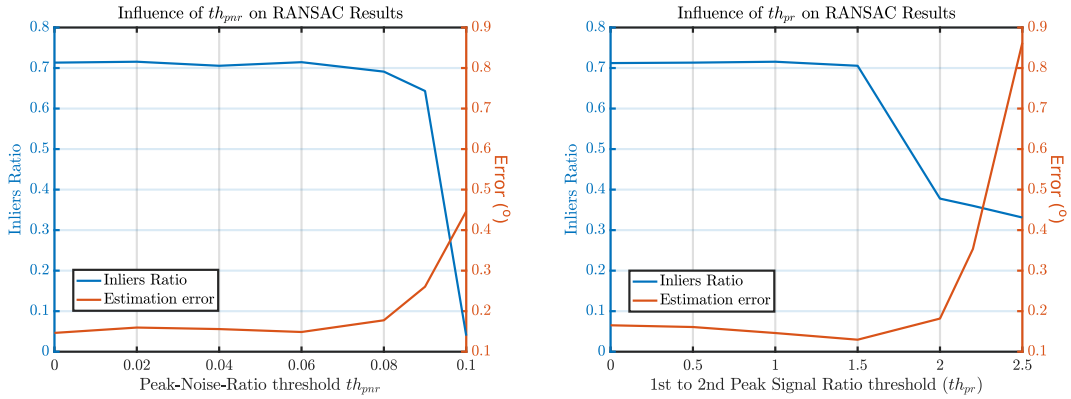


(a) *office*　　**(b)** *lawn*　　**(c)** ***CVLIBS***　　**(d)** ***OVMIS***

图 2.5 四个数据集中的图像示例

Figure 2.5 Example images from four datasets

示当前配准不成功时，当前子图对将无法生成相应的一致点对。当信噪比阈值较大时，匹配成功的点对数较少，RANSAC 内点的比例将下降，同时位姿估计算法会因为内点数太少而无法给出准确的位姿。



(a) *PNR* 阈值 th_{pnr} 对 RANSAC 结果的影响　　(b) *PR* 阈值 th_{pr} 对 RANSAC 结果的影响

图 2.6 两种信噪比阈值对位姿估计的影响

Figure 2.6 Influence of two signal-to-noise ratios on pose estimation

图2.6中的两幅图分别展示了 *PNR* 和 *PR* 的阈值 th_{pnr} 和 th_{pr} 的选取对内点比例和估计误差的影响。图中所示结果是在 *office* 数据集中 10 帧图像上测试所得。该 10 帧图像为相机发生俯仰角变化的结果，每连续的两帧图像之间的相机俯仰角为 5°。同时，为了减少偶然性，每个测试都重复进行了 10 次，最终图上显示的结果为 10 次测试的平均值。

图2.6所示结果有两点值得关注。一是 *PNR* 和 *PR* 这两种方法对算法的影响十分相似；二是两种方法都存在一段范围，阈值在该范围内时均可以获得较高的 RANSAC 内点比例和较低的估计误差，也就是说只有当阈值设置得非常大的时候，会导致一致点对集合 $\mathbb{S}$ 中的点对很少，可能会引发没有足够的点进行位姿估计。

基于以上观察结果，接下来的实验将仅使用其中一种方法，本文中选择 *PNR* 方法，相应的阈值 th_{pnr} 设置为 0.06。

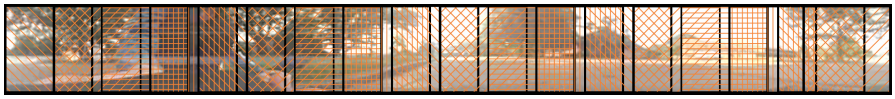
2.5.2.2　子图选取策略评估

2.2.1节讨论了如何将全景图像 ${}^1I_p$ 和 ${}^2I_p$ 划分为 n 张 $N_a \times N_a$ 大小的子图 ${}^1a_i$ 和 ${}^2a_i$ $(i \in \{1, ..., n\})$，其中每对子图 ${}^1a_i$、${}^2a_i$ 时从原全景图像中的同一窗口 w_i 获得的。这些子图窗口 w_i 以某种规律分布在全景图像上，从而生成子图。根

据2.3.2节中提到的递归划分策略，这些子图窗口也可能会被进一步地细分。

对于子图窗口 w_i 窗口而言，有两个参数和其分布有关，分别是窗口的大小 N_a 以及步长 s_a，其中步长 s_a 决定了下一个子图窗口 w_{i+1} 与当前窗口 w_i 在 x 方向和 y 方向上的相对偏移。根据窗口的大小和步长的选择，窗口的划分策略可分为重叠划分和非重叠划分两种。例如，图2.7a展现了在全景图像 1100×110 的非重叠划分，其中窗口大小为 $N_a=110$，步长为 $s_a=110$；图2.7b 中的示例为重叠划分，其中窗口大小为 $N_a=110$，步长为 $s_a=60$。

(a) 非重叠示例（$N_a=110$，$s_a=110$）



(b) 重叠示例（$N_a=110$，$s_a=60$）

图 2.7 通过重叠和非重叠窗口生成子图集合

Figure 2.7 Non-overlapping and overlapping windows to generate sub-images

本节接下来将通过对比实验，来解释窗口大小 N_a 与步长 s_a 对整个 3D 频域 VO 算法的计算时间与位姿估计误差的影响。关于递归划分子图策略的影响，将在2.5.2.3节中着重测试与评估。

在本实验中，主要用到的数据集是 *office* 和 *lawn* 数据集，每个数据集中都包含横滚、俯仰和偏航运动，其中每连续两帧间的角度分别为 2°，5° 和 10°（见表2.1）。这两个数据集中包含的图像为全向图像，其大小为 1280×720，转换成全景图像后，其尺寸为 1100×110。然后根据设置的窗口大小 110×110，64×64 和 32×32 以及不同步长，在竖直和水平方向上划分子图。本实验测试了每种窗口大小和步长的不同组合，相应的计算时间和位姿估计误差记录在表2.1中。

需要注意的是，在本实验中一致点对集合 $\mathbb{S}$ 中点对的最大个数 $\#_{\max}\mathbb{S}$ 由窗口的步长决定。给定全景图像的大小 ${}^{y}N\times{}^{x}N$ 和窗口步长 s_a，则窗口个数为 $n=\lceil\frac{{}^{y}N}{s_a}\rceil\cdot\lceil\frac{{}^{x}N}{s_a}\rceil$，需要进行 n 次子图 ${}^{1}a_i$ 和 ${}^{2}a_i$ 之间的 FMT 配准。在该划分算法中，全景图像的高度和长度不一定是步长的倍数，即 $\frac{{}^{y}N}{s_a}$ 和 $\frac{{}^{x}N}{s_a}$ 不一定要是整数，在这种情况下，在划分子图快到边缘时，只需简单地将右/下角的边对齐即可。在本对比实验中，一致点对的最大个数 $\#_{\max}\mathbb{S}$ 跨度从最稀疏的 10 到最稠密的 315，

分别对应窗口大小为 $N_a = 110$，步长为 $s_a = 110$ 以及窗口大小为 $N_a = 32$，步长为 $s_a = 10$。

理论上而言，窗口的大小 N_a 将影响 FMT 配准的鲁棒性及其计算时间。窗口大小的选择将影响子图之间的重叠区域，进而影响 FMT 的鲁棒性。当窗口大小 N_a 越大时，FMT 配准越鲁棒，相应的计算时间也会增加。

基于表格2.1中的结果，可以得到以下两点结论：

（1）当连续两帧之间的旋转较大时，子图的大小 N_a 和位姿估计的准确度之间有一定的关联。例如，当两帧之间的旋转大于 5° 时，如果子图的尺寸是 32×32，那么估计误差较大，如偏航角的误差接近 5°，相比之下，子图尺寸为 64×64 时则可以较为准确地估计该偏航角。同样地，当偏航角更大，如 10° 时，64×64 尺寸的子图也没有办法准确估计该偏航角，而 110×110 尺寸的子图则可以。

（2）在给定子图大小 N_a 时，子图之间的重叠大小，即步长 s_a 对算法准确度的影响较小。表2.1显示，当算法可以有效估计位姿时，误差都比较小。对固定子图大小 N_a 而言，不同步长情况下，位姿估计的平均误差相差大约在 $\approx 0.01°$。也就是说误差基本是随机分布的，更小的步长并不一定能得到更小的位姿误差。但是步长 s_a 对计算时间有着重要的影响，即步长越大，子图数越少，所需的计算时间越少。因此，用较为稀疏的一致点对集合既可以加快计算速度，又能获得较为准确的估计结果。比如在该实验中，步长 $s_a = 110$ 可以至多得到 10 组一致点对，当横滚、俯仰和偏航角的瞬时变化达到 10° 时，也能得到较为准确的位姿估计结果。总体来说，在该实验中，只要给定一个足够大的子图大小 N_a，位姿估计误差就可以小至 0.1°。

相机的平移和旋转都会影响图像之间的重叠区域，重叠区域的大小将直接影响 FMT 的鲁棒性。因此，无论相机发生平移或旋转，需要选择合适的窗口大小来保证子图之间有足够大的重叠区域。

2.5.2.3 递归子图划分策略的评估

如2.3.2节中所述，递归子图划分策略用在本章的所有实验中。按照2.5.2.1中得出的结论，递归子图划分策略使用的信噪比阈值为 $th_{pnr} = 0.06$，并且递归终止条件中的最小子图尺寸为 32。本节通过进一步的实验来评估了该策略的优势。通常情况下，当图像中包含不同深度的物体或者动态物体时，往往 FMT 配准的

(a) 相机旋转间隔：2°

大小	步长	偏航			俯仰			横滚		
		$\epsilon(°)$	σ^2	$t(s)$	$\epsilon(°)$	σ^2	$t(s)$	$\epsilon(°)$	σ^2	$t(s)$
110	110	0.046	0.0000	**0.110**	**0.010**	0.0015	0.122	0.083	0.0006	0.112
	60	0.048	0.0000	0.174	0.089	0.0000	0.180	0.226	0.0289	0.190
	30	0.050	0.0000	0.298	0.079	0.0002	0.313	0.830	0.1156	0.301
	20	0.043	0.0000	0.434	0.126	0.0056	0.524	0.378	0.0556	0.439
	10	0.044	0.0000	0.842	0.191	0.0225	0.960	0.372	0.0752	0.822
64	64	0.029	0.0000	0.114	0.057	0.0003	0.153	**0.044**	0.0002	**0.105**
	30	0.035	0.0000	0.269	0.043	0.0002	0.382	0.057	0.0001	0.255
	20	**0.028**	0.0000	0.486	0.044	0.0000	0.716	0.051	0.0003	0.472
	10	0.039	0.0000	1.338	0.039	0.0001	1.344	**0.044**	0.0002	1.300
32	32	0.102	0.0012	0.119	0.063	0.0005	**0.121**	0.051	0.0003	0.115
	20	0.105	0.0004	0.196	0.079	0.0006	0.191	0.076	0.0006	0.195
	10	0.094	0.0007	0.588	0.058	0.0004	0.564	0.072	0.0011	0.562

(b) 相机旋转间隔：5°

大小	步长	偏航			俯仰			横滚		
		$\epsilon(°)$	σ^2	$t(s)$	$\epsilon(°)$	σ^2	$t(s)$	$\epsilon(°)$	σ^2	$t(s)$
110	110	0.029	0.0000	**0.116**	0.149	0.0003	**0.118**	0.260	0.2138	**0.117**
	60	0.031	0.0000	0.203	0.312	0.0736	0.187	0.213	0.1386	0.185
	30	**0.027**	0.0000	0.345	0.179	0.0226	0.336	0.105	0.0017	0.332
	20	0.028	0.0000	0.469	0.560	0.4221	0.479	0.153	0.0038	0.489
	10	**0.027**	0.0000	0.901	0.146	0.0088	0.907	0.136	0.0011	0.867
64	64	0.076	0.0004	0.129	0.096	0.0011	0.130	0.157	0.0010	0.124
	30	0.052	0.0001	0.288	**0.077**	0.0003	0.303	0.098	0.0008	0.282
	20	0.051	0.0001	0.526	0.106	0.0006	0.549	0.118	0.0005	0.521
	10	0.055	0.0002	1.430	0.078	0.0003	1.393	**0.093**	0.0001	1.385
32	32	5.032	0.0828	0.129	0.138	0.0010	0.148	0.151	0.0015	0.130
	20	4.923	0.0912	0.205	0.112	0.0004	0.210	0.121	0.0009	0.201
	10	4.935	0.0272	0.598	0.112	0.0008	0.761	0.131	0.0011	0.630

(c) 相机旋转间隔：10°

大小	步长	偏航			俯仰			横滚		
		$\epsilon(°)$	σ^2	$t(s)$	$\epsilon(°)$	σ^2	$t(s)$	$\epsilon(°)$	σ^2	$t(s)$
110	110	0.095	0.0001	0.110	**0.169**	0.0034	**0.099**	0.542	0.0704	0.116
	60	0.068	0.0005	0.190	0.531	0.0120	0.153	0.421	0.0080	0.161
	30	0.084	0.0003	0.325	0.709	0.2078	0.288	0.311	0.0122	0.289
	20	0.069	0.0005	0.437	0.611	0.9730	0.362	0.551	0.6600	0.410
	10	**0.052**	0.0001	0.793	0.481	0.1668	0.760	**0.300**	0.0455	0.754
64	64	7.960	4.5076	**0.106**	0.273	0.0453	0.129	0.363	0.0842	0.113
	30	9.318	7.3541	0.241	0.518	0.1076	0.252	0.473	0.0797	0.251
	20	7.355	3.5246	0.414	0.358	0.0356	0.465	0.644	0.1699	0.445
	10	5.410	8.6105	1.127	0.268	0.0415	1.238	0.590	0.1415	1.187
32	32	9.997	0.0711	0.113	8.765	4.0711	0.114	7.593	3.2097	**0.111**
	20	10.089	0.1219	0.171	8.350	3.6449	0.175	8.673	2.7465	0.174
	10	10.077	0.0731	0.488	7.875	1.8814	0.527	6.063	2.6723	0.508

表 2.1 不同子图窗口参数设置下的旋转估计误差 ϵ、方差 σ^2 以及计算时间 t

Table 2.1 Rotation estimation error ϵ, covariance σ^2 and run-time t with different settings

相移图噪声较大，此时该递归划分策略将会被触发。图2.8中展示了一个包含动态物体（手）和不同距离物体（墙和桌子）的子图对。当利用傅里叶梅林变换配准该对子图时，递归划分子图策略就会被触发。

图 2.8 递归划分策略被触发的子图对示例

Figure 2.8 An example of the sub-images pair where the sub-division strategy is triggered

接着，为了评估该递归策略的性能，本实验在 *office* 的俯仰数据集上对比了使用和不使用该策略时的位姿估计误差，初始设置时子图的尺寸为 $N_a = 110$，步长为 $s_a = 110$。图2.9表明该策略可以略微地改善所提出的 3D 频域 VO 算法的性能。性能提升不是很高的原因之一是在这个数据集上，基于傅里叶梅林配准的结果，该递归划分策略被触发的次数约为总配准次数的五分之一。

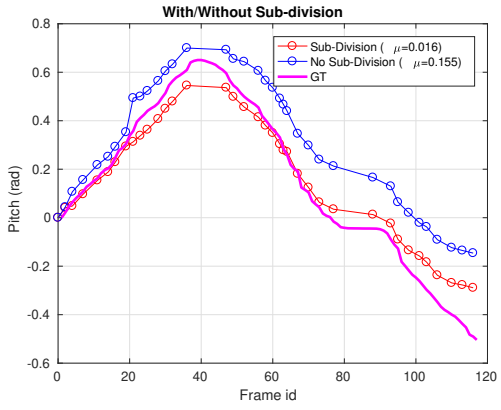


图 2.9 使用/不使用该递归划分策略对 3D 频域 VO 算法性能的影响

Figure 2.9 Performance on the algorithm with/without the sub-division strategy

2.5.3 与其他方法的对比实验

本实验中比较了本章所提出的 3D 频域 VO 算法与其他方法。由于本章主要考虑的是提升算法在全向图像上特征配准的性能，因此本实验与两种具备代表

性的特征点法进行了比较。另一方面，本实验还与另一种常见的算法——光流法进行比较。在特征匹配后，参与比较的算法都将和本章所提出的 3D 频域 VO 算法最后一步一样，即算法1的第 19 行：通过 RANSAC 框架下的五点法估算相对位姿。为了保证算法的公平性，其实现均由 OpenGV[1]工具包中的 Stewenius 五点法实现。

关于比较方法中所用到的特征点，本实验中选择 ORB[25] 和 AKAZE[26,27] 作为特征点法的代表。这两个特征点各有优势，文献[126] 中指出 ORB 的计算效率非常高，而 AKAZE 则是专门为非线性变换所设计的特征点，因此它在畸变较大的图像上表现更为鲁棒。为了不失公平性，基于特征点 ORB 和 AKAZE 的 VO 和基于光流法的 VO 算法，其特征提取与匹配部分将均由 OpenCV[2]中的相关函数实现。

另一方面，为了更进一步地评估本章的 3D 频域 VO 算法的性能，本实验中还将与开源的半直接法（SVO2）[88] 比较。SVO2 是现有最先进的 VO 算法之一，并且其很好地支持了全向相机。

2.5.3.1　算法的重投影/几何误差及其限制

给定三帧全向图像 ${}^1I_o$、${}^2I_o$、${}^3I_o$，其对应的特征点几何分别记为 ${}^1\mathbb{K} = \{{}^1k_i\}$，${}^2\mathbb{K} = \{{}^2k_i\}$, ${}^3\mathbb{K} = \{{}^3k_i\}$，则其重投影误差可定义为：

$$\epsilon_R = \frac{1}{L}\sum_{i\in L} \| {}^3k_i' - {}^3k_i \| , \tag{2.7}$$

其中，L 是集合 ${}^3\mathbb{K}$ 的长度，${}^3k_i'$ 是全向图像 ${}^3I_o$ 上的重投影点，其计算公式为

$${}^3k_i' = K {}^1_3T P_i \tag{2.8a}$$

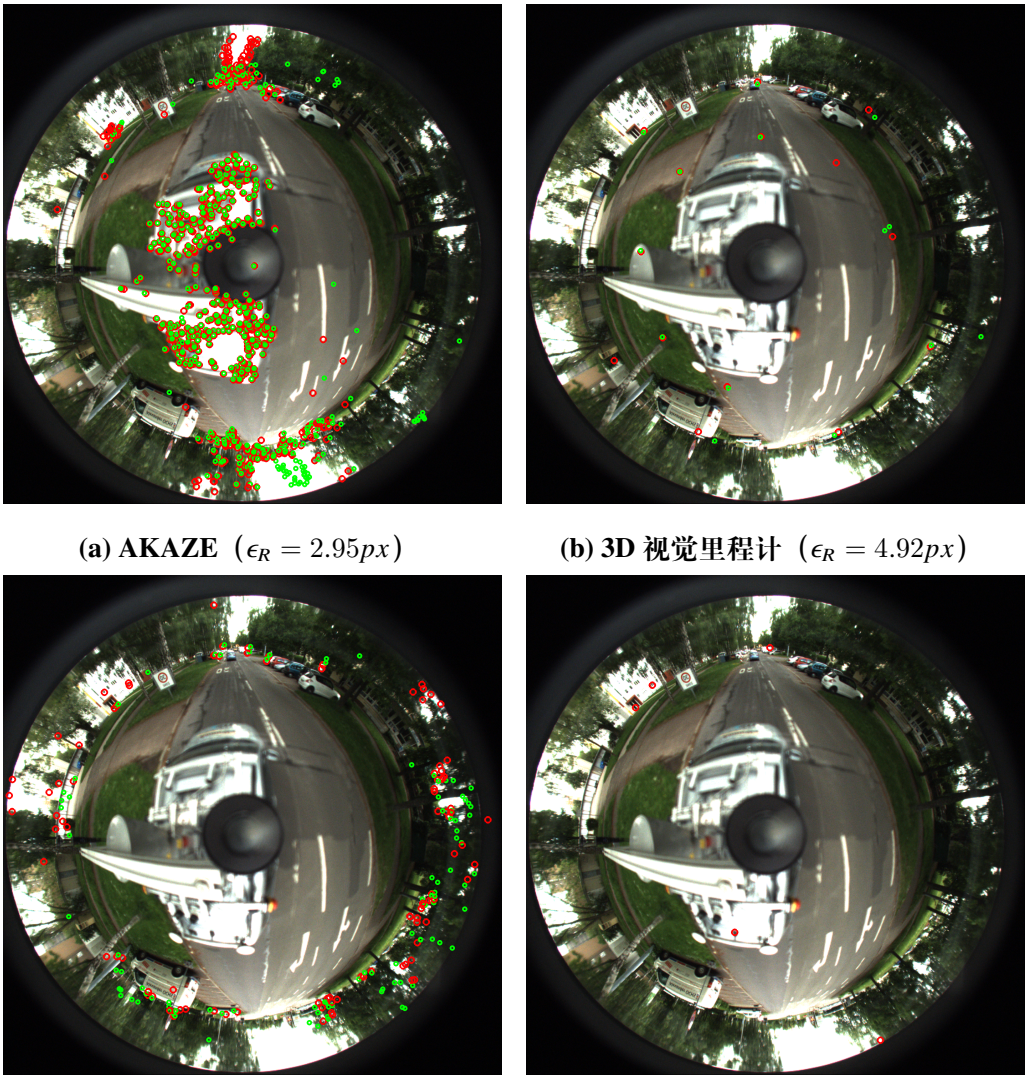
$$P_i = Triangulate({}^1k_i, {}^2k_i, {}^1_2T) \tag{2.8b}$$

此处的 ${}^1_2T$ 和 ${}^1_3T$ 由算法计算出的相对位姿变换给出。

图2.10展示了四种不同方法在 *CVLIBS* 数据集上的重投影误差的一帧示例，其中原本的特征点用红色圆圈表示，重投影后得到的点用绿色圆圈表示。从图2.10中可看出，基于 AKAZE 的方法重投影误差最小，重投影误差为 $\epsilon_R = 2.95px$，其次是本章提出的 3D 频域 VO，重投影误差大约是 $\epsilon_R = 4.92px$，然后

[1] https://laurentkneip.github.io/opengv/

[2] https://docs.opencv.org/



(a) AKAZE（$\epsilon_R = 2.95px$） **(b) 3D 视觉里程计（$\epsilon_R = 4.92px$）**

(c) 光流法（$\epsilon_R = 10.28px$） **(d) ORB（$\epsilon_R = \infty px$）**

图 2.10 在 *CVLIBS* 数据集[3,4] 上的重投影误差示例

Figure 2.10 Illustrative examples of the reprojection error in the CVLIBS dataset[3,4]

是光流法，其重投影误差为 $\epsilon_R = 10.28px$。而基于 ORB 的方法在该示例中重投影失败。

值得重视的一点是，这个利用重投影误差进行评估的方法有一个重要的缺陷。基于 AKAZE 的方法得到的大部分特征点都集中在采集平台即车顶上（见图2.10a），在车移动的过程中，由于车顶和相机的相对位置是固定的，因此这些特征点一直在采集到的全景图像的同一位置。这种情况下得到的重投影误差虽然较小，但是这些特征点对于估计帧与帧之间的变换没有贡献，这些特征点的存在反而会导致位姿估计的结果没有运动。如果不考虑这些汽车上的特征点，AKAZE 在该算法中的重投影误差约为 $8.64px$，此时 3D 频域 VO 表现最好。虽然在利用 AKAZE 计算视觉里程计时，可以将图像中的这一部分去除，但是在有动态物体的场景中，还是很有可能出现这种虽然重投影误差很小，但是位姿变换的估计误差很大的情况。

因此，在后续的实验中，与真值之间的绝对误差作为算法性能评估的评价标准，至少对于旋转估计而言，相应的惯性测量可以提供较好的比较基准。至于平移的估计，在后续实验中设计较少，但也会在部分提供平移真值的数据集上进行评估。

2.5.3.2　不同方法在不同数据集上的对比

如2.5.2.2节得出的规律，本章提出的 3D 频域 VO 算法在子图尺寸较大、分布较为稀疏时性能较好。对于作者自己采集的数据集，其全向图像的分辨率为 1280×720，相应的全景图像分辨率为 1100×110，该实验中的子图大小和步长都设为 110 像素。*CVLIBS* 数据集中的全向图像尺寸为 1400×1400，对应的全景图像尺寸为 2670×450，相应的子图大小和步长都设为 $N_a = s_a = 256$。*OVMIS* 数据集中的全向图像尺寸为 620×620，对应的全景图像尺寸大小为 1100×250，相应的子图大小和步长分别设为 $N_a = 250, s_a = 50$，此处 N_a 与 s_a 的取值不同主要是出于五点法的算法要求考虑。

在对比实验中，本节首先评估了不同方法在三个数据集 *office*、*lawn* 和 *CVLIBS* 上的旋转估计结果。图2.11展示了在 *office* 数据集上的旋转估计结果，可以看到本章中提出的 3D 视觉里程计算法和基于光流的 VO 最接近真值，其次是基于 AKAZE 的方法。在此数据集上，相机视野中有着各种不同物体，如书、架子、椅子等，因而特征较为丰富。

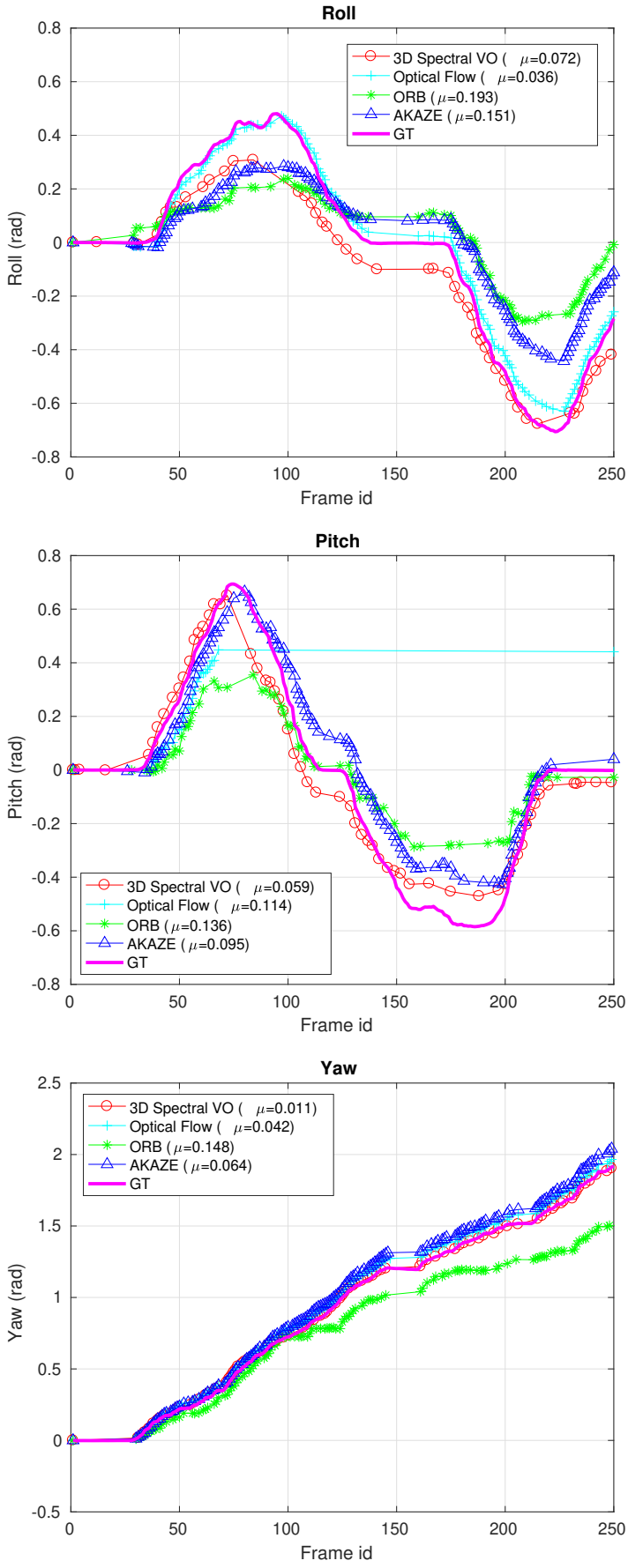


图 2.11 在 *office* 数据集上的旋转估计及其误差 μ

Figure 2.11 Rotation estimation on the *office* dataset and average error μ

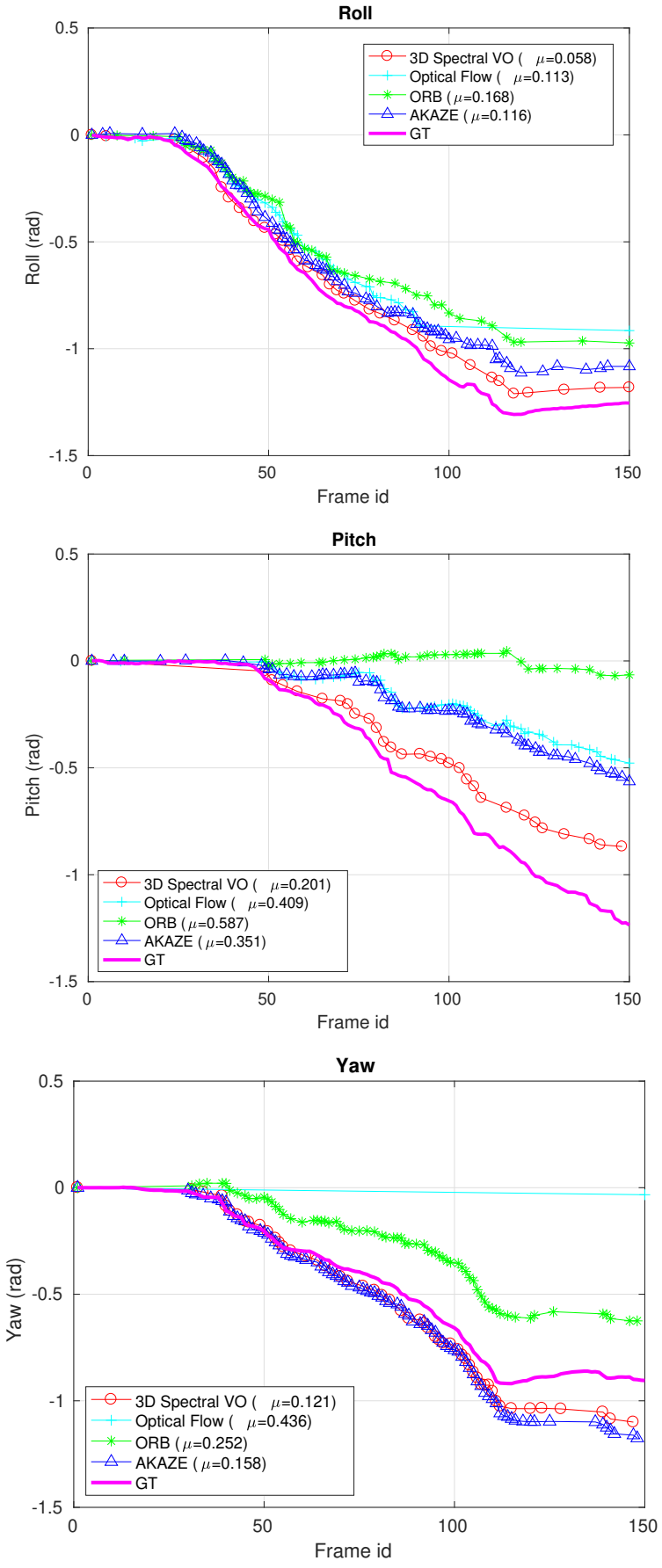


图 2.12 在 *lawn* 数据集上的旋转估计及其误差 μ

Figure 2.12 Rotation estimation on the *lawn* dataset and average error μ

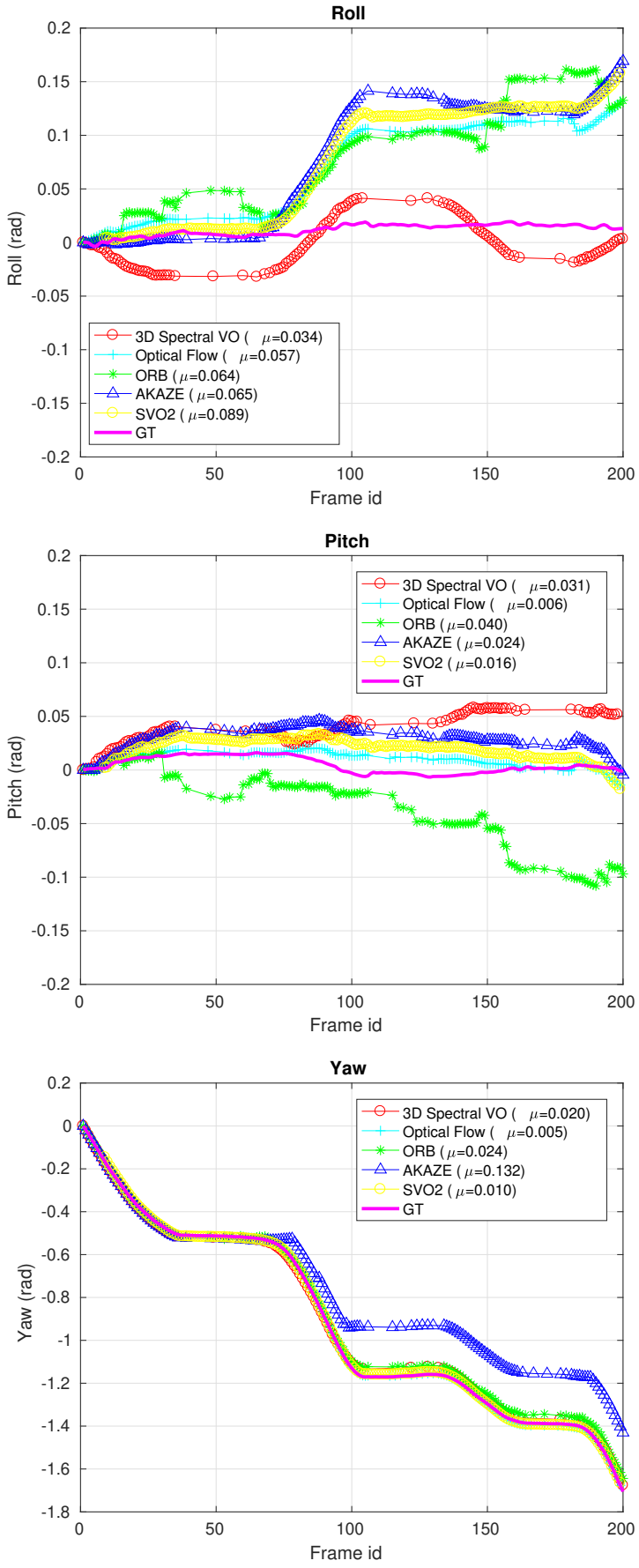


图 2.13 在 *MPI-omni* 数据集[3,4] 上的旋转估计及其误差 μ

Figure 2.13 Rotation estimation on the *MPI-omni* dataset[3,4] and average error μ

然而在类似 *lawn* 数据集的场景中，相机视野中大都为草（见图2.12），图像上的特征非常的类似，故特征法较难正确匹配。在 *lawn* 数据集上，AKAZE 在每帧图像上大约能提取 520 个特征，通过特征匹配每两帧间约有 365 对特征，在最后的五点法估计位姿环节大约有 190 对匹配能够被选为内点；ORB 在每帧图像上大约能提取 500 个特征，通过特征匹配每两帧间约能匹配 365 对特征，在最后的五点法估计位姿环节大约有 95 对能够被选为内点；同样地，光流法的平均特征点个数为 23，只有 17 对匹配上的特征点对。从以上数据来看，该数据集上特征过于类似，对这些方法提出了较大的挑战。图2.12展示了这些不同方法在 *lawn* 数据集上的表现，其中基于 ORB 的视觉里程计没有能够正确估计相机的俯仰（pitch）运动，光流法未能正确估计相机的偏航（yaw）角，而本章所提出的 3D 频域 VO 表现最好。

最后，为了不失一般性，这些算法也在公共数据集 *CVLIBS* 上进行了评估，结果如图2.13所示。从图中可以看出，3D 频域 VO 算法在该室外数据集上取得了较好的效果。在该数据集上，SVO2 也参与了评估。需要注意的是，SVO2 需要在高质量、高分辨率的图像上才能较好的工作，即它在作者采集的低分辨率全景数据集上并不能工作，因此只在该数据集的评估中给出了 SVO2 的结果。

表 2.2 在所有数据集上进行旋转估计的评估误差

Table 2.2 Average error of Rotation estimation on all datasets

		3D 频域 VO	光流法	ORB	AKAZE
roll	$\epsilon[rad]$	**0.061** ± 0.038	0.074 ± 0.045	0.163 ± 0.105	0.131 ± 0.075
pitch	$\epsilon[rad]$	**0.113** ± 0.090	~~0.240 ± 0.207~~*	0.293 ± 0.248	0.202 ± 0.164
yaw	$\epsilon[rad]$	**0.084** ± 0.065	~~0.057 ± 0.050~~*	0.218 ± 0.153	0.152 ± 0.097
$\mu(\epsilon)$	$[rad]$	**0.088** ± 0.068	~~0.136 ± 0.120~~ *	0.227 ± 0.174	0.163 ± 0.115
Fail	[%]	0	28	0	0

* 在某些帧上，光流法未能成功跟踪相机的俯仰和偏航运动

表2.2给出了在这三个数据集上旋转估计的均方误差 (Root Mean Square Error, RMSE) 及其标准差。虽然表中显示光流法的 RMSE 非常小，但是这并不能体现光流法的实际性能，原因是光流法在部分数据集上未能准确跟踪相机运动，而这些失败情形并未记入 RMSE 计算中。如表2.2中的最后一行所示，光流法的失败

频率大约为 28%，分别对应图2.11和2.12中的情况。

总体来说，3D 频域 VO 表现得非常好，它在各种不同环境中都表现得较为鲁棒，其中包括有大量重复特征的草坪环境。若以位姿估计误差的倒数来描述准确度，则该算法的准确度大约是 ORB 算法准确度的三倍，是基于 AKAZE 的 VO 算法准确度的两倍。

在评估算法的旋转估计性能时，估计的位姿和旋转真值之间的均方误差是主要评估指标。然而该指标不能直接用于平移估计的评估，因为单目相机不能恢复出绝对尺度，所以估计得到的平移的模是不确定的。并且和旋转测量不同，惯性测量装置通常测得的绝对平移运动非常不精确，不适合作为比较基准。因此，本实验用估计的平移的方向和真值方向的误差来作为平移评估的标准，在下文中缩写为均方角度误差（Root Mean Square Angle Error, RMSAE）。

图2.14比较了不同方法在 *OVMIS* 数据集[5] 上平移估计的 RMSAE 及其标准差。本实验中所用的图像序列为该数据集中的一个室外场景图像序列，包含 318 张图像，运动轨迹大约为 25.27 米。为了减少累积误差的影响，本实验中仅考虑帧间的平移误差。相关结果展示在图2.14中，本章提出的 3D 频域 VO 算法仍是其中表现最好的。

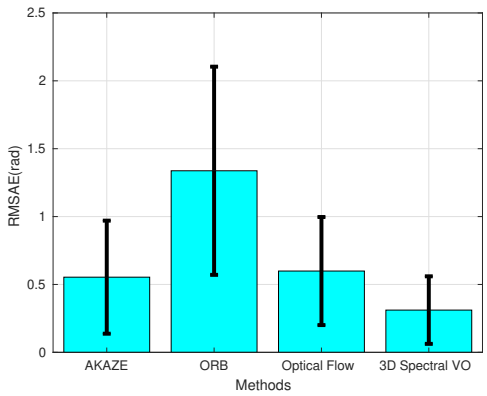


图 2.14 不同方法在 *OVMIS* 数据集[5] 上的平移误差

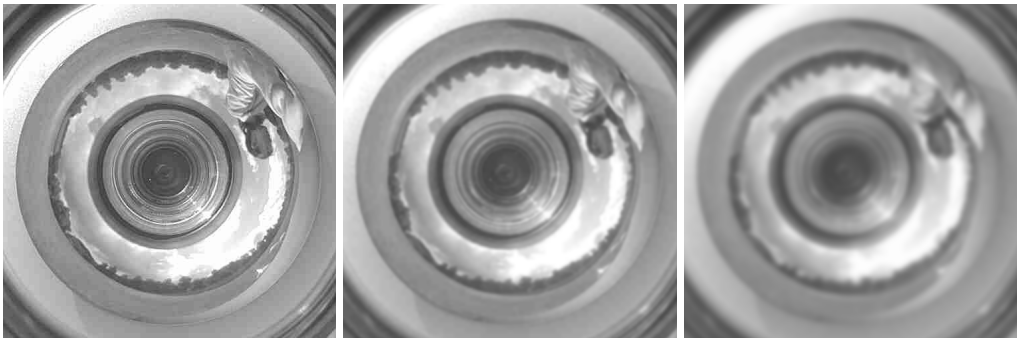
Figure 2.14 Translation error of the different methods on the *OVMIS* dataset[5]

2.5.3.3 可见度较低条件下的鲁棒性测试

正如第 1章所讨论的，频域配准方法在可见度较为苛刻的条件下仍能表现得很好，如运动模糊、烟雾场景下、浑浊的水下环境等。为了系统性地评估本章的 3D 频域 VO 算法，本实验在 *office*、*lawn* 和 *CVLIBS* 数据集上增加不同程度的高

斯模糊来仿真这些可见度苛刻的场景。图2.15展示了不同程度的高斯模糊下各个数据集的图像示例。

(a) 办公室场景 *office*



(b) 草坪场景 *lawn*

(c) ***CVLIBS*** **数据集**

图 2.15 模糊大小分别为 0、10 **和** 20 **像素的模糊图像**

Figure 2.15 Blur images with blur size 0, 10 and 20 pixels

本实验从每个数据集中选取了五对图像，每对图像中有一帧加了高斯模糊，每组实验中包含了 30 种不同程度的高斯模糊。为了减少 RANSAC 算法随机性的影响，每组实验都重复进行了 10 次。本实验中分别测试了算法对横滚、俯仰和偏航角的估计，也对比了 3D 频域 VO 和基于 ORB、AKAZE 和光流法的 VO 的效果。

实验结果如图2.17所示，主要展示了不同算法的位姿估计的 RMSE 和标准差。从图中可看出，基于 AKAZE（绿色）的方法在该实验中表现得最为鲁棒，主要得益于其针对高斯尺度空间的设计。3D 频域 VO（深蓝色）的表现仅次于 AKAZE（绿色），两者的误差均小于基于 ORB（浅蓝）的方法。此外，基于光流法（玫红）的 VO 也没有 3D 频域 VO（深蓝）和基于 AKAZE（绿色）的方法鲁棒。

图2.17的图例上还给出了各个算法的平均运行时间，可以看到光流法的速度最快，大约比 AKAZE 快十倍；其次就是 3D 频域 VO 和 ORB，两者的运行速度相差不多，大约都比 AKAZE 快七倍。

综上所述，本章所提出的 3D 频域 VO 方法对于高斯模糊的鲁棒程度，与 AKAZE 不相上下，但是 3D 频域 VO 的运行速度更快。尽管基于 ORB 的方法比 3D 频域 VO 运行速度稍快，但是 ORB 在2.5.3.2节的实验中误差较大。至于光流法，本实验中得出的结论类似，即虽然其运行速度快于 3D 频域 VO，但是没有 3D 频域 VO 算法鲁棒。尽管不在本文讨论范围之内，但 FMT 本身及其在常规的使用中特别适合通过图形处理单元（GPU）[127] 或数字信号处理器（DSP）来进行加速，甚至只使用现场可编程门阵列（FPGA）[128] 的硬件，都可以实现几个数量级的加速。

图 2.16 含动态物体的图像示例

Figure 2.16 An image with dynamic objects

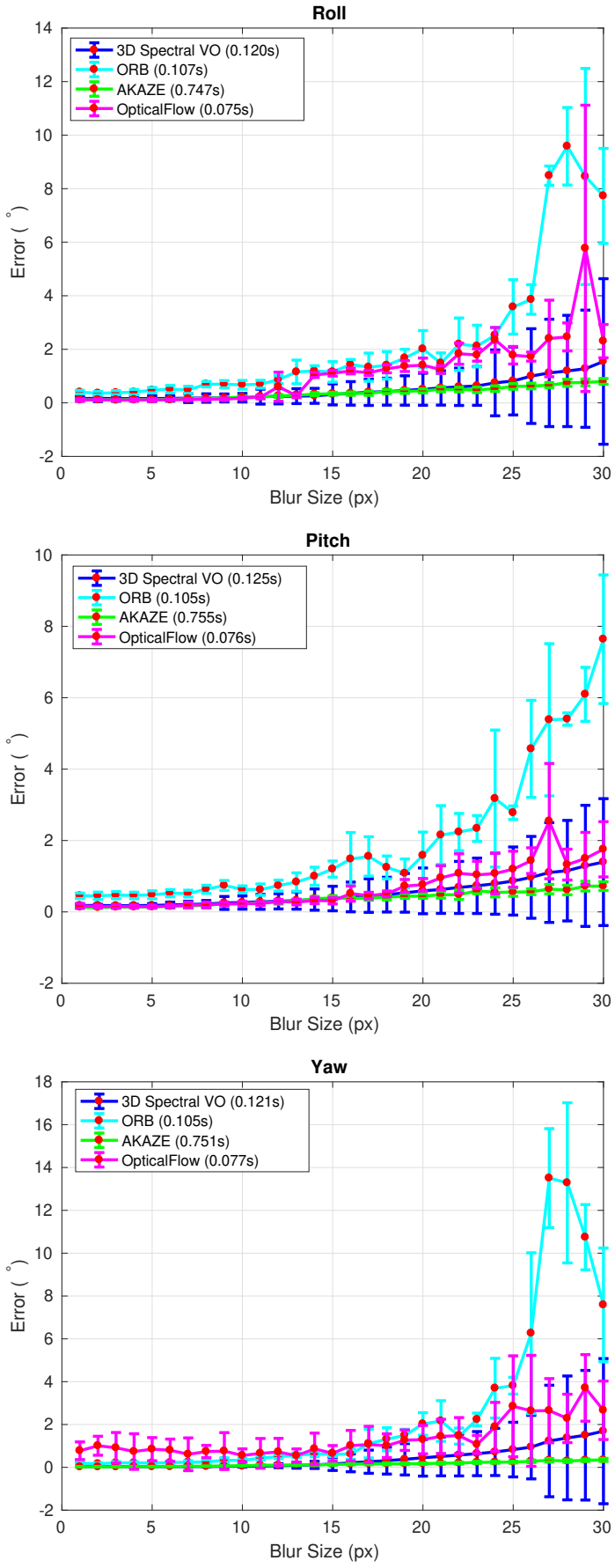


图 2.17 不同算法在模糊图像上的估计误差与标准差

Figure 2.17 Estimation error and standard deviation of algorithms on blurred images

2.5.3.4 在动态环境下的性能评估

为了更进一步地分析算法的鲁棒性，本节中进行了动态场景下的实验。为此，作者在有一个动态物体的场景下分别采集了横滚、俯仰和偏航运动的数据集。图2.16给出了该数据集的一帧示例，动态物体位于红色框中。

图2.18展示了不同算法在该含动态物体的数据集上的性能。可以看到，光流法未能成功跟踪横滚、俯仰运动，主要原因是匹配上的特征点对从 30 对降至了 5 对左右。尽管 ORB 在该数据集上表现地比光流法鲁棒，但是在相机发生俯仰运动时，其累计误差仍然是比较大的。ORB 在该数据集表现得比光流法鲁棒主要得益于更多匹配的特征点对，其在每帧图像上平均可提取 500 个特征，其中约有 150 个特征可以被匹配上，然后五点法大约将其中的一半视作了内点用来估计位姿。相比之下，AKAZE 可以提取到更多的特征点，约为 750 个，其中约有 2/3 为匹配点对，匹配点对中又有超过 1/2 的部分在 RANSAC 五点法中作为内点进行位姿估计，因此 AKAZE 表现得比 ORB 更好。最后，可以看到本章所提出的 3D 视觉频域 VO 在该动态环境给出了最为鲁棒和准确的结果，比以上三者都表现得更好。

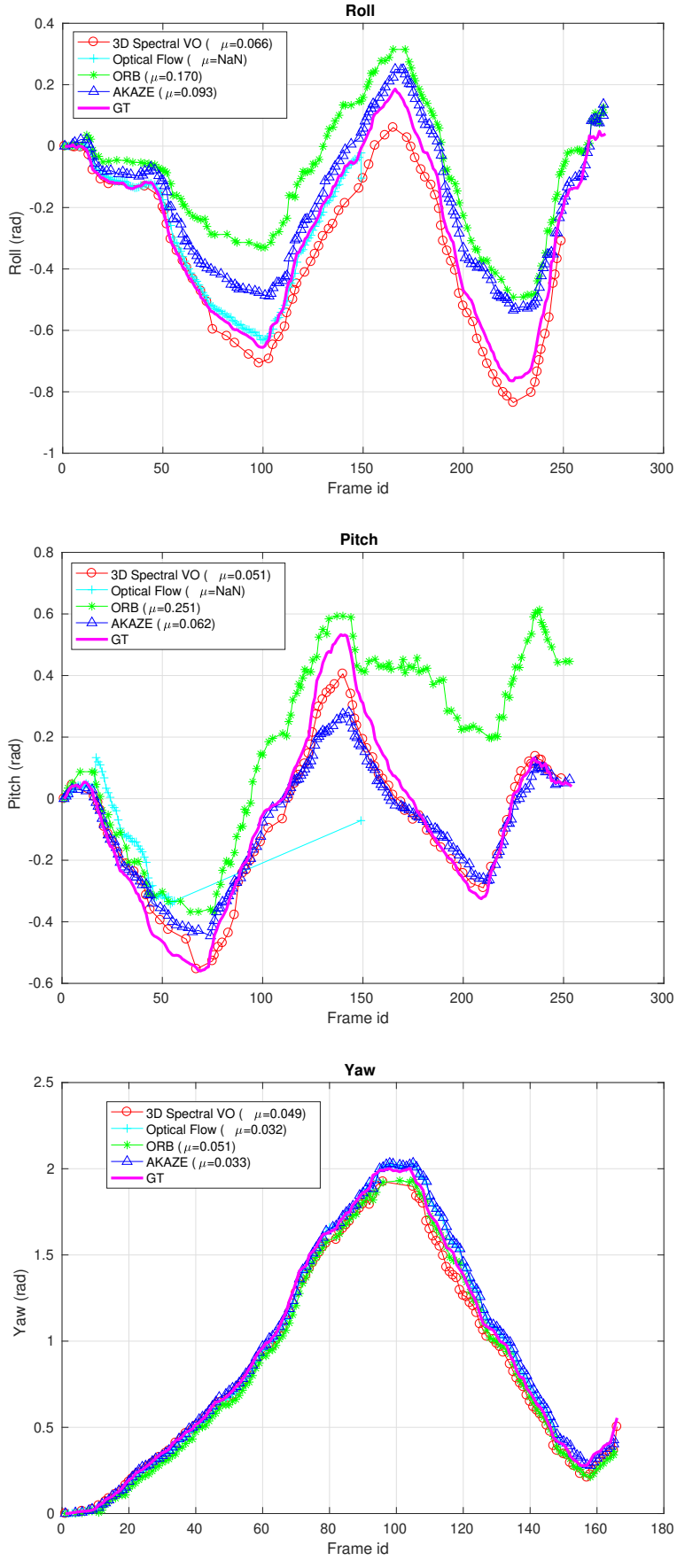


图 2.18 不同方法在动态物体数据集上的性能评估及其平均误差 μ

Figure 2.18 Performance evaluation on the dataset with dynamic objects and average error μ

2.6 小结

本章提出了 3D 频域 VO 算法，它主要利用 FMT 在稀疏子图对集合上配准，从而得到匹配的一致点对，进而估计相机位姿。该方法主要应用于全向图像。实验表明该 3D 频域 VO 算法在有挑战性的环境中存在较大潜力，比如特征较少的场景，多烟多雾、运动模糊、浑浊的水下等可见度较低的条件下，有动态物体的环境等。在本章中的几个测试数据集上，该算法的表现比传统的特征点法（AKAZE 和 ORB）和光流法更为鲁棒，位姿估计的结果也更准确。

第 3 章　基于正弦曲线拟合的全向相机姿态估计

第2章中展示了 FMT 在弱纹理场景下能够进行鲁棒、准确的全向图像配准，在利用配准结果估计相机姿态时，采用了传统的五点法，对相机标定的精度要求较高。本章基于全向相机的自身性质，将全向相机的旋转建模成正弦曲线的拟合问题，从而提出一种基于正弦曲线拟合的全向相机姿态估计方法。该方法只需要简单的标定即可，因此在廉价相机和高端相机上均可使用。文献[63,129] 中都未采用几何法进行全向相机的姿态估计，而是将全向图像转换成全景图像，然后通过水平方向的像素移动来估计全向相机在 2D 平面上的朝向。受此启发，本章中重新考虑了全向相机在三维空间中的旋转，发现当全向相机发生横滚或者俯仰时，对应的全景图像上的像素会在竖直方向上以某种规律运动，当全向相机发生偏航时，全景图像上的像素会在水平方向上以某种规律运动，而当三种旋转同时发生时，这些规律仍然存在。

本章通过数学推导证明当全向相机发生旋转时，全景图像上像素的移动符合正弦曲线形状，而且如果将全向图像沿水平方向分割成不同的子图，这些子图的平面旋转角度大小也符合正弦曲线形状。因此，本章将全向相机的旋转建模成了正弦曲线拟合问题，通过拟合像素的移动以及子图的旋转可以直接估计出帧间的旋转，从而估计出相机的朝向。为了不失一般性，在建模过程中本章也考虑了全向相机的平移对模型的影响。但是由于相机的平移对应的像素运动和深度有关，会给平移估计带来一定难度，因此在本章中并不涉及用正弦曲线模型进行全向相机的平移估计，但在相关实验中将评估建模过程中考虑平移项和不考虑平移项对算法性能的影响。详细的建模过程以及平移估计的难点将在3.1.2节中介绍。

此外，为了保证正弦曲线拟合算法能够正常工作，需要保证像素的移动和子图的旋转估计具有较高的置信度。本章将采用 FMT 和光流法两种方法估计像素的移动，并对比它们对相机姿态估计的影响。

本章的主要贡献总结如下：

- 发现全向相机的旋转不仅使得像素按正弦曲线运动，还会使得子图的旋转符合正弦曲线规律；

- 提出了一种新颖的基于正弦曲线拟合的相机姿态估计方法；
- 在不同数据集上，对比了该基于正弦曲线拟合的方法和几何法进行相机姿态估计的性能。

3.1 算法设计

3.1.1 相机模型和校准

为了简化计算，本章将全向相机建模成圆柱模型。由于相机的特性和制造工艺的差异，无法保证图像上每个像素都是方的。换句话说，在真实世界中，同一深度下的同一长度可能会在 u 和 v 方向上投影的像素长度不同。为了保持分辨率一致性，在一开始就要校准宽度和高度之间的像素比，来保证每个像素都是方的。圆柱模型以及像素校准均在1.1.1节中进行了介绍，本章不再赘述。该模型的参数较少，易于标定，因此即使是像素分辨率较低的廉价相机也能进行很好地标定。

值得注意的是，在全景图像中，如果 $u = 0$ 的列对应这相机的正 x 轴（即机器人的正面），则左边开始的 $u = \frac{1}{4}u_{max}$ 的像素则对应着相机的正 y 轴（即机器人的左侧），相应地，负 x 轴方向则在全景图像的 $\frac{1}{2}u_{max}$ 处。在不失一般性的前提下，为了简化公式，本章假设 x 轴始终位于 $u = 0$（图3.1中的灰色虚线），因而 y 轴位于 $\frac{1}{4}u_{max}$（图3.1中的绿线）。本章接下来的部分假设每个像素都是方的，相机模型为标定好的圆柱模型。

3.1.2 全景图像的运动模型

本小节从数学上证明当相机发生运动时，其对应的全景图像上像素的移动遵循正弦曲线规律，然后3.1.3节将介绍算法中关于运动向量的提取。

关于运动模型的假设如下：

1. 相机移动很小，如旋转角度的大小 $\|\Theta\| \leq 5°$；
2. 每个像素都是方的。

当相机发生运动时，连续两帧之间的移动很小，因此第一个假设可以满足；在进行相机的标定以及分辨率一致性的调整之后，第二个假设可以被满足。

本章将折反射全向相机的运动建模为正弦曲线，该模型可表示为：

$$y = A\sin(x + \phi) + B \tag{3.1}$$

其中 A、B 和 ϕ 分别表示正弦曲线的幅度、偏移和初相；相位 x 由全景图像列索引 u_p 和频率给定，该正弦曲线的频率 γ 已知，可由全景图像的宽度及其对应的视角确定。相机的运动会导致全景图像上 u 和 v 方向上的像素移动，但是 u 和 v 方向上的运动对应的正弦曲线参数不同，可以表示为：

$$\Delta v(u_p) = \gamma \|\Theta_{xy}\| \cdot \sin\left(\gamma u_p - \hat{\Theta}_{xy}\right) + \lambda_i \mathrm{t}_z \,, \tag{3.2a}$$

$$\Delta \theta(u_p) = \|\Theta_{xy}\| \cdot \sin\left(\gamma u_p - \hat{\Theta}_{xy} + \frac{\pi}{2}\right) \,, \tag{3.2b}$$

$$\Delta u(u_p) = \lambda_i \|\mathrm{t}_{\mathrm{xy}}\| \cdot \sin\left(\gamma u_p + \hat{\mathrm{t}}_{xy}\right) + \gamma \Theta_z \,. \tag{3.2c}$$

其中，Θ 代表相机的旋转，t 代表相机的平移，θ 代表区域图像的旋转。这也是本章的**核心公式**。图像的旋转和相机姿态中的旋转是不同的，本章使用 θ 和 Θ 来区分这两者。

需要关注的是，这三个正弦函数（公式(3.2a)、(3.2b)和(3.2c)）涵盖了相机运动的六个自由度参数。可以看到，公式(3.2b)与公式(3.2a)十分相似，除了初相相差 $\frac{pi}{2}$、幅度上的放缩因子 γ 和平移项 $\lambda_i \mathrm{t}_z$ 不同外，待估计的参数 $\|\Theta_{xy}\|$ 和 $\hat{\Theta}_{xy}$ 是相同的，这意味着它们可以进行联合优化。尽管这些正弦函数也涵盖了平移项，但是在与像素深度相关的参数 λ_i 未知情况下，平移项很难被恢复。而对于旋转来说，像素的深度并不会产生影响，因此本章仅将平移项用作旋转估计的辅助，而不进行平移估计。关于平移项对旋转估计的影响将在3.3.2节进行详细讨论。

3.1.2.1　直观分析

首先，我们直观地分析了四种情况下的相机运动，来说明像素运动实际上是如何遵循正弦函数的：

- **绕 x 轴或 y 轴的旋转 (Θ_{xy})**：（*以绕 x 轴的旋转：横滚为例，见图3.1*）全景图像的列索引 u_p 越靠近 x 轴（不管正轴或负轴），在 v 方向上的像素移动越小；列索引 u_p 越靠近 y 轴（不管正轴或负轴），在 v 方向上的像素移动越大。见公式(3.2a)。

对于横滚运动（绕 x 轴的旋转）而言，公式(3.2b)表示的图像旋转 $\Delta\theta(u_p)$，将在 $u_p = 0$ 和 $u_p = \frac{1}{2}u_{max}$ 处取得最大值，这两列分别对应相机模型的 x 轴正方向和负方向。列索引 u_p 越靠近 y 轴（不管正轴或负轴），旋转越小.

- **沿着 z 轴方向的平移 (t_z)**：此时，每一列 u_p 上像素的移动都是一样的，均是 v 方向上的运动，因此对应着正弦曲线的偏移量，见公式(3.2a)。

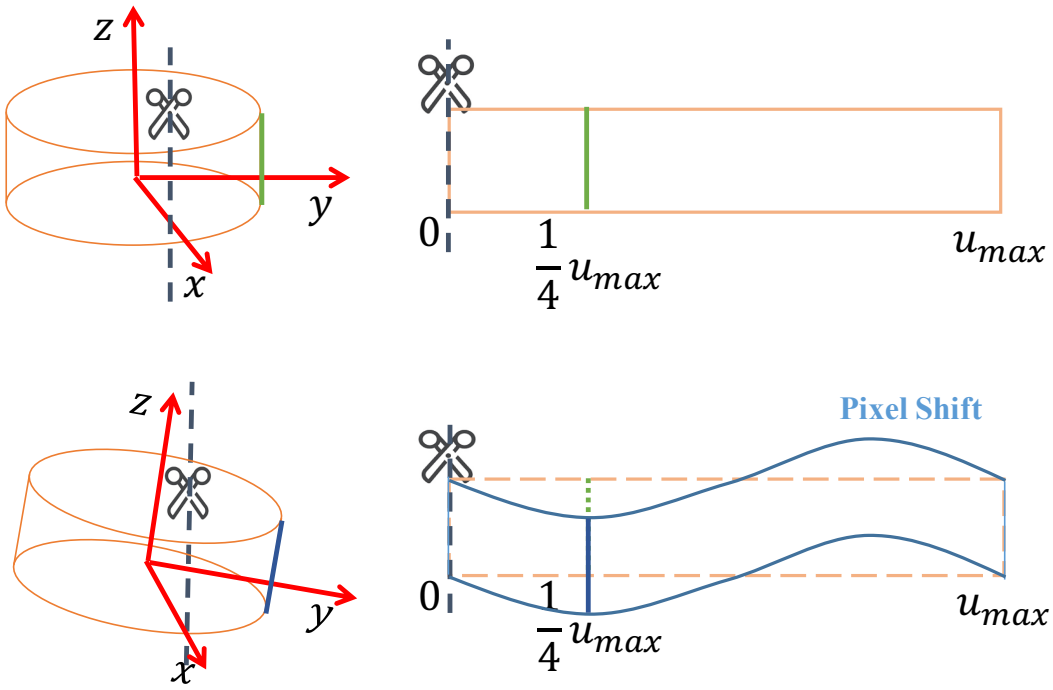


图 3.1 绕 x 轴旋转（横滚运动）的直观分析示例

Figure 3.1 Intuitive demonstration of a rotation around the x-axis (roll)

• **绕 z 轴的旋转** (Θ_z)：当相机绕 z 轴旋转，即发生偏航时，每一列 u_p 的像素都会沿图像 u 方向发生位移，在公式(3.2c)中对应着偏移量。

• **沿 x 轴或 y 轴的平移** (t_{xy})：（以沿着 x 轴方向的平移为例）和绕 x 轴旋转的情况类似，全景图像的列索引 u_p 越靠近 x 轴（不管正轴或负轴），在 u 方向上的像素移动越小；列索引 u_p 越靠近 y 轴（不管正轴或负轴），在 u 方向上的像素移动越大。该规律和公式(3.2c)一致。

接下来，将对该正弦曲线运动模型进行数学解释，并给出相机旋转下该正弦模型的详细推导。假定两帧图像 ${}^1I_p$ 和 ${}^2I_p$，其对应的相机位姿之间的变换为 ${}^1_2T = [{}^1_2\Theta, {}^1_2t]$。假设第二帧全景图像 ${}^2I_p$ 上任意一点 ${}^2p = (u, v)^\top$，它的圆柱坐标 ${}^2P$ 为：

$$ {}^2P = \begin{bmatrix} x_2 \\ y_2 \\ z_2 \end{bmatrix} = \begin{bmatrix} r\cos\frac{u}{r} \\ r\sin\frac{u}{r} \\ \frac{H}{2} - v \end{bmatrix} \tag{3.3} $$

本节接下来的内容将主要分析当相机发生旋转时，全景图像上每一行、每一列的位移。思路如下：首先，利用给定的变换矩阵 ${}^1_2T$ 将三维点 ${}^2P$ 转换为 ${}^1P = [x_1, y_1, z_1]^\top$；其次，找到射线 ${}^1PO$ 与圆柱 $\{C : x^2 + y^2 = r^2\}$ 之间的交点 ${}^1\bar{P} =$

$[\bar{x}_1, \bar{y}_1, \bar{z}_1]^\top$ (O 为圆柱中心)，然后将其转换到全景图 ${}^1I_p$ 上的点 ${}^1p$；最后，分别计算 ${}^1p$ 和 ${}^2p$ 之间的行位移和列位移。

3.1.2.2　数学分析

为了简化数学推导过程，该模型只关注全景图像最中间的行，即 $v = \frac{H}{2}$。此外，为了单独分析绕各个轴的旋转，将相机旋转 Θ 的欧拉角定义为 $(\alpha_x, \alpha_y, \alpha_z)^\top$。

绕 x 轴的旋转　两帧之间的相机位姿变换矩阵 T 为

$$ {}^1_2T = \left[\begin{array}{ccc:c} 1 & 0 & 0 & 0 \\ 0 & \cos\alpha_x & -\sin\alpha_x & 0 \\ 0 & \sin\alpha_x & \cos\alpha_x & 0 \end{array}\right] \tag{3.4a} $$

经过变换后的点 ${}^1P$ 为

$$ {}^1P = \left[\begin{array}{c} r\cos\frac{u}{r} \\ r\sin\frac{u}{r}\cos\alpha_x - (\frac{H}{2} - v)\sin\alpha_x \\ r\sin\frac{u}{r}\sin\alpha_x + (\frac{H}{2} - v)\cos\alpha_x \end{array}\right] = \left[\begin{array}{c} r\cos\frac{u}{r} \\ r\sin\frac{u}{r}\cos\alpha_x \\ r\sin\frac{u}{r}\sin\alpha_x \end{array}\right] \tag{3.4b} $$

则交点 ${}^1\bar{P}$ 为

$$ {}^1\bar{P} = \left[\begin{array}{c} \frac{r}{\sqrt{1+k^2}} \\ \frac{kr}{\sqrt{1+k^2}} \\ \frac{r\sin\frac{u}{r}\sin\alpha_x}{r\cos\frac{u}{r}}\ \frac{r}{\sqrt{1+k^2}} \end{array}\right] \tag{3.4c} $$

其中 $k = \frac{r\sin\frac{u}{r}\cos\alpha_x}{r\cos\frac{u}{r}}$。从而可以得到列方向的位移 Δv：

$$ \begin{aligned} \Delta v =& (\frac{H}{2} - z_2) - (\frac{H}{2} - z_1') \\ =& \bar{z}_1 - z_2 \\ =& \frac{r(r\sin\frac{u}{r}\sin\alpha_x)}{\sqrt{(r\cos\frac{u}{r})^2 + (r\sin\frac{u}{r}\cos\alpha_x)^2}} \\ \approx& \frac{r(r\sin\frac{u}{r}\alpha_x)}{\sqrt{r^2}} \\ =& \alpha_x r\sin\frac{u}{r} \end{aligned} \tag{3.4d} $$

需要注意的是只有当假设相机的运动较小成立时，公式(3.4d)中的约等号才成立，因为运动较小时，$\sin\alpha_x \approx \alpha_x$，$\cos\alpha_x \approx 1$。类似地，可以推导出行位移 Δu：

$$\begin{aligned}\Delta u &= r\arctan\frac{y_2}{x_2} - r\arctan\frac{\bar{y}_1}{\bar{x}_1}\\ &= r\arctan\left(\tan\frac{u}{r}\right) - r\arctan k\\ &= u - r\arctan\frac{r\sin\frac{u}{r}\cos\alpha_x}{r\cos\frac{u}{r}}\\ &\approx u - r\arctan\frac{r\sin\frac{u}{r}}{r\cos\frac{u}{r}}\\ &= 0\end{aligned} \tag{3.4e}$$

Δu 与 Δv 成立的条件相同。

绕 z 轴的旋转　两帧之间的相机位姿变换矩阵 T 为

$${}^1_2T = \left[\begin{array}{ccc:c}\cos\alpha_z & -\sin\alpha_z & 0 & 0\\ \sin\alpha_z & \cos\alpha_z & 0 & 0\\ 0 & 0 & 1 & 0\end{array}\right] \tag{3.5a}$$

经过变换后的点 ${}^1P$ 为

$${}^1P = \begin{bmatrix} r\cos(\frac{u}{r}+\alpha_z)\\ r\sin(\frac{u}{r}+\alpha_z)\\ \frac{H}{2}-v\end{bmatrix} = \begin{bmatrix} r\cos(\frac{u}{r}+\alpha_z)\\ r\sin(\frac{u}{r}+\alpha_z)\\ 0\end{bmatrix} \tag{3.5b}$$

交点 ${}^1\bar{P}$ 和 ${}^1P$ 为同一个点：

$${}^1\bar{P} = {}^1P = \begin{bmatrix} r\cos(\frac{u}{r}+\alpha_z)\\ r\sin(\frac{u}{r}+\alpha_z)\\ 0\end{bmatrix} \tag{3.5c}$$

从而得到了行方向上的位移 Δu：

$$\Delta u = r\arctan\frac{y_2}{x_2} - r\arctan\frac{\bar{y}_1}{\bar{x}_1} = -r\alpha_z \tag{3.5d}$$

此时，由于在 z 轴方向上，点 ${}^2P$ 和点 ${}^1\bar{P}$ 没有区别，因此 Δv 等于 0。

混合旋转　对于混合旋转的情况，变换矩阵是横滚角 α_x、俯仰角 α_y 和偏航角 α_z 的组合，即

$$
{}^1_2T = \left[\begin{array}{ccc:c} czcy & czsysx - szcx & czsycx + szsx & 0 \\ szcy & szsysx + czcx & szsycx - czsx & 0 \\ -sy & cysx & cycx & 0 \end{array}\right] \tag{3.6a}
$$

其中，为了方便推导，作如下简写：$sx = \sin\alpha_x$, $cx = \cos\alpha_x$, $sy = \sin\alpha_y$, $cy = \cos\alpha_y$, $sz = \sin\alpha_z$ 以及 $cz = \cos\alpha_z$。然后，对点 ${}^2P$ 应用该变换矩阵，得到的变换之后的点 ${}^1P$ 为：

$$
{}^1P = \begin{bmatrix} r\cos(\frac{u}{r})czcy + r\sin(\frac{u}{r})(czsysx - szcx) \\ r\cos(\frac{u}{r})szcy + r\sin(\frac{u}{r})(szsysx - czcx) \\ -r\cos(\frac{u}{r})sy + r\sin(\frac{u}{r})cysx \end{bmatrix} \tag{3.6b}
$$

然后，将该点投影到圆柱上，得到 ${}^1\bar{P}$：

$$
{}^1\bar{P} = \begin{bmatrix} \frac{r}{\sqrt{1+k^2}} \\ \frac{kr}{\sqrt{1+k^2}} \\ \frac{-r\cos(\frac{u}{r})sy + r\sin(\frac{u}{r})cysx}{r\cos(\frac{u}{r})czcy + r\sin(\frac{u}{r})(czsysx - szcx) + f_x}\frac{r}{\sqrt{1+k^2}} \end{bmatrix} \tag{3.6c}
$$

其中 $k = \frac{r\cos(\frac{u}{r})szcy + r\sin(\frac{u}{r})(szsysx - czcx) + f_y}{r\cos(\frac{u}{r})czcy + r\sin(\frac{u}{r})(czsysx - szcx) + f_x}$。然后，利用相机模型和合理假设的近似，可以得到相应的行位移与列位移。具体而言，

$$
\begin{aligned}
\Delta u &= r\arctan\frac{y_2}{x_2} - r\arctan\frac{\bar{y}_1}{\bar{x}_1} \\
&= r\arctan\frac{\sin\frac{u}{r}}{\cos\frac{u}{r}} - r\arctan k \\
&= r\arctan\frac{(\frac{\sin\frac{u}{r}}{\cos\frac{u}{r}}) - k}{1 + k\frac{\sin\frac{u}{r}}{\cos\frac{u}{r}}} \\
&\quad\vdots \\
&\approx r\arctan\frac{r\sin^2\frac{u}{r}\alpha_x\alpha_y - r\alpha_z - \frac{r}{2}\sin\frac{2u}{r}\alpha_x\alpha_y\alpha_z}{r + \frac{r}{2}\sin\frac{2u}{r}\alpha_x\alpha_y + r\sin^2\frac{u}{r}\alpha_x\alpha_y\alpha_z} \\
&\approx r\arctan(-\alpha_z) \\
&\approx -r\alpha_z
\end{aligned} \tag{3.6d}
$$

$$
\begin{aligned}
\Delta v &= \bar{z}_1 - z_2 \\
&= \frac{r(\sin(\frac{u}{r})cysx - \cos(\frac{u}{r})sy)}{r\cos(\frac{u}{r})czcy + r\sin(\frac{u}{r})(czsysx - szcx)} \frac{r}{\sqrt{1+k^2}} \\
&\quad \vdots \\
&\approx \frac{r(\sin\frac{u}{r}\alpha_x - \cos\frac{u}{r}\alpha_y)}{\sqrt{(1+\alpha_z^2)(1+\sin^2\frac{u}{r}\alpha_y^2\alpha_x^2 + 2\cos\frac{u}{r}\sin\frac{u}{r}\alpha_y\alpha_x)}} \\
&\approx r\sin\frac{u}{r}\alpha_x - r\cos\frac{u}{r}\alpha_y \\
&= \sqrt{\alpha_y^2+\alpha_x^2}r\sin(\frac{u}{r} + \arctan\frac{\alpha_x}{\alpha_y})
\end{aligned}
\tag{3.6e}
$$

图像旋转 首先说明，这里的图像旋转也是相对于不同的列而言的，由于单独的一列谈不上旋转，此处每一列的图像旋转是指以该列为中心线的的子图的旋转。关于如何估算图像旋转将在3.1.3节中详细说明。

对于图像旋转 $\Delta\theta$(公式(3.2b))，仅考虑图像中的一个像素不足以进行推导。因此，将每列的方向向量作为参考进行推导。为了从相机运动中提取图像旋转信息，需要将旋转后的方向矢量投影到参考列索引的切平面上。接下来是详细推导。

假设相机直立，即 z 轴朝上，则在三维空间中每列的方向向量都是 $v_2 = (0,0,1)^\top$。由于相机的偏航和位移不会造成子图旋转，因此为简化起见，推导过程中仅考虑了横滚和俯仰，相应的旋转变换矩阵为：

$$
{}^1_2\Theta = \begin{bmatrix} cy & sysx & sycx \\ 0 & cx & -sx \\ -sy & cysx & cycx \end{bmatrix} \tag{3.7}
$$

旋转方向向量 v_1 为

$$
\begin{aligned}
v_1 = {}^1_2\Theta v_2 &= \begin{bmatrix} cy & sysx & sycx \\ 0 & cx & -sx \\ -sy & cysx & cycx \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} \sin\alpha_y\cos\alpha_x \\ -\sin\alpha_x \\ \cos\alpha_y\cos\alpha_x \end{bmatrix}
\end{aligned}
\tag{3.8}
$$

为了接下来推导，本节首先回顾平面上直线的投影的性质。假设在三维空间中，A 为任意一条直线，B 是某一平面的法向量，则直线 A 在平面 B 上的投影为

$$\mathrm{A}\|\mathrm{B} = \mathrm{B} \times (\mathrm{A} \times \mathrm{B}) \tag{3.9}$$

或

$$\mathrm{A}\|\mathrm{B} = (\mathrm{B} \times \mathrm{A}) \times \mathrm{B} \tag{3.10}$$

其中，“×”表示两个向量之间的内积。

对于图像的列索引 u_p，它的切平面法向量为 $\mathrm{n} = (\cos(\gamma u_p), \sin(\gamma u_p), 0)^\top$。利用公式(3.9)，可将旋转后的方向向量 v_1 投影到切平面上，得到其在切平面上的投影方向向量为

$$\begin{aligned}\hat{v}_1 &= \mathrm{n} \times (v_1 \times \mathrm{n}) \\ &= \begin{bmatrix} \sin\alpha_x \cos\gamma u_p \sin\gamma u_p + \sin\alpha_y \cos\alpha_x \sin^2\gamma u_p \\ -\sin\alpha_x \cos^2\gamma u_p - \sin\alpha_y \cos\alpha_x \cos\gamma u_p \sin\gamma u_p \\ \cos\alpha_y \cos\alpha_x \end{bmatrix}\end{aligned} \tag{3.11}$$

然后可以用如下公式提取每个列索引对应的图像旋转角度。

$$\begin{aligned}\cos\Delta\alpha &= \frac{v_2 \cdot \hat{v}_1}{\|v_2\|_2 \|\hat{v}_1\|_2} \\ &\approx \frac{1}{\sqrt{(\alpha_x \cos\gamma u_p + \alpha_y \sin\gamma u_p)^2 + 1}} .\end{aligned} \tag{3.12}$$

当相机旋转较小时，这里的约等号成立。然后，可以得到

$$\begin{aligned}\sin\Delta\alpha &= \frac{\alpha_x \cos\gamma u_p + \alpha_y \sin\gamma u_p}{\sqrt{(\alpha_x \cos\gamma u_p + \alpha_y \sin\gamma u_p)^2 + 1}} \\ &= \frac{\sqrt{(\alpha_x^2 + \alpha_y^2)}\sin(\gamma u_p + \arctan\frac{\alpha_x}{\alpha_y})}{\sqrt{(\sqrt{(\alpha_x^2 + \alpha_y^2)}\sin(\gamma u_p + \arctan\frac{\alpha_x}{\alpha_y}))^2 + 1}} \\ &= \frac{\sqrt{(\alpha_x^2 + \alpha_y^2)}\sin(\gamma u_p + \arctan\frac{\alpha_x}{\alpha_y})}{\sqrt{(\alpha_x^2 + \alpha_y^2)\sin(\gamma u_p + \arctan\frac{\alpha_x}{\alpha_y}) + 1}} \\ &\approx \sqrt{(\alpha_x^2 + \alpha_y^2)}\sin(\gamma u_p + \arctan\frac{\alpha_x}{\alpha_y}) .\end{aligned} \tag{3.13}$$

当三维旋转足够小的时候，最后一个约等号成立。最后可以得到如下结论

$$\Delta\theta \approx \sqrt{(\alpha_x^2 + \alpha_y^2)}\sin(\gamma u_p + \arctan\frac{\alpha_x}{\alpha_y}) . \tag{3.14}$$

3.1.3 运动向量的提取

根据上述推导，可以从像素的位移中提取相机姿态。为此，首先需要计算像素的位移，即运动向量，然后通过正弦曲线拟合估算出正弦曲线的参数，从而估计相机姿态。本节将简要介绍两种计算像素位移的方法：光流法和 FMT。关于正弦曲线拟合的详细介绍将在3.1.4节中给出。

光流法[130] 是用于目标跟踪的经典方法之一。其假设在一段时间间隔内光度是恒定的，则像素强度也会保持相同。基于该假设，可利用两帧之间的像素强度一致性来计算像素在 u 和 v 方向上的位移，即 Δu 和 Δv。由于不能直接从光流中提取出两帧之间的旋转 $\Delta\theta$，因此使用光流法进行实验时，不考虑公式(3.2b)。

FMT 算法[118] 已被成功应用于缺乏特征的环境，如水下环境。它首先将空间域图像变换到频域，然后将频域的图像采样到对数坐标系来估计两个图像之间的旋转 $\Delta\theta$ 和缩放。根据估计得到的旋转和缩放参数，可以逆旋转和逆缩放第二张图像，然后通过相位相关法即可估算出两帧图像之间的平移 $\Delta u, \Delta v$。但是，FMT 算法通常只能用于 2D 图像配准，并要求图像对应的真实场景是平面的且平行于成像平面的。第2章中利用滑动窗口将全向图像被分成子图，由于这些子图像对应的三维环境较少，可以被认为是平面的，因此 FMT 可以用于计算两帧子图之间的变换。为利用 FMT 来计算运动向量（$\Delta u, \Delta v, \Delta\theta$），本章使用了类似的子图策略。

需要说明的是，尽管 FMT 也可以估计图像之间的缩放，但是在本章中我们并未使用缩放。

3.1.4 拟合算法

曲线拟合主要是为了估计公式(3.1) 中的一些未知参数 $\Phi = \{A, \phi, B\}$，然后从这些参数中估计出相机的姿态。估计偏航角最简单的方法就是将公式(3.2c)中的 Δu 取平均，即便相机同时在 $x-y$ 平面内发生平移。但是，类似的方法不能用来估计横滚和俯仰，即公式(3.2a)中的 Θ_{xy}，因为 z 轴的位移会影响幅度 $\|\Theta_{xy}\|$ 的估计。也就是说，公式(3.2a)中的的偏移项（z 轴平移）对于横滚和俯仰估计是必需的，3.3.2节中将通过实验来说明平移项对姿态估计的影响。尽管可以通过取平均值的方法来估计偏航角，但从算法的普适性考虑，本章还是利用优化的方法来估计未知参数 Φ，即将其建模成非线性最小二乘问题，然后从估计得到的 Φ

中计算旋转 Θ。

为了估计公式(3.2a)、(3.2b)和(3.2c)中相应的参数 $\Phi_v = \{\|\Theta_{xy}\|, \hat{\Theta}_{xy}, \lambda_i \mathrm{t}_z\}$，$\Phi_u = \{\lambda_i \|\mathrm{t}_{xy}\|, \hat{\mathrm{t}}_{xy}, \Theta_z\}$，构建了如下两个目标函数：

$$r_v(u_p, \Phi_v) = \Delta v(u_p; \Phi_v) - y_v \tag{3.15a}$$

$$r_\theta(u_p, \Phi_v) = \Delta\theta(u_p; \Phi_v) - y_\theta \tag{3.15b}$$

$$r_u(u_p, \Phi_u) = \Delta u(u_p; \Phi_u) - y_u \tag{3.15c}$$

$$\min_{\Phi_v} L_v(u_p; \Phi_v) = \min_{\Phi_v} L_v^1(u_p; \Phi_v) + \eta L_v^2(u_p; \Phi_v) \tag{3.15d}$$

$$= \min_{\Phi_v} \frac{1}{2}(\left\|r_v(u_p, \Phi_v)\right\|_2^2 + \eta \left\|r_\theta(u_p, \Phi_v)\right\|_2^2) \tag{3.15e}$$

$$\min_{\Phi_u} L_u(u_p; \Phi_u) = \min_{\Phi_u} \frac{1}{2} \left\|r_u(u_p, \Phi_u)\right\|_2^2 \tag{3.15f}$$

其中 η 是优化的权重系数；r_v, r_θ 和 r_u 是平移的残差；y_θ 是图像的旋转，y_v 和 y_u 是 FMT 算法[118] 或光流法估计的行方向或者列方向上的位移；L_v 和 L_u 是标准最小二乘形式的损失函数。基于鲁棒性考虑，本章使用非线性优化 Levenberg-Marquardt (LM) 算法[131] 最小化目标函数，也可以使用其他优化方法代替 LM 算法。

$$L_v^1(u_p; \Phi_v, \delta) = \delta \cdot \left(\sqrt{1 + \left(\frac{r_v(u_p, \Phi_v)}{\delta} \right)^2} - 1 \right) \tag{3.16a}$$

$$L_v^2(u_p; \Phi_v, \delta) = \delta \cdot \left(\sqrt{1 + \left(\frac{r_\theta(u_p, \Phi_v)}{\delta} \right)^2} - 1 \right) \tag{3.16b}$$

$$L_u(u_p; \Phi_u, \delta) = \delta \cdot \left(\sqrt{1 + \left(\frac{r_u(u_p, \Phi_u)}{\delta} \right)^2} - 1 \right) \tag{3.16c}$$

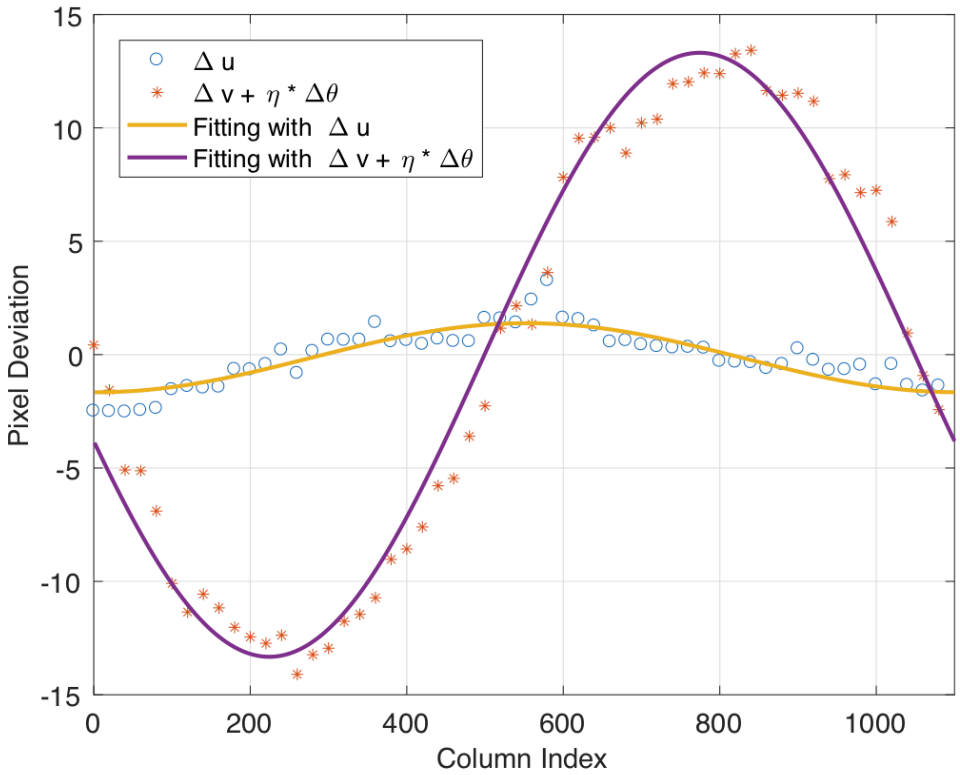
为了处理离群值问题，本章选用了一个鲁棒损失函数：Pseudo-Huber 损失函数。文献[132] 中指出，与 L2 损失函数相比，Huber 损失函数对数据中的离群值更不敏感。而 Pseudo-Huber 损失函数[133] 结合了 L2 损失函数和 Huber 损失函数的最佳特性：当接近于最小值时，它是强凸的，而对于异常值也较为平缓。因此，在实际实现时，用其来替换公式(3.16a)、(3.16b)和(3.16c)中的最小二乘损失形式(公式(3.15d)、(3.15f))。

3.2 算法实现

算法2描述了本章的正弦曲线拟合算法实现，其中对于不同的数据集，W, L, H 和 d 的取值不同。

(a) 图像 1

(b) 图像 2



(c) 拟合结果

图 3.2 在 u 和 v 方向上的位移拟合示例

Figure 3.2 An example of fitting results in u and v direction

首先，用一正方形窗口在全景图像 ${}^1I_p$ 和 ${}^2I_p$ 上沿 u 方向的滑动，提取出一系列子图对。然后利用 FMT 算法来计算每个窗口提取出子图对的二维运动，即一系列的 $\Delta u, \Delta v$ 和 $\Delta\theta$ 以及对应的列索引 $u_p = \frac{1}{2}L + k \times d$；类似地，光流法

算法 2 利用正弦曲线拟合估计全向相机的旋转

1: **输入：** 全向图像 ${}^1I_o, {}^2I_o$；滑动窗口尺寸 $L \times L$ 及其步长 d
2: 通过笛卡尔坐标系与极坐标系之间的转换，得到全景图像 ${}^1I_p, {}^2I_p$，其大小为 $W \times H$
3: **while** $L + k \times d \leq W, k \in \mathbb{N}$ **do**
4: 　　利用第 k 个窗口从 ${}^1I_p$ 和 ${}^2I_p$ 上提取子图，通过 FMT 或光流法计算子图之间的旋转和平移 t_{uv}
5: 　　将 t_{uv} 放入图像运动向量集合 $\mathbb{M}$
6: **end while**
7: 通过 LM 算法对集合 $\mathbb{M}$ 中的值进行正弦曲线拟合，估计参数 Φ_v 和 Φ_u（公式(3.15)）
8: 从估计得到的参数 Φ_v, Φ_u 计算出旋转 Θ（公式(3.1)）
9: **输出：** Θ

也可用于计二维运动，但仅限于位移 Δu 和 Δv。然后，按算法 2中描述的，将这些值进行正弦函数拟合来估计参数 Φ。在本章的实现与实验中，公式(3.16) 中的 $\delta = 0.5$。图3.2c给出了两帧子图进行曲线拟合的一个示例，从图中可以发现优化算法能够有效地进行正弦曲线拟合。可以看到曲线 $\Delta v + \eta\Delta\theta$ 的最大值位于 $\frac{1}{4}u_{max}$ 和 $\frac{3}{4}u_{max}$ 之间。按照之前的约定，x 轴位于 $u = 0$，这意味着图像在 y 轴方向的位移很大，也就是说相机发生了绕 x 轴的旋转，即横滚。

3.3　实验与分析

为了评估所提出算法的性能，本节进行了不同的实验来分析该方法的鲁棒性、计算速度和准确性。本节首先在3.3.1节中分析了加入 FMT 估计得到的旋转进行联合优化的优势；其次，通过实验说明所提出的正弦曲线拟合方法对平移干扰具有鲁棒性（3.3.2）。此外，本节还将该算法与两种基于几何的方法进行比较，即 STEWENIUS 五点算法和 n 点算法，其中后者是专门为纯旋转场景设计的[123]，两种几何算法的实现都利用 OpenGV[1]库；接着，本章将所提出的正弦曲线拟合法和 STEWENIUS 五点法、n 点法在不同数据集上进行了测试比较，主要比较指标为准确度和计算速度，其中五点法和 n 点法的归一化匹配点对由不同的特征点以及光流法提供（3.3.3节、3.3.4节）。最后，在3.3.5节中对算法的两个步骤（运动矢量提取和旋转估计）进行了不同方法的运行时间分析。

实验中使用的数据集涵盖从室内到室外场景，其中包含一个具有模糊特征

[1] https://github.com/laurentkneip/opengv

的数据集。除了公开数据集 *OVMIS*[5] 和 *CVLIBS*[3,4] 外，作者还使用廉价全景相机采集了图像，以增加相机运动的丰富度，廉价全景相机由低成本全镜头（Kogeto Dot Lens）和手机（Oneplus 5）上的相机组成。作者采集图像时场景与公开数据集场景相似，主要考虑图像分别率以及相机质量对图像以及姿态估计算法的影响。由于这些公开数据集通常是使用高端传感器进行采集的，因此图像具有高分辨率，相较而言，廉价全景相机采集的图像分辨率较低且质量较差（请参见图 3.3）。

在作者采集数据集时，将手机固定在三脚架上，然后旋转三脚架上的万向球来获取图像序列。从图 3.3中可以看到，手机及其支架在全景图像中占据相当大的空间，在进行相机姿态估计时，这些部分可能会导致算法认为相机未发生运动。尽管如此，后续的实验将说明针对这种欺骗性数据，所提出的基于正弦曲线拟合的方法依然很鲁棒。

表3.3中总结了本节实验中用到的数据集，以及它们是在哪一小节被用到的，不同数据集获得真值的方式也展示在了该表格中。在自己采集的数据集中，手机的 IMU 被当做作旋转的真值来源。为了确认手机 IMU 数值的可靠性，我们在室内比较了 OptiTrack 跟踪系统的测量值与 IMU 的测量值，发现两者相差小于 0.5°，因此手机 IMU 的测量值可以用作评估真值。图3.3和3.4分别给出了作者采集的数据集和公开数据集的图像示例。

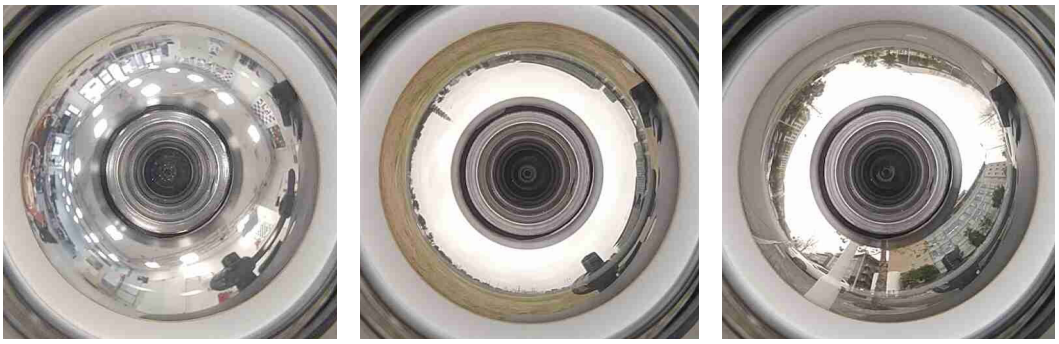
在接下来的实验中（3.3.2节除外），横滚、俯仰和偏航角的均方误差 (RMSE) 被作为评估标准，表征真值和估计值之间的差异。在计算均方误差时，首先将真值和估计值的欧拉角表示转换为旋转矩阵，然后使用旋转矩阵来计算相对变换，接着将相对旋转矩阵转化成轴角的表示形式，其中角度的大小记为误差，用来表征算法的性能。对于 VO 而言，实时误差过大可能会导致 VO 系统不能正常工作，因此绝对的定位精度比帧与帧之间的相对误差更有意义，本章实验中使用绝对 RMSE 用作评估标准，而不是相对 RMSE。只有在3.3.2节中，使用了相对误差，因为在该小节中主要考虑的是帧与帧之间的旋转估计，在这种情况下绝对误差和相对误差之间没有差异，更多细节将在3.3.2节中详细说明。

此外，本实验还对比了本章提出的正弦曲线拟合方法和几何法。按照本章前述介绍的，正弦曲线拟合方法所需的运动向量由光流法或 FMT 算法提供；几何法包括 n 点法和五点法，其匹配点对可以由 ORB、AKAZE、光流法和 FMT 实现。

表 3.1 数据集概览

Table 3.1 Overview of datasets

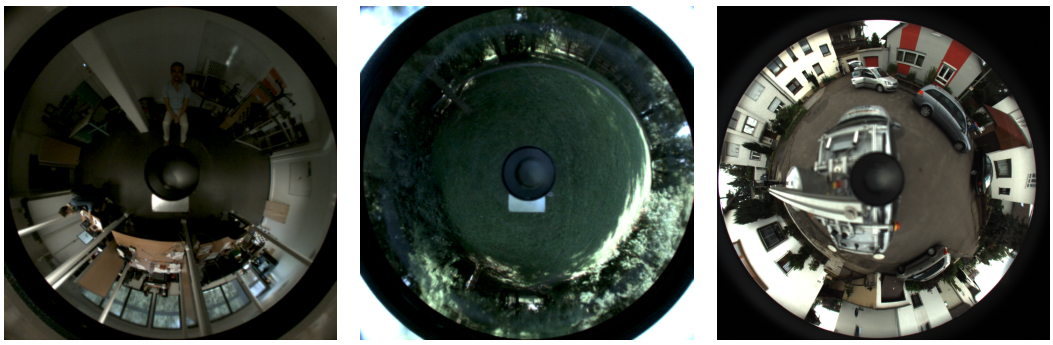
场景	运动	名字	图像帧数	使用小节	真值
作者采集的					
室内	横滚	indoor_single_roll	155	3.3.3 单一旋转下的算法性能评估	手机 IMU
草地	横滚	grass_single_roll	158	3.3.3 单一旋转下的算法性能评估	手机 IMU
街道	横滚	street_single_roll	160	3.3.3 单一旋转下的算法性能评估	手机 IMU
室内	俯仰	indoor_single_pitch	116	3.3.3 单一旋转下的算法性能评估	手机 IMU
草地	俯仰	grass_single_pitch	178	3.3.3 单一旋转下的算法性能评估	手机 IMU
街道	俯仰	street_single_pitch	174	3.3.3 单一旋转下的算法性能评估	手机 IMU
室内	偏航	indoor_single_yaw	200	3.3.3 单一旋转下的算法性能评估	手机 IMU
草地	偏航	grass_single_yaw	200	3.3.3 单一旋转下的算法性能评估	手机 IMU
街道	偏航	street_single_yaw	200	3.3.3 单一旋转下的算法性能评估	手机 IMU
室内	横滚 & 俯仰 & 偏航	indoor_rpy	200	3.3.1 关于使用图像旋转进行联合优化的评估 3.3.4 混合旋转下的算法性能评估	手机 IMU
草地	横滚 & 俯仰 & 偏航	grass_rpy	183	3.3.1 关于使用图像旋转进行联合优化的评估 3.3.4 混合旋转下的算法性能评估	手机 IMU
街道	横滚 & 俯仰 & 偏航	street_rpy	191	3.3.1 关于使用图像旋转进行联合优化的评估 3.3.4 混合旋转下的算法性能评估	手机 IMU
室内	横滚 & 俯仰 & 偏航 & z	office_zrpy	11	3.3.2 相机发生平移时的鲁棒性测试	手机 IMU
OVMIS[5]					
室内	横滚 & 俯仰 & 偏航	OVMIS_1	160	3.3.4 混合旋转下的算法性能评估	Robotic Platform (ARIA Library)
草地	yaw	OVMIS_2	156	3.3.4 混合旋转下的算法性能评估	Robotic Platform (ARIA Library)
草地	偏航 & x & y	OVMIS_3	200	3.3.2 相机发生平移时的鲁棒性测试	Robotic Platform (ARIA Library)
CVLIBS[3,4]					
街道	偏航 & x & y	CVLIBS	200	3.3.4混合旋转下的算法性能评估	IMU/GPS



(a) 室内　　(b) 草地　　(c) 街道

图 3.3 由 Oneplus 5 手机采集的图像示例

Figure 3.3 Image examples captured by Oneplus 5



(a) *OVMIS*_1　　(b) *OVMIS*_2　　(c) *CVLIBS*

图 3.4 公开数据集中的图像示例

Figure 3.4 Image examples from public datasets

ORB 和 AKAZE 是具有代表性的特征，ORB 是最快的检测器之一，而 AKAZE 则被设计用于检测非线性空间中的特征[126]。基于 FMT 的五点法已在第2章中详细介绍过，此处不再赘述。

由于 FMT 只能用于透射投影的二维图像配准，无法直接用于全向图像，因此不管是正弦曲线拟合还是几何法，如果需要用 FMT 来计算运动向量，则需要先将全向图像转换为全景图像，然后从全景图像中划分子图，再利用 FMT 计算不同列的运动向量，从而得到全景图像帧间的 Δu, Δv 和 $\Delta\theta$。类似地，在测试基于光流法的正弦曲线拟合时，利用光流法计算像素位移也是在全景图像上进行的。此外，其他的算法都是在原始全向图像上计算的，以避免图像变换产生额外误差。

本章所有的实验计算均在 Intel Core i7-4790 CPU、16 GB 内存的 PC 上进行，

算法用 C++ 单线程实现。

3.3.1　关于使用图像旋转进行联合优化的评估

本章在进行正弦曲线拟合的过程中，如果前续计算图像运动向量时采用的是 FMT，则在拟合时使用 3DoF 的运动向量 $(\Delta u, \Delta v, \Delta\theta)$。本节主要研究考虑利用 $\Delta\theta$ 进行联合优化时对姿态估计产生的影响。正如在3.1.4节中所分析的，与像素偏差类似，滑动窗口得到的每个子图的旋转与其列索引的关系也遵循正弦曲线，因而这两项可以联合优化。在该实验中，公式(3.15)中联合优化参数 $\Delta\theta$ 的权重系数 η 设置为 0.1，在使用了 $\Delta\theta$ 的所有其他实验中也会使用此值。

当相机朝向 Θ 变化时，子图对应的的 $\Delta\theta$ 遵循正弦曲线变换，因此所提出的算法将其包含在联合优化中。

具体来说，本实验包含两种不同的设置：在正弦拟合中使用和不使用旋转 $\Delta\theta$ 进行联合优化，即公式(3.15d)中是否设置 $\eta \neq 0$。而且每种设置都可以考虑是否加入平移项，3.3.2节中还将专门讨论平移项的问题。也就是，本节中共比较四种不同的方法, 分别为 `ours_fmt_wotrans_wrot`, `ours_fmt_wotrans_worot`, `ours_fmt_wtrans_wrot`, `ours_fmt_wtrans_worot`。`worot` 代表未使用子图的旋转进行联合优化，`wrot` 表示使用子图的旋转进行联合优化，`wotrans` 代表不考虑平移项，`wtrans` 表示使用了平移项。关于是否需要使用平移项将在3.3.2中进行讨论。本节的实验不包括光流法，因为它无法直接估计图像旋转。

表 3.2 使用/不使用图像旋转进行联合优化的姿态估计 RMSE (rad)

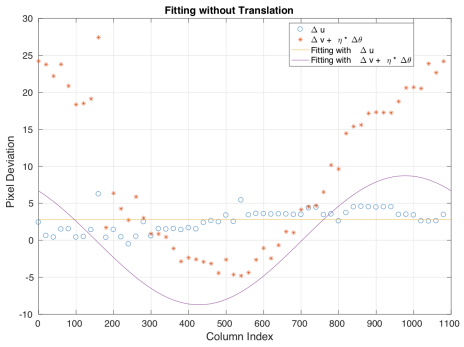
Table 3.2 RMSE (rad) of with/without rotation for joint optimization

	indoor_rpy	grass_rpy	street_rpy
`ours_fmt_wotrans_wrot`	**0.212**	**0.314**	**0.499**
`ours_fmt_wotrans_worot`	0.214	0.336	0.522
`ours_fmt_wtrans_wrot`	**0.204**	**0.567**	**0.504**
`ours_fmt_wtrans_worot`	0.217	0.592	0.533

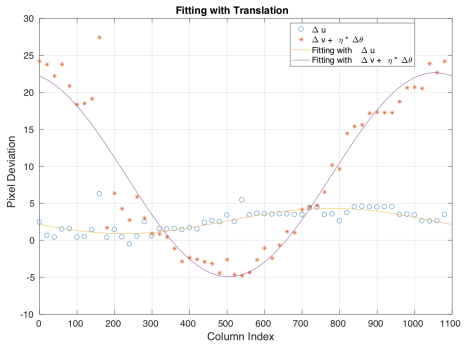
所有方法都在三个数据集上进行了测试：indoor_rpy, grass_rpy and street_rpy，实验结果见表 3.2。从表中可以发现，使用旋转进行联合优化可以在一定程度上提升性能。因此，以下各节中仅使用联合优化方法 `ours_fmt_wtrans_wrot` 和

(a) 输入图像 1

(b) 输入图像 2



(c) 不使用平移项拟合



(d) 使用平移项拟合

图 3.5 在 office_zrpy 数据集上进行使用/不使用平移项的正弦曲线拟合

Figure 3.5 Sinusoid fitting with/without translation on the office_zrpy dataset

`ours_fmt_wotrans_wrot` 进行比较。

3.3.2 相机发生平移时的鲁棒性测试

3.1.2节中介绍的旋转估计模型中使用了平移项。本节主要展示了关于加入平移项进行拟合的实验。直觉上来说，当 $z-$ 轴上的平移、横滚和同时发生时（见公式(3.2a)），如果拟合时去除平移项 λt_z，拟合结果会出现很大的误差。为验证该推测，首先进行了有无平移项的简单实验，如图3.5a 和3.5b所示，两帧图像对应的相机变换为旋转与沿 z 轴方向的平移，图3.5d和3.5c分别表示有无平移项的拟合结果。从图3.5c中可以看出，在没有平移项的情况下进行曲线拟合时，紫色曲线不能很好地拟合出位移 $\Delta v + \eta\Delta\theta$（红色点）；相较之下，在图3.5d中，使用平移项进行拟合时，紫色曲线能够更好的拟合位移 $\Delta v + \eta\Delta\theta$。因此当有较大的平移、旋转时，使用平移项进行拟合比不使用平移项具有更好的性能。

基于这样的初步测试，本节使用 office_zrpy 数据集中的 11 帧图像进行了一个小实验，定量地比较在存在较大平移和旋转时，本章所提出的正弦曲线拟合算法、n 点法和五点算法的性能。以其中一帧图像作为参考帧，其余图像均与参考

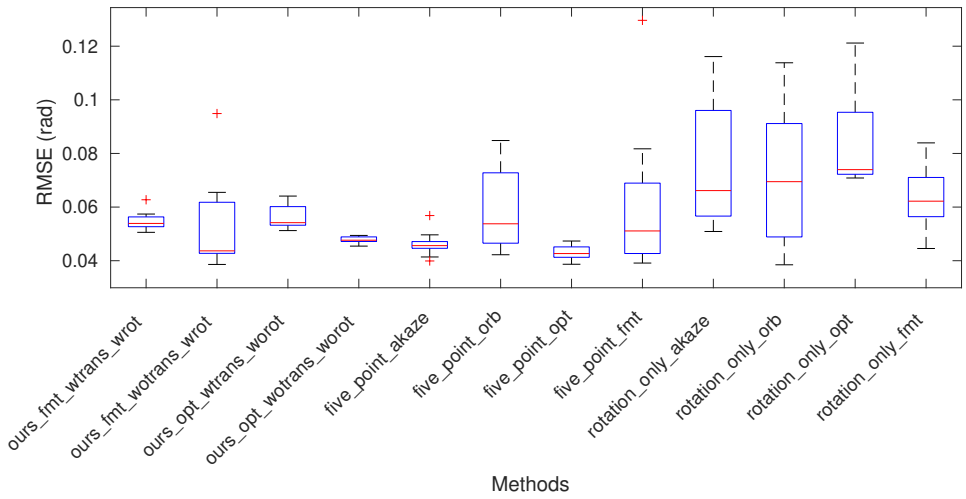


图 3.6 在 office_zrpy 数据集上不同算法进行姿态估计的结果

Figure 3.6 Performance of rotation estimation testing on the office_zrpy dataset

图像进行配准，每帧图像和参考图像之间的相机运动包括旋转和 z 轴方向上的平移。对于正弦曲线拟合方法，本节的实验中将包含分别使用 FMT 和光流法计算图像运动；对于基于几何的两种方法：n 点法和五点法，本节实验中也将涵盖采用不同的方法寻找匹配点对。对这两种几何方法，n 点法通常用于没有平移或平移很小的情况下的旋转估计，而在既有旋转又有平移的情况下，一般使用五点算法。

图3.6用箱形图展示了相关的实验结果，每个蓝色箱格的上边界和下边界分别为上下四分位数，中间的红色线代表了中值，最大最小值分别由上下的黑色短线表示，离群点由红色 + 表示。从图中的中值（中间的红色线）可以发现，在相机既发生旋转又发生平移运动时，五点法的表现最好，其次是我们提出的方法，n 点法表现得最差。

图3.6中还展示了一个比较有趣的结果，比较好的拟合效果不一定能保证更好的估计结果。比如基于光流法的正弦曲线拟合中“不加入平移项”的设置 `ours_opt_wotrans_worot` 表现得比“加入平移项”的设置 `ours_opt_wtrans_worot` 更好，但前者的拟合结果并不佳。一个可能的原因是两帧之间相机的旋转角度非常小，这种情况下“不加入平移项”可能会比“加入平移项”的效果更好。需要注意的是，本章算法基于的模型假设是旋转角应该小于 5°，当角度过大时，实验将变得不可控。不过在相机的连续运动中，可认为该假设是成立的。

综上而言，和其他几何法相比，本章所提出的正弦曲线拟合方法不管是否加入平移项，都能取得不错的效果。对于几何法而言，在相机有平移的情况下，五点法的表现更好。由于在接下来的几节（3.3.3和3.3.4节）中用到的数据集都不是纯旋转，可能有一些平移，因此这四种方法都会被评估。

3.3.3 单一旋转下的算法性能评估

本节主要评估单一旋转下各种算法的性能，即相机仅绕一个轴（x、y 或 z 轴）旋转。在进行下一节较为复杂的混合旋转估计之前，本节先进行各种算法的基本性能测试。本节中要评估的方法与3.3.2节中相同，共十二种方法。12 种方法方法分别是：使用 AKAZE，ORB，光流法和 FMT 的 n 点法（图3.8）中的前四行），使用这四个特征算子的五点法（中间四行）以及使用 FMT 和光流法的正弦曲线拟合方法（最后四行）。如表3.1所示，这些实验是在由配有全景镜头的手机采集的 9 个数据集上进行的，包括室内、草地和街道场景。

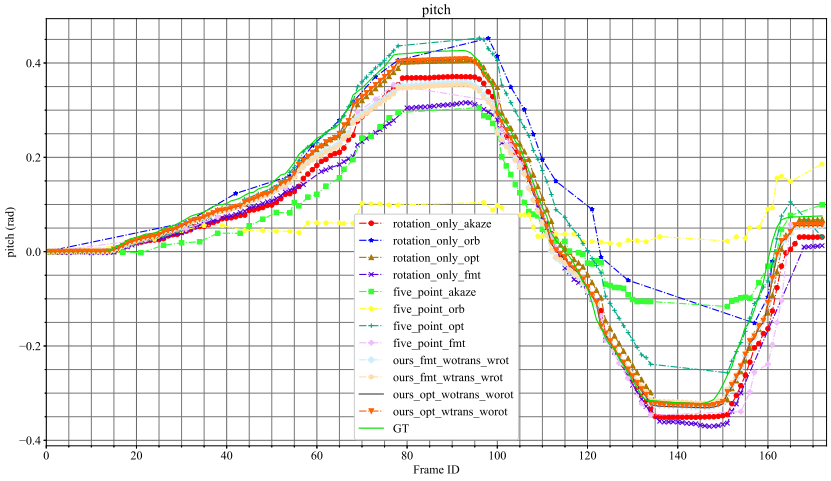


图 3.7 在 street_single_pitch 数据集上进行单一旋转估计的定性评估结果示例

Figure 3.7 Qualitative results for single rotation estimation on the street_single_pitch dataset

图3.7定性地展示了不同算法在 street_single_pitch 数据集上的旋转估计结果，可以看到，几乎所有方法估算的旋转都接近真值（绿色实心）。其他数据集的定性结果在附录中展示。

图3.8定量地展示了在手机采集的数据集上不同算法估计的单一旋转的结果。颜色的强度表示误差的大小：颜色越深，误差越大。正如3.3节一开始提到的，跟踪过程可视为 VO 系统，因此在这里使用绝对 RMSE 而不是相对 RMSE 作为评

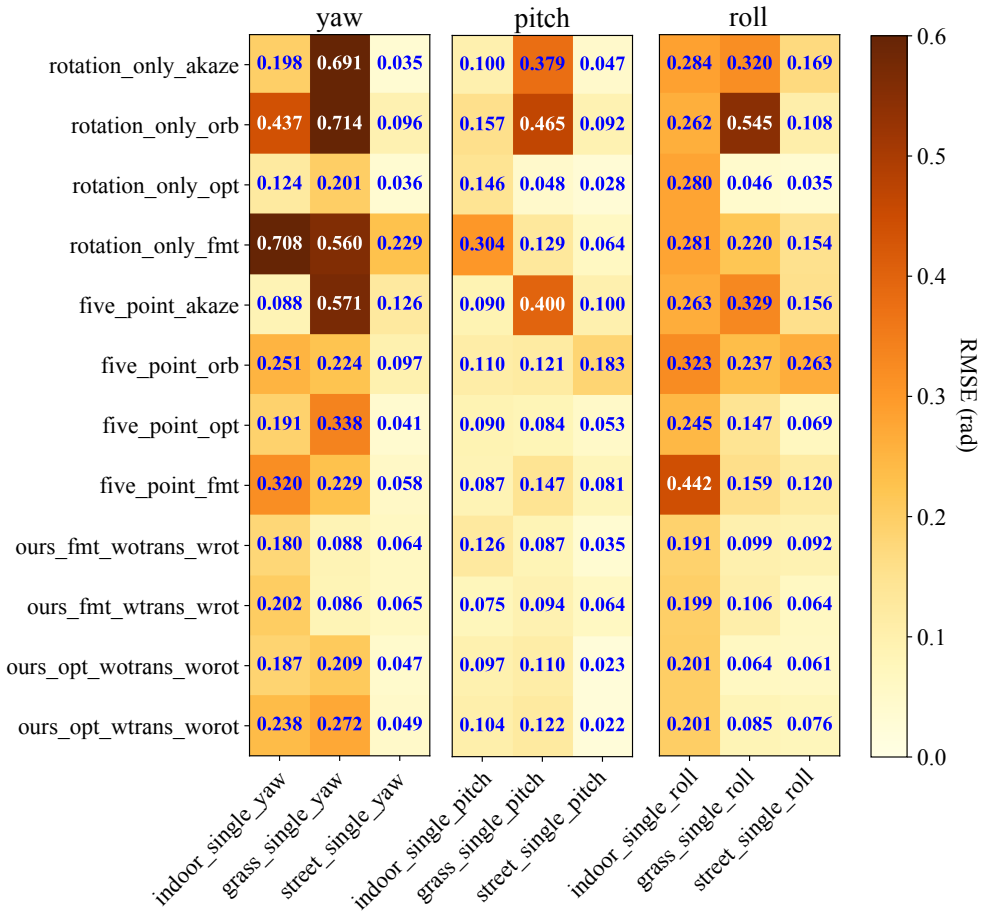


图 3.8 不同数据集上单一旋转的定量评估结果

Figure 3.8 Quantitative results for single rotation estimation on different datasets

估标准。因而尽管图3.9中显示的 RMSE 值看起来很大，但实际上仍然在合理范围内。在这里给出一个最坏情况的简单示例：如果每帧的估计误差为 $+1°$，并且总共有 100 帧，则累积误差将为 $5050°$，因此平均误差约为 $50°$，即 $0.87rad$。相比之，图3.9中的 RMSE 并不是很大。因此在进行不同算法的比较时，只有 RMSE 的相对大小需要被关注，而不是值本身。3.3.4节中也采用了相同的评估方式。

从图3.8中可以看到，几乎每种方法在街道场景中都能获得较好的结果。这种场景下，图像纹理通常包含丰富的特征，这有助于找到匹配点对，因此正弦曲线拟合法和基于几何的方法都可以很好地用于旋转估计。室内场景也能提供了很多特性，因此所有方法在其相关俯仰和横滚的数据集上都取得了很好的效果。但是，在草坪场景时，特征非常相似，有些特征无法被正确匹配，尤其是对于 ORB、AKAZE 和光流法而言，性能差异就会体现出来。例如，在 grass_yaw 数据集上，基于 AKAZE 的几何方法的误差很大，`rotation_only_orb` 的误差也很大。从图3.8的前四行中，可以看到 `rotation_only_opt` 是 n 点法中鲁棒性最好、给出的旋转估计结果精确度最高的，从中间四行中可看出 `five_point_opt` 是五点算法中最好的。在不同设置下，基于 FMT 和光流法的正弦曲线拟合法具有相似的性能。而图3.8中的浅色分布表示，正弦曲线拟合的性能一般优于几何方法。

3.3.4 混合旋转下的算法性能评估

本节将进行相机绕多个轴旋转下的算法性能评估，在多个不同的数据集上进行了测试，包括使用手机采集的数据集：indoor_rpy、grass_rpy 和 street_rpy，两个公开数据集：*OVMIS*[5] 和 *CVLIBS*[3,4]，其中 *OVMIS* 提供了室内（*OVMIS*_1）和草坪（*OVMIS*_2）场景，而 *CVLIBS* 的图像是在街道场景采集的。手机采集的数据集和公开数据集都涵盖了室内、草地和街道场景。

图3.9展示了手机采集的（前三列）和公开数据集（后三列）上不同算法进行旋转估计的结果，较大的 RMSE 用深色表示。在手机采集的数据集上，所有方法在室内场景中均表现更佳，而在草地和街道数据集上的它们的性能下降了。通过比较不同的算法，发现 `five_point_orb` 和 `rotation_only_fmt` 表现最差。公开数据集中的图像都是高质量的全向图像，因此所有方法均取得了良好的效果，而所提出的正弦曲线拟合法（最后六行）更加可靠。通过横向比较可以发现，每种方法在公开数据集上的性能要比在手机采集的数据集上表现更好。原因有两

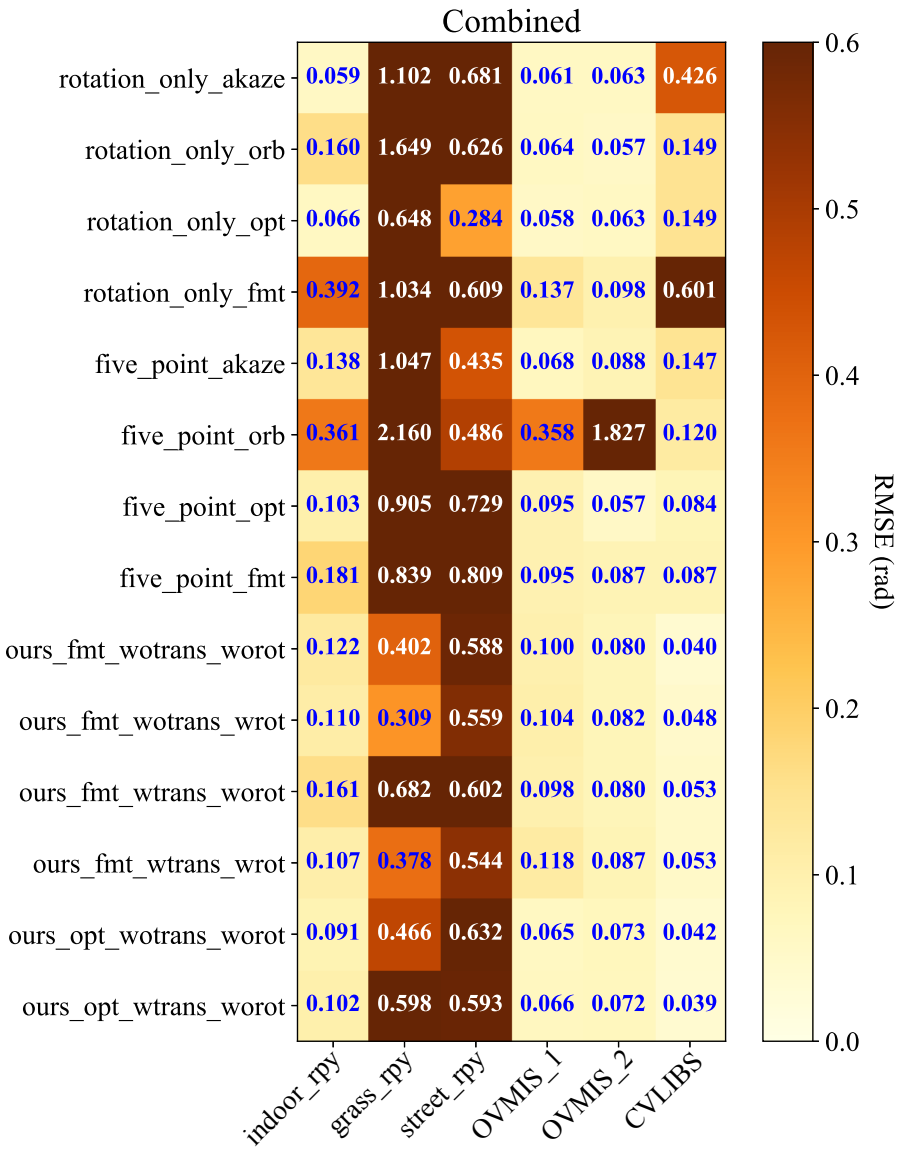


图 3.9 多个数据集上不同算法进行混合旋转估计的结果

Figure 3.9 Different methods evaluation on multiple datasets

个：1）用来采集数据的手机相机是卷帘式快门，且分辨率低；2）廉价的全向镜头生产工艺粗糙。

考虑到在所有数据集上的整体性能，本文所提出的正弦曲线拟合方法（最后六行）在准确性和鲁棒性上表现更好。尽管正弦曲线拟合并不总是最好的，但在不同设置下，除了 street_rpy 场景之外，它在所有数据集上都取得了很好的结果。而几何方法则依赖配对情况，偶尔会出现跟踪旋转失败的情况，例如 grass_rpy 数据集上的 `five_point_orb`。当都使用光流法来实现正弦拟合和几何法时，我们可以发现 `rotation_only_opt` 性能最好，`ours_opt_wotrans_worot` 其次。此外，当两者都使用 FMT 时，`ours_fmt_wotrans_wrot` 性能最好。

3.3.5 运行时间分析

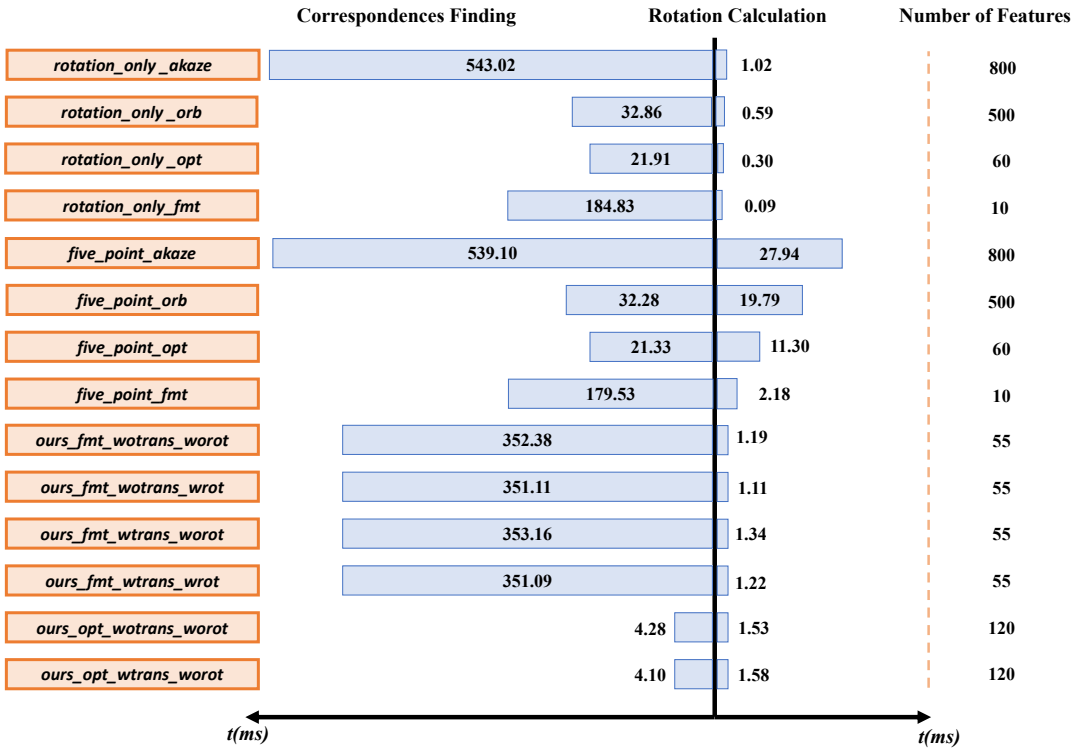


图 3.10 每帧的平均运行时间分析

Figure 3.10 Run-time analysis per frame

为了分析哪个步骤需要更长的运行时间，本节分别测试了旋转估计的两个步骤的运行时间：点对匹配和旋转估计。第一步可分为两类：AKAZE、ORB 和光流法的特征点匹配和 FMT 算法的运动向量计算（包括像素偏差 $\Delta u, \Delta v$ 和旋转 $\Delta\theta$）；第二步，分别是利用 n 点法、五点法以及正弦拟合法进行相机朝向 Θ

估计。

图3.10显示了两个步骤的时间消耗。每种方法在寻找匹配点上的耗时比旋转估计更长。从图3.10的左侧，可以看到：光流法和 ORB 比 FMT 和 AKAZE 更快。在本实验的设置下，基于 FMT 的正弦拟合的点对匹配数量是固定的；而几何方法的点对匹配数量除了 five_point_fmt，都取决于图像的外观丰富度；基于光流法的正弦曲线拟合的点对匹配数量也不是固定的，因为光流法同时跟踪了子图的中心点和特征提取的特征点，因此也与场景有关。值得注意的是，所提出的方法中使用的光流法比几何方法中的光流法快，原因是几何法中的光流法是在原始图像上计算光流，而正弦曲线拟合中用到的光流法是在全景图像上计算的。当然，后续可以通过图像掩模来提升几何法中光流法的运行速度。通过比较 five_point_fmt 和使用 FMT（前四行）的正弦拟合法，可以发现正弦拟合法比五点法更快。总得来说，不考虑寻找匹配点对的过程，n 点法是最快的，其次是正弦曲线拟合。

3.4　小结

本章提出了一种基于正弦拟合的旋转估计方法，它利用全景图像中的像素位移与旋转来估计全向相机的姿态。该方法支持使用 FMT 和光流法来计算像素偏差，其中，FMT 可以计算两个子图之间的旋转 $\Delta\theta$，计算结果可用于联合优化，而光流法只能计算像素位移。这些运动向量被拟合为两条正弦曲线，用来估计相机的旋转。

通过大量的实验比较了所提出的正弦曲线拟合方法和基于特征的几何方法：n 点法和五点法。不同数据集上的实验展示了正弦曲线拟合算法的鲁棒性，其主要归功于点对匹配和正弦曲线拟合。点对匹配会删除不匹配的特征，正弦曲线拟合可以过滤像素偏差较大的异常值。与第2章的基于 FMT 的五点法相比，本章的正弦曲线拟合算法在旋转估计中的准确度更高。此外，实验中还采用了其他的方法估计子图运动，并与 n 点法进行了对比，结果表明正弦拟合方法与基于几何法的方法一样好，且在分辨率低、场景难度较大的图像（例如草地）上表现出更好的性能。

不同方法的运行时间分析表明，点对匹配会占用大部分的运行时间，其中光流法最快；正弦曲线拟合和基于几何的方法都可以非常快速地进行旋转计算。

基于光流法的正弦拟合方法在所有测试的算法中具有最快的运行时间，同时在准确性和鲁棒性方面也显示出很好的结果。

该方法的缺点是无法进行平移估计，因此无法实现 6DoF 的位姿估计。并且由于 FMT 只能在单一深度下正常工作，在多深度场景下不能准确估计子图的平移，这也会给 6DoF 的位姿估计带来麻烦。在未来的工作中，本文作者计划解决该方法无法进行平移估计的问题。为此，本文将首先在第4章中讨论如何拓展FMT，使其可以估计多深度场景下的小孔相机运动。

第 4 章　多深度场景下的傅里叶梅林变换

第2和3章都利用 FMT 进行了运动向量的估计，但是也都遇到了 FMT 无法应用于多深度场景的问题。第2章提出递归划分子图的策略来避开该问题，第3章中则在此基础上利用拟合算法的鲁棒性来剔除离群点。本章从 FMT 本身出发，提出扩展傅里叶梅林算法 (extended Fourier-Mellin transform, eFMT)，用于放宽单一深度场景的限制。如果场景中物体的深度不同，那么尽管相机的运动是相同的，由于透射投影的影响对应的像素运动也是不同的。FMT 估算出来的运动是视野中大部分深度对应的运动，因此如果大部分深度改变了，那么相机的运动速度就不能从 FMT 的估算结果中正确推算出了，因此基于 FMT 的 VO 不能在多深度场景中给出正确结果。

本章提出的 eFMT 算法在不改变 FMT 旋转估计的基础上，扩展了 3D 平移估计（缩放和平行于成像平面的平移），主要考虑的相机模型为小孔相机。在本章的4.1节中将会介绍相机的旋转对深度不敏感，但是多深度场景将会导致多种缩放和平移的发生。由于 FMT 已经成功应用于多种场景，如遥感、图像批准、定位与建图等，因此 eFMT 存在着巨大的应用前景，它可以进一步拓展 FMT 的应用场景。本章的实验部分也将介绍 eFMT 的一个应用示例：用 eFMT 来估计无人机的运动轨迹。无人机上装载有一朝向向下的针孔相机，无人机从校园上空飞过时采集图像，然后用 eFMT 来估计无人机的轨迹。采集的场景中包括建筑物、树木、河流等高度不同的物体，用以测试 eFMT 在多深度场景中的性能。此外，实验中还将对比当下流行的基于特征点法和直接法的 VO 框架，用来评估 eFMT 的鲁棒性和准确性。

本章的主要贡献总结如下：

- 首次提出将 FMT 中的缩放和平移估计拓展到多深度场景中；eFMT 比 FMT 更加通用但却保持了 FMT 鲁棒的优点；
- 实现了一个基于 eFMT 的 VO 框架；
- 在多深度场景下，评估 eFMT、FMT 和[134] 中总结的几种代表性 VO 算法的性能，包括 ORB-SLAM3[86]、SVO[88] 和 DSO[135]。

4.1 问题描述

本节主要对多深度场景下、相机的 4DoF 运动导致的图像变换进行了建模。

给定图像 ${}^1I$ 上的一点 $p=[x,y]^\top$，其可以利用相机内参进行归一化

$$\bar{p}=\begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}^{-1}\begin{bmatrix} x \\ y \\ 1 \end{bmatrix}=\begin{bmatrix} \frac{1}{f_x}(x-c_x) \\ \frac{1}{f_y}(y-c_y) \\ 1 \end{bmatrix} \tag{4.1}$$

其中，f_x、f_y 为相机的焦距，(c_x,c_y) 是图像的中心坐标。假设该像素 p 对应的 3D 点为 P，其深度为 δ，则在 ${}^1I$ 的坐标系下，点 P 的坐标可表示为

$$P=\begin{bmatrix} \frac{\delta}{f_x}(x-c_x) \\ \frac{\delta}{f_y}(y-c_y) \\ \delta \end{bmatrix}. \tag{4.2}$$

假设图像 ${}^1I$ 和 ${}^2I$ 对应的相机位姿之间的变换为 4DoF 的运动，即绕着相机主轴的旋转——偏航角 θ、在成像平面上的 2D 平移 $(\Delta x,\Delta y)$ 和垂直于成像平面的平移 $\Delta\delta$，则点 P 投影在 ${}^2I$ 的点 p' 上，其坐标为：

$$\begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}P+\begin{bmatrix} \Delta x \\ \Delta y \\ \Delta\delta \end{bmatrix}, \tag{4.3a}$$

即

$$p'=\begin{bmatrix} \frac{\delta}{\delta+\Delta\delta}(x\cos\theta-y\sin\theta) \\ +\frac{1}{\delta+\Delta\delta}(-\delta c_x\cos\theta+\delta c_y\sin\theta+f_x\Delta x)+c_x \\ \\ \frac{\delta}{\delta+\Delta\delta}(x\sin\theta+y\cos\theta) \\ +\frac{1}{\delta+\Delta\delta}(-\delta c_x\sin\theta-\delta c_y\cos\theta+f_y\Delta y)+c_y \end{bmatrix}. \tag{4.3b}$$

从而得到一个在 4DoF 下通用的方程

$$\begin{aligned} {}^2I(x,y) \quad &= {}^1I(z_\delta(x\cos\theta_0-y\sin\theta_0)+x_\delta, \\ &\qquad z_\delta(x\sin\theta_0+y\cos\theta_0)+y_\delta) \end{aligned} \tag{4.4}$$

用来描述 ${}^1I$ 和 ${}^2I$ 之间的像素变换，其中 $\theta_0=\theta$，

$$z_\delta=\frac{\delta}{\delta+\Delta\delta}, \tag{4.5}$$

$$x_\delta = \frac{1}{\delta + \Delta\delta}(-\delta c_x \cos\theta + \delta c_y \sin\theta + f_x \Delta x) + c_x \tag{4.6}$$

以及

$$y_\delta = \frac{1}{\delta + \Delta\delta}(-\delta c_x \sin\theta - \delta c_y \cos\theta + f_y \Delta y) + c_y \ . \tag{4.7}$$

从以上分析中，可以看到每个像素的缩放 z_δ 和平移 (x_δ, y_δ) 和其深度 δ 有关，而旋转 θ_0 则和深度无关。对多深度场景下的两帧图像 ${}^1I$ 和 ${}^2I$，公式(4.4)-(4.7)有多种不同的解，每种深度对应一种解，因此有多种缩放和平移。相移图中每个单元格的能量都和落在该单元格内的像素 (x_δ, y_δ) 个数成正相关，这些像素的深度均为 δ。由于 FMT 假设等距环境，即认为每个像素的深度 δ 相同，那么 FMT 假设所有像素 p 的平移 (x_δ, y_δ) 和缩放 z_δ 都是相同的。因此在 FMT 算法中所有的像素 (x_δ, y_δ) 都会落在相移图的同一个单元格中，形成了一个峰。

本章提出了 eFMT，放松了等距环境的限制，通过求解不同深度 z_δ 情况下的式(4.4)来估计相机位姿。

4.2　算法设计

本节首先分别讨论只有平移或者只有缩放下如何求解式(4.4)，然后再讨论如何处理 4DoF 运动下的一般场景。此外，由于单目相机无法恢复绝对尺度，因此本节中也将讨论如何保证平移和缩放的尺度一致性。

为了不失一般性，在本章中用帧号 1、2 和 3 代表任意连续的三帧。

4.2.1　纯平移场景

FMT 将平移的估计从旋转和缩放的计算中解耦出来了，因此可以先只考虑纯平移场景，即相机在 $x-y$ 平面（平行于成像平面）运动。此时，公式(4.4)可简化为

$${}^2I(x, y) = {}^1I(x + x_\delta, y + y_\delta) \ . \tag{4.8}$$

如公式(4.6)和(4.7)所示，由于多深度环境的原因，平移 (x_δ, y_δ) 不再对应单个能量峰。图4.1展示了多深度场景下的一个平移相移图示例，可以看到该相移图上有多个峰，在 $x-y$ 视角下可以看到这些高的峰落在同一条线上。该共线性质可以从 x_δ 和 y_δ 的定义中得出。在纯平移的情况下，公式(4.6)和(4.7) 可以简化为

$$x_\delta = \frac{f_x \Delta x}{\delta}, y_\delta = \frac{f_y \Delta y}{\delta} \ . \tag{4.9}$$

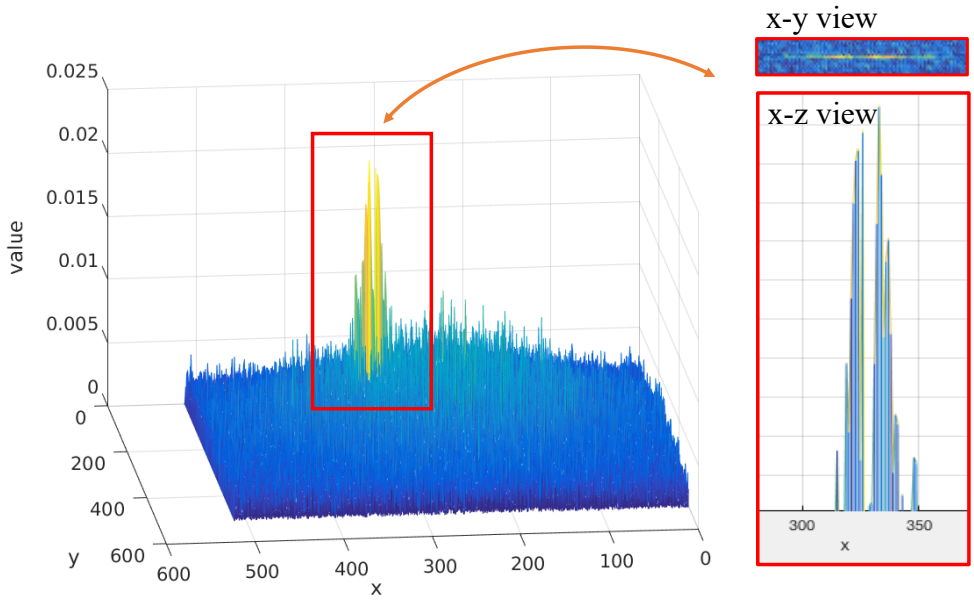


图 4.1 在多深度场景下的平移相移图示例

Figure 4.1 An example of a translation phase shift diagram in a multi-depth environment

可以看到，每个平移 (x_δ, y_δ) 向量的方向都是相同的，即

$$\left(\frac{f_x\Delta x}{\sqrt{(f_x\Delta x)^2+(f_y\Delta y)^2}}, \frac{f_y\Delta y}{\sqrt{(f_x\Delta x)^2+(f_y\Delta y)^2}}\right), \tag{4.10}$$

与像素深度 δ 无关的。此外，还可以发下这些平移 (x_δ, y_δ) 都落在一条线上：

$$f_y\Delta y \cdot x - f_x\Delta x \cdot y = 0\,. \tag{4.11}$$

也就是，这些能量较高的峰都落在一条通过相移图中心的线上。而且由于这些像素不会沿着相反方向运动，因此这些较高的能量峰可以看做落在从相移图中心出发的射线上。最极限的情况就是相机视野内有一倾斜平面，那么将没有可分辨的峰，这些高的能量值会是相移图上的一段连续线段。为了不失一般性并避免检测能量峰，本文将在考虑连续线段的基础上，提出多深度场景下平移估计的方法。

给定相机的运动，所有像素的平移运动是共线的，其方向独立于像素的深度，但是平移的大小与像素的深度、相机的运动有关。因而，不同于只在相移图上找到最高峰，本章将以一种全新的方式来进行平移估计。具体而言，相移图的中心点代表没有平移运动，从中心点出发，将相移图按角度大小均分成扇形等份，然后通过极搜索找到能量总和最大的扇形 r_{max}。该扇形的方向代表平移向

量的方向，缩写为 t。由于单目相机无法恢复绝对尺度，因此本章并不估计运动的大小，而是只估计平移向量的方向，并将其归一化为单位向量，在本文中称之为**单位平移向量**。

正如第1章中所介绍的，用 FMT 来实现 VO 有一弱点，因为 FMT 只计算两帧之间的运动，而不考虑如何将 ${}^1I$ 和 ${}^2I$ 之间的平移与 ${}^2I$ 和 ${}^3I$ 之间的平移放到同一尺度下，即 VO 所需的尺度一致性。为了克服这一弱点，eFMT 利用 r_{max} 扇形中包含的能量来计算尺度一致因子。为此，本节将从相移图的 r_{max} 扇形中采样得到平移能量向量 $\mathbb{V}_t$。给定相机平移的情况下，图像中不同深度的区域对应平移能量向量中的不同索引，且某个深度的像素越多，该深度对应的能量值越大。假设图像 ${}^1I$ 和 ${}^2I$ 之间的平移能量向量为 ${}^2_1\mathbb{V}_t$，图像 ${}^2I$ 和 ${}^3I$ 之间的平移能量向量为 ${}^3_2\mathbb{V}_t$。由于第二张图像 ${}^2I$ 在两者的平移估算中都被用到了，那么在两者的平移估计中，${}^2I$ 的像素深度是不变，对平移估计的贡献也是相同的，则平移能量向量 ${}^2_1\mathbb{V}_t$ 和 ${}^3_2\mathbb{V}_t$ 之间的不同都来自于平移大小的不同，和平移的方向无关。事实上，这两个向量之间就是简单的缩放关系，缩放系数为平移大小的比例，缩放之后平移能量向量中的能量大小还是和不同深度的图像区域相互对应的。基于此，重新缩放因子 ${}^{2\to1}_{3\to2}s_t$ 可以通过 ${}^2_1\mathbb{V}_t$ 和 ${}^3_2\mathbb{V}_t$ 之间的模式匹配得到，计算公式如下：

$$ {}^{2\to1}_{3\to2}s_t = \arg\min_s ||{}^2_1\mathbb{V}_t - f({}^3_2\mathbb{V}_t, s)||^2_2 , \tag{4.12} $$

其中，$f(\cdot)$ 是利用缩放因子 s 对向量 ${}^3_2\mathbb{V}_t$ 的长度和相应的能量进行缩放。详细算法将在4.3节中展示。

由于相机运动导致的视角变换，使得两帧图像有非重叠部分，从而导致相移图上的噪声变大。但正如文献[120]中所分析的，经典 FMT 对图像的重叠程度有要求，只要在该重叠要求内，非重叠部分导致的噪声就不会对平移估计有影响，可忽略之。

4.2.2　纯缩放场景

从公式(1.19)中可以看到，旋转和缩放共享一个相移图（见图4.2）。并且，公式(4.4)说明旋转是独立于像素深度的，所有像素的旋转都是相同的，从而在 eFMT 算法中，旋转的估计与 FMT 算法中采用的方法相同。本节只考虑纯缩放的情况，即相机垂直于成像平面发生平移，图像发生缩放。此时，公式(4.4)可简

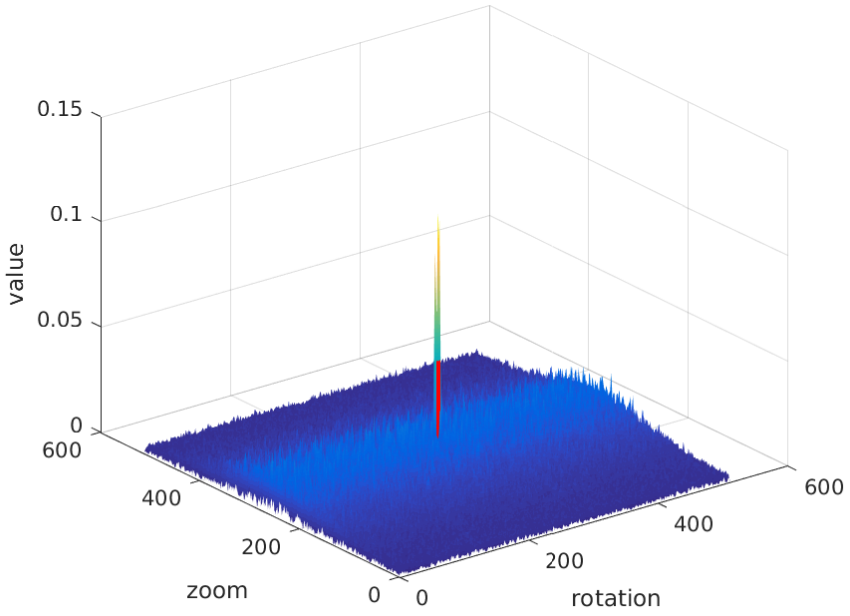


图 4.2 旋转和缩放的相移图示例

Figure 4.2 An example of rotation and zoom phase shift diagram

化为

$$^{2}I(x,y) = {}^{1}I(z_{\delta}x, z_{\delta}y)\ . \tag{4.13}$$

同时，公式(1.19)变成了

$$^{2}M(\xi,\theta) = {}^{1}M(\xi - d_{\delta},\theta)\,, \tag{4.14}$$

其中，$d_{\delta} = \log z_{\delta}$。可以看到，不同的深度会有不同的缩放，所有像素的旋转相同，即所有的缩放对应同一个旋转，即相移图上的同一列号，因此，这些缩放对应的峰都在旋转和缩放相移图上的某一列。需要注意的是，在实际应用中，有时像素深度是连续变化的，此时这些缩放对应的峰就会连在一起，变成了相移图上能量值较大的部分，而不再是单独的峰了。为此，本章不再检测多个峰值，而是首先找到相移图上能量总和最大的列 $\mathbb{C}_z^*$，然后从该列中找到最大的缩放 z_{max} 与最小的缩放 z_{min}，再从 z_{max} 与 z_{min} 之间均匀地采样出一个多种缩放值的集合 $\mathbb{Z} = \{z_{\delta}\}$。能量总和最大的列 $\mathbb{C}_z^*$ 的计算公式如下：

$$\mathbb{C}_z^* = \arg\max_{\mathbb{C}_z}\{\mathbb{C}_z \in q_z\}\,, \tag{4.15}$$

其中，q_z 是旋转与平移的相移图。另一方面，如4.1节中所推导的，缩放 z_{δ} 与像素的深度 δ 成反比例关系，如公式(4.5)所示。因此，从相移图中估算得到的最小

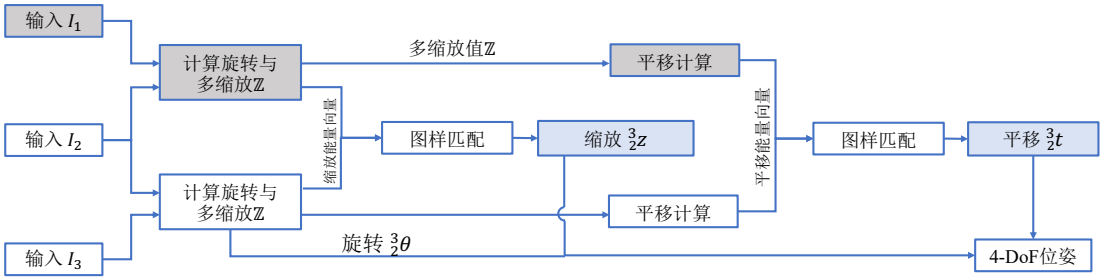


图 4.3 扩展傅里叶梅林算法的流程图

Figure 4.3 Pipeline of eFMT

和最大缩放，分别对应着最大和最小的像素深度。由于平移相移图中的能量大小也和像素深度有关，可以建立起缩放能量和平移能量之间的相应关系，相关内容将在4.2.3节中讨论。

此外，在纯缩放的场景中，保证缩放的尺度一致对 VO 也非常重要。为此首先从 $\mathbb{C}_z^*$ 中提取 $\mathbb{V}_z$，$\mathbb{V}_z$ 是 $\mathbb{C}_z^*$ 中能量更高的那一半。这主要是考虑到所有的像素要么缩小要么放大，因此其对应的能量应该只在相移图的上半部分或者下半部分。假设图像 ${}^1I$、${}^2I$、${}^3I$ 之间只有缩放，且 ${}^1I$ 和 ${}^2I$ 之间的缩放能量向量为 ${}^2_1\mathbb{V}_z$，${}^2I$ 和 ${}^3I$ 之间的缩放能量向量为 ${}^3_2\mathbb{V}_z$，则 ${}^2_1\mathbb{V}_z$ 和 ${}^3_2\mathbb{V}_z$ 之间的尺度缩放因子可通过下式计算得到：

$$ {}^{2\to1}_{3\to2}s_z = \arg\min_s ||{}^2_1\mathbb{V}_z - g({}^3_2\mathbb{V}_z, s)||_2^2 , \tag{4.16} $$

其中，$g(\cdot)$ 是将向量 ${}^3_2\mathbb{V}_z$ 进行平移的函数。它是计算平移尺度缩放因子的所用到的模式匹配的一个变种，两者的主要区别在于匹配平移能量向量用的是放缩该向量进行匹配，匹配缩放能量向量用的是通过平移该向量进行匹配。两者相应的算法都将在4.3中展示。

4.2.3　一般的 4DoF 运动情况

当相机发生 4DoF 运动时，本章所提出的 eFMT 算法估算相机位姿之间变换的流程和原来的 FMT 算法相似，eFMT 的流程如图4.3所示。由于基于单目相机的 VO 的平移需要是尺度一致的，因此本章中通过每三帧确定尺度一致的变换。

和 FMT 算法[136] 的流程类似，eFMT 首先计算两帧之间的旋转和缩放，不同于经典 FMT 只搜索旋转和缩放相移图中的最高峰，eFMT 用到了相移图上半列的所有信息，生成了多种缩放值 $\mathbb{Z}=\{z_\delta\}$ 和缩放能量向量 $\mathbb{V}_z$（见4.2.2节）。而且，eFMT 更多考虑的是能量而不是单一峰值，如多种缩放值 $\mathbb{Z}=\{z_\delta\}$ 是通过相

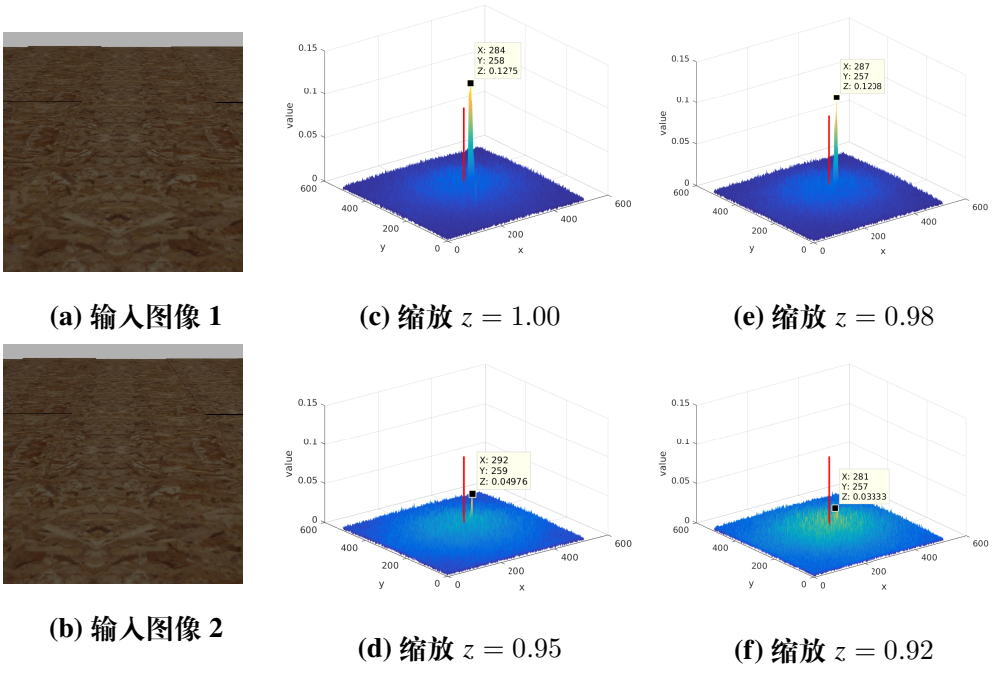
移图上的最小和最大缩放值均匀采样得到的，而不是直接从峰的位置计算得出的。从而使得该算法对相移图中的连续能量更为鲁棒。得到旋转 θ_0 和缩放 z_δ 之后，和 FMT 类似地，对第二帧图像进行逆旋转和逆缩放：

$$ {}^2I' = {}^2I(z_\delta x\cos\theta_0 - z_\delta y\sin\theta_0, z_\delta x\sin\theta_0 + z_\delta y\cos\theta_0)\,, \tag{4.17} $$

然后，利用公式(1.22)的方法在图像 ${}^1I$ 和 ${}^2I'$ 上进行相位相关计算，通过4.2.1节中介绍的方法，从平移相移图中提取平移能量向量 $\mathbb{V}_{t,z_\delta}$。然后将缩放的能量值作为权重，把多个平移能量向量组合起来，公式如下：

$$ {}^2_1\mathbb{V}_t = \sum_{z_\delta\in\mathbb{Z}} \frac{\mathbb{V}_z[z_\delta]}{U} * {}^2_1\mathbb{V}_{t,z_\delta}\,, \tag{4.18} $$

其中，$\mathbb{V}_z[\cdot]$ 的功能是找到缩放值 z_δ 对应的能量；$U = \sum_{z_\delta\in\mathbb{Z}}\mathbb{V}_z[z_\delta]$。缩放值的能量越大说明越多的像素对应于该缩放值，那么相应的平移能量向量应该获得更高的权重，因此公式(4.18) 的成立是合理的。



(a) 输入图像 1 **(c) 缩放 $z = 1.00$** **(e) 缩放 $z = 0.98$**

(b) 输入图像 2 **(d) 缩放 $z = 0.95$** **(f) 缩放 $z = 0.92$**

图 4.4 不同缩放值下的平移相移图

Figure 4.4 Translation PSDs with different zoom values

4.2.4 关于一般的 4DoF 运动的补充说明

经典的 FMT 可以简单地将旋转和缩放估计从平移估计中解耦出来，然而对于 eFMT 却没有这么简单，因为当相机沿着 z 轴（垂直于成像平面）运动时，不

同深度的物体其缩放程度不一样。在既有缩放又有平移的情况下，每个像素的视在运动和它的深度、缩放和平移都有关系。但是为了使平移能量向量的模式匹配（公式(4.12)）可以仅基于一个简单的尺度缩放因子，那么像素移动的能量只能和像素的深度以及平移的速度有关，需要独立于缩放。正如上文所述，eFMT会为每个缩放值都计算一个平移能量向量 $\mathbb{V}_{t,z_\delta}$，这说明在逆缩放第二帧图像时，多深度图像中总是有些部分没有被正确缩放，但是没有被正确缩放的部分却同样被用作了相位相关法（公式(1.22)）的输入，最终通过公式(4.18)合到平移能量向量中。

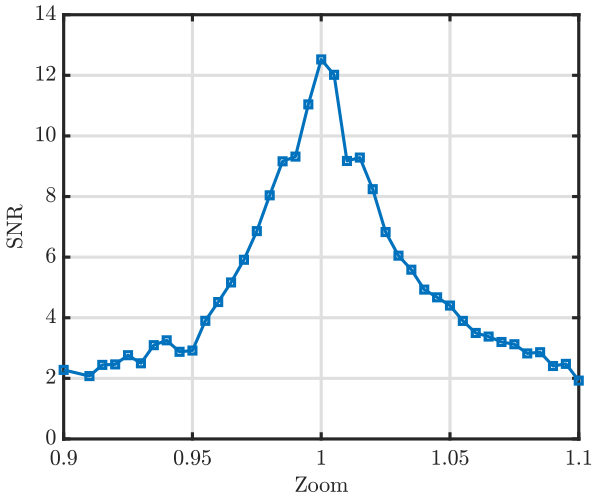


图 4.5 不同缩放值对应的信噪比（准确缩放值为 1）

Figure 4.5 Signal-to-Noise ratio with different zoom. (Correct zoom is 1.)

可能有读者会认为错误缩放的图像部分，会导致错误的像素平移估计，从而得到一个折中的平移能量向量，即峰的位置可能会改变。但是事实并不是这样的，相位相关法（公式(1.22)）对缩放较为敏感，即它只会挑选缩放相同的像素，形成相移图上能量较高的值，其他缩放不同的部分只是相移图上的噪声。这是因为如果缩放错误，公式(4.8)就不再成立。图4.4展示了错误的缩放会如何影响平移相移图。可以看到错误的缩放会降低正确平移的能量，将错误的能量分布到整张相移图上，并且如果缩放只是和正确的缩放有些许不同，平移相移图变化不大，而当缩放与正确的缩放值相差较大时，则会导致结果不对。为了更好地解释该现象，本文在图4.5中展示了当缩放值设置不同时，平移相移图的信噪比变化趋势。此处信噪比的计算为平移能量向量中能量较高的值的均值与相移图上其他值的均值之间的比。图4.5和4.4说明当缩放值的偏差达到 0.08 时就会导致信

噪比小于 2.6，此时相移图已经上的噪声已经很大了。在 eFMT 中，当对多个缩放值 $\mathbb{Z}$ 进行循环计算时，平移能量向量 $\mathbb{V}_{t,z_\delta}$ 只会挑选正确缩放的像素，从而组合得到正确的平移能量向量 $\mathbb{V}_t$。

4.2.5 视觉里程计中的实际考虑

本节中将 eFMT 应用到相机的位姿估计中，即视觉里程计，以此来展现 eFMT 相较于 FMT 的优势。对于 eFMT 和 FMT 而言，在 VO 中主要需要考虑的是如何将平移和缩放放到同一尺度下，即平移和缩放的一致性。

为此，本节再次分析了图像变换和相机运动之间的关系。如图4.6所示，假设相机 C 在 Pose 1 处，其视野里的第 i 个物体大小为 l_i，与相机的距离（即深度）为 δ_i，接着相机在 $x-y$ 平面内移动了 $(\Delta x, \Delta y)$，沿着 z 轴方向运动了 $\Delta\delta$，从而到达了 Pose 2。

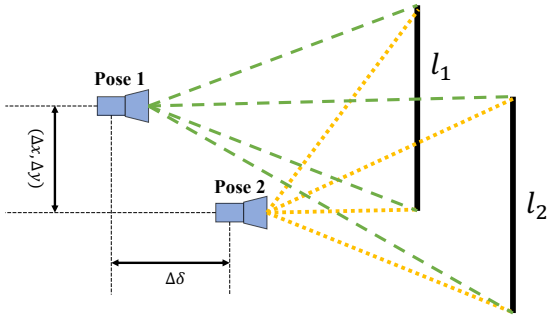


图 4.6 像素运动与物体深度的关系示意图

Figure 4.6 Objects of different depths in the FoV of the camera

根据小孔相机的特性，在 Pose 1 和 Pose 2 处采集到的图像 ${}^1I$ 和 ${}^2I$ 之间的缩放为 $z_{\delta_i} = \frac{\delta_i}{\delta_i+\Delta\delta}$。类似地，可以得到图像 ${}^1I$ 和 ${}^2I$ 之间不同深度 δ_j 的像素平移为 $(\frac{f\Delta u_j}{\delta_j}, \frac{f\Delta v_j}{\delta_j})$，其中 f 是相机的焦距。此处讨论平移时深度下标使用 j 而不是讨论缩放时的 i，主要是考虑到缩放和平移是独立计算的。通过上述分析，可以发现垂直于成像平面（沿 z 轴）的平移与平行于成像平面（$x-y$ 平面内）的平移的比例为：

$$\frac{(\frac{1}{z_{\delta_i}}-1)f}{\|(\Delta u_j, \Delta v_j)\|}, \tag{4.19}$$

当且仅当 $i=j$，说明在讨论相同深度的物体的平移与缩放，即 $\delta=\delta_i=\delta_j$。

为找到对应的 i 和 j，可以对缩放能量向量和平移能量向量进行模式匹配。本章中为了简化该步骤，采用了最大能量匹配，即先找到缩放能量向量 $\mathbb{V}_z$ 中对

应最大能量值的缩放 z_{peak}，假设其对应物体的大小为 l_i、深度为 δ_i，然后在对应于该缩放 z_{peak} 的平移能量向量中找到最大峰值对应的平移向量 $(\Delta u', \Delta v')$，对应于物体的大小为 l_j、深度为 δ_j，此处 $l_j = l_i$、$\delta_j = \delta_i$。上述关系对同样深度的所有像素都成立，并不需要这些像素在一连续平面上。

综上，可以得到相机位姿之间的 3D 平移 t：

$$t = \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta \delta \end{bmatrix} = \begin{bmatrix} \Delta u \\ \Delta v \\ \frac{(\frac{1}{z_{peak}} - 1) f}{\|(\Delta u', \Delta v')\|} \end{bmatrix}, \tag{4.20}$$

其中，$(\Delta u, \Delta v)$ 是单位平移向量。

4.2.6　关键点总结

eFMT 的关键点概述如下：

• 观察到多种深度的场景会导致缩放和平移相移图上有多个很强的能量值，并且这些能量在相移图上是共线的。

• 不同于经典 FMT 只在平移相移图上找到一个最大峰，eFMT 用一维平移能量向量来表示平移，该能量向量上能量的不同位置代表了对应于不同深度的像素的移动，能量值的大小代表了符合该平移（或该深度）的像素的数量。此外，eFMT 中还将平移的大小和方向分开考虑。平移的方向为采样得到的平移方向向量在平移图上的方向，以单位平移向量表示。缩放的表示也是类似的。因此，eFMT 保持了 FMT 相比于特征法和直接法的优势，即准确性和鲁棒性。

• 通过模式匹配找到对应的缩放和平移，然后将其尺度对应，使得缩放和平移在同一参考系下。

• 最后，对连续的三帧，通过平移能量向量之间的模式匹配找到后两帧之间的单位平移向量相对于前两帧之间的单位平移向量的比值，即尺度一致因子，将其乘到后两帧之间的单位平移向量上，以保证平移的尺度一致性。对于连续三帧之间的缩放尺度一致性也是类似操作。因此 eFMT 的尺度一致性比 FMT 更好。

4.3　算法实现

本节主要介绍了基于 eFMT 的 VO 算法的实现。本节首先给出了算法框架，然后讨论了平移和缩放的尺度一致性实现细节。

算法3展示了基于 eFMT 的 VO 算法的主要步骤。在第一二两帧的计算中，直接利用 FMT 计算出了两帧之间的旋转 θ_0、缩放 z 和平移，其中平移被归一化为单位平移向量 t。同时，从相应的相移图中，分别提取缩放能量向量 $\mathbb{V}_z$ 和平移能量向量 $\mathbb{V}_t$，以用作下一轮迭代中的模式匹配。从第三帧开始，eFMT 被用来计算缩放和平移的尺度一致因子，并将其应用的缩放和单位平移向量上，从而估计出每两帧之间的 4DoF 的运动。最后，通过链式法则生成相机的运动轨迹。

如4.2.1节所述，对于平移的计算，eFMT 不再搜索相移图上的最高峰，而是寻找相移图上能量总和最大的扇形区域 r_{max}。具体来说，先将相移图从图像中心 b 出发，分成 n 个等角度的扇形区域，扇形张开角度记为 o，在本章中 $o = 2°$；然后将扇形区域中每个单元格的能量相加，从而找到能量最大的区域 r_{max}；接着，从图像中心 b 到具有最大能量的扇形区域 r_{max} 中的最大能量值的方向被视作平移方向，即单位平移向量 ${}^{2}_{1}t_i$。此外，为用一维能量向量 ${}^{2}_{1}\mathbb{V}_{t,z_\delta}$ 来表示最大能量扇形 ${}^{2}_{1}r_{max}$ 中的值，在具体实现时从最大能量扇形 r_{max} 等距地采样能量值填充到 ${}^{2}_{1}\mathbb{V}_{t,z_\delta}$ 中，然后通过公式(4.18)将这些不同深度的平移能量向量合并为 ${}^{2}_{1}\mathbb{V}_t$。

算法4和5分别表示了计算平移和缩放的尺度一致性因子中用到的模式匹配算法。有多种方法可以实现模式匹配算法，比如相位相关法、搜索算法和动态规划等。考虑到针对相移图信号中外点的鲁棒性，本章采用一种搜索算法，具体实现如算法4和5所示。

算法 3 基于扩展傅里叶梅林变换的视觉里程计

输入： $\mathbb{I} = \{{}^{i}I | i \in \mathbb{N} \wedge 0 \leq i < \text{帧数}\}$

for i in $[1..len(\mathbb{I})]$ **do**

　　if $i = 1$ **then**　　▷ 和傅里叶梅林变换相似

　　　　估计旋转 ${}^{1}_{0}\theta_0$、缩放 ${}^{1}_{0}z$ 和平移 ${}^{1}_{0}t$

　　　　从旋转和平移的相移图中生成缩放能量向量 ${}^{1}_{0}\mathbb{V}_z$

　　　　从平移的相移图中生成平移能量向量 ${}^{1}_{0}\mathbb{V}_t$

　　else　　▷ 多个缩放与多个平移

　　　　计算 ${}^{i-1}I$ 和 ${}^{i}I$ 之间的旋转和缩放相移图

　　　　从相移图上估计旋转 ${}^{i}_{i-1}\theta_0$ 和存有多个缩放的向量 ${}^{i}_{i-1}\mathbb{Z}$

　　　　从相移图上生成 ${}^{i}_{i-1}\mathbb{V}_z$

　　　　for j in $[0..len({}^{i}_{i-1}\mathbb{Z})]$ **do**

　　　　　　计算平移能量向量 ${}^{i}_{i-1}\mathbb{V}_{t,j}$ 和单位平移向量 ${}^{i}_{i-1}t_j$

　　　　end for

　　　　通过公式(4.18)，组合平移能量向量得到 ${}^{i}_{i-1}\mathbb{V}_t$

　　　　估计 3D 平移，如4.2.5节所述

　　　　通过模式匹配计算 ${}^{i-1}_{i-2}\mathbb{V}_z$ 和 ${}^{i-1}_{i-2}\mathbb{V}_z$ 之间的缩放尺度一致因子

　　　　通过模式匹配计算 ${}^{i-1}_{i-2}\mathbb{V}_t$ 和 ${}^{i-1}_{i-2}\mathbb{V}_t$ 之间的平移尺度一致因子

　　　　更新相应的缩放和平移

　　　　在两帧之间的 4DoF 变换上执行链式法则

　　end if

end for

输出： 和输入 $\mathbb{I}$ 对应的相机位姿

算法 4 平移的尺度一致因子计算

输入： ${}^{2}_{1}\mathbb{V}_t$ 和 ${}^{3}_{2}\mathbb{V}_t$
初始化距离 d 为无穷大
for $s = 0.1 : 0.002 : 10.0$ **do**
    根据 s 将 ${}^{3}_{2}\mathbb{V}_t$ 放缩成 ${}^{3}_{2}\mathbb{V}'_t$
    计算 ${}^{2}_{1}\mathbb{V}_t$ 和 ${}^{3}_{2}\mathbb{V}'_t$ 之间的欧式距离 d_s
    if $d_s < d$ **then**
        $d \leftarrow d_s$
        ${}^{2\rightarrow1}_{3\rightarrow2}s_t \leftarrow s$
    end if
end for
输出： 平移尺度一致因子 ${}^{2\rightarrow1}_{3\rightarrow2}s_t$

算法 5 缩放的尺度一致因子计算

输入： ${}^{2}_{1}\mathbb{V}_z$ 和 ${}^{3}_{2}\mathbb{V}_z$
初始化距离 d 为无穷大
for $\Delta = -r : 1 : r$ **do**　　▷ r 是 ${}^{2}_{1}\mathbb{V}_z$ 的长度
    根据 Δ 将 ${}^{3}_{2}\mathbb{V}_z$ 平移至 ${}^{3}_{2}\mathbb{V}'_z$
    计算 ${}^{2}_{1}\mathbb{V}_z$ 和 ${}^{3}_{2}\mathbb{V}'_z$ 之间的欧氏距离 d_s
    if $d_s < d$ **then**
        $d \leftarrow d_s$
        ${}^{2\rightarrow1}_{3\rightarrow2}s_z \leftarrow$ shift_to_scale$\{\Delta\}$
    end if
end for
输出： 缩放尺度一致因子 ${}^{2\rightarrow1}_{3\rightarrow2}s_z$

4.4　实验与分析

本节在仿真和真实世界的多深度环境中评估了所提出的 eFMT 算法。再次说明，虽然 FMT 有多种不同的实现与变种，但出于鲁棒性和准确度的考虑，本章使用文献[95,96] 中所述的改进 FMT。然而无论是 FMT 的实现是哪种，它们在相移图上都只搜索一个峰，因此在多深度场景中，它们都会遇到困难。

本节首先在仿真环境中进行了最基础的实验，测试了缩放和平移的尺度一致性。该仿真场景中仅包含两个不同深度的平面，用以测试 eFMT 的有效性。然后在真实场景的实验中，本节比较了 eFMT、FMT 和其他最先进的 VO 算法，包括 ORB-SLAM3[86]、SVO[88] 和 DSO [135]。这三种最先进的 VO 算法不依赖于 FMT，但是文献[134] 中指出这三种算法是当下最流行、最具代表性的 VO 算法。真实世界的实验包括两部分：一个简单实验和一个大规模的 UAV 数据集[1]。其中简单实验的场景和仿真场景类似，都是两个深度不同的木板，尽管场景中的元素较少，但由于木板上的场景都非常相似，因此该场景会比一般的室内场景难一些。为了评估 eFMT 在更一般情况下的性能以及为其提供一个可能的应用场景，第二个真实场景实验的数据是由一个装在 UAV 上的、朝向向下的相机采集得到，该场景包括许多不同的元素，例如建筑类的屋顶、草地和溪流等等。由于相机视野中有多种不同深度，尤其是透射投影会使得建筑物变成一个倾斜的平面，从而导致了无穷种深度，因此该场景对于经典的 FMT 难度很大；另外该场景中的特征较少的屋顶以及特征模糊的草坪等也会对经典的 VO 算法提出挑战。本节将在该实验中展示 eFMT 可以同时解决这两大难题：1）多深度场景；2）特征难以提取或匹配。

本章所有的实验都是在 Intel Core i7-4790 CPU@3.6GHz、16GB 内存的计算平台上进行的，相关算法使用 C++ 单线程实现。所有输入图像的大小均为 512×512。

4.4.1　仿真场景下的实验

在该仿真测试中，所有的图像都由 Gazebo 仿真器生成，以获得准确的位姿真值。仿真场景如图4.7所示，相机被装在一个机械臂的末端，从而可以通过控制机械臂来达到移动相机的目的。

[1] https://robotics.shanghaitech.edu.cn/static/datasets/eFMT/ShanghaiTech_Campus.zip

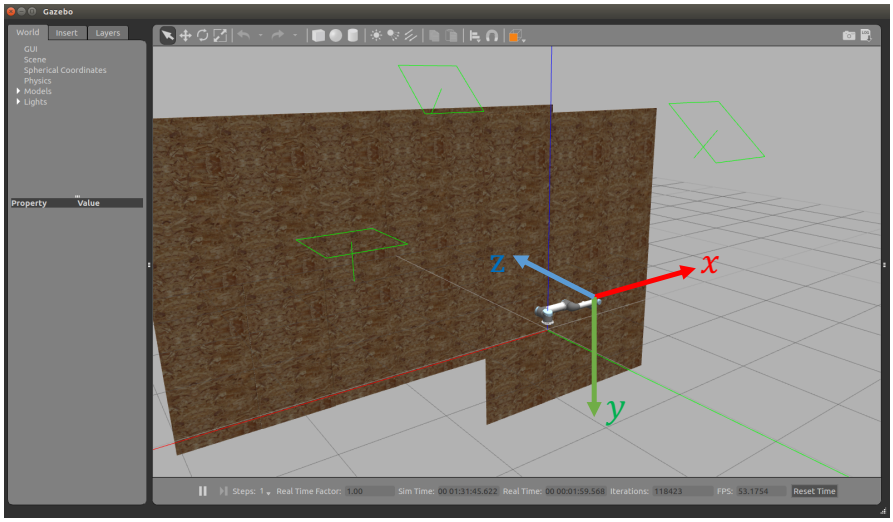


图 4.7 仿真场景示意图

Figure 4.7 Simulated environment

4.4.1.1 缩放的尺度一致性

在这种仿真场景下，沿着 z 轴移动机械臂，可以用相机采集到不同深度的两个平面的图像。这里选取三帧这样的图像，如图4.8b、4.8c和4.8d所示，从上到下，图像被放大了。在每张图像上，图像的左半部分离相机更远，右半部分离相机较近。每两帧之间的旋转和缩放相移图展示在图4.8的第二列中了。可以看到每张相移图上都有两个峰，代表了视野中有不同的深度，而且高一点的峰并不是总在左边，说明视野中的主要深度改变了，这将破坏 FMT 的尺度一致性。经典的 FMT 只使用最高峰，但是，本章提出的 eFMT 考虑缩放能量向量，并通过计算尺度一致性因子将所有的缩放值放到同一尺度下。

表 4.1 缩放估计的闭环分析

Table 4.1 Loop Closure for Zoom Estimation

	${}_{0}^{1}z$	${}_{1}^{2}z$	${}_{0}^{2}z$	$\|\|{}_{0}^{1}z * {}_{1}^{2}z/{}_{0}^{2}z\|\|$
eFMT	0.889	0.889	0.768	1.029
FMT[95,96]	0.889	0.881	0.902	0.868

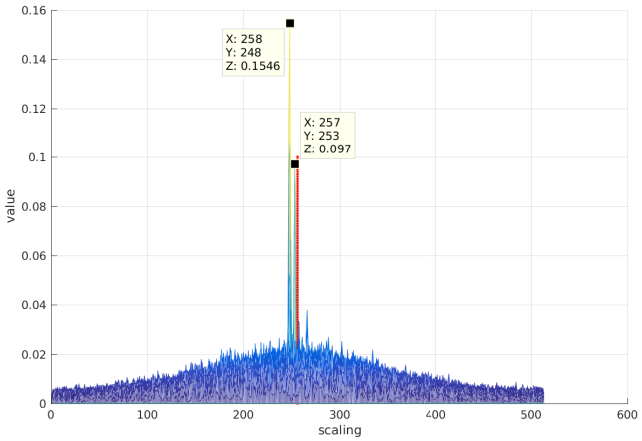
本实验中通过三帧图像组成的一个局部闭环来展示 eFMT 在尺度一致性上优于 FMT，理论上而言，图像 0 和 2 之间的缩放 ${}_{0}^{2}z$ 应该等于图像 0、1 之间的缩放 ${}_{0}^{1}z$ 和图像 1、2 之间的缩放 ${}_{1}^{2}z$ 之间的乘积。表4.1展示了 eFMT 和 FMT 估算得

(a) 视觉辅助输入图像 0

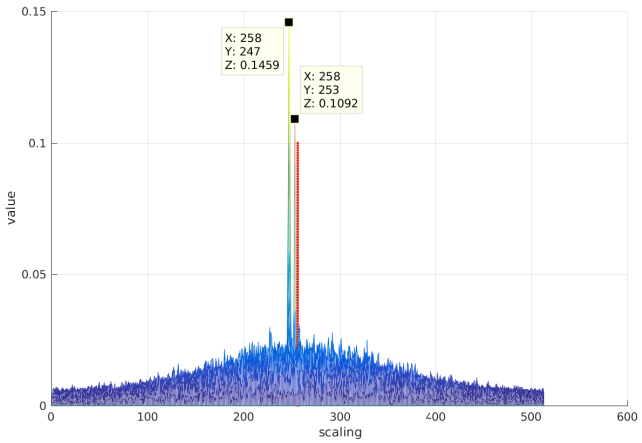
(b) 输入图像 0

(c) 输入图像 1

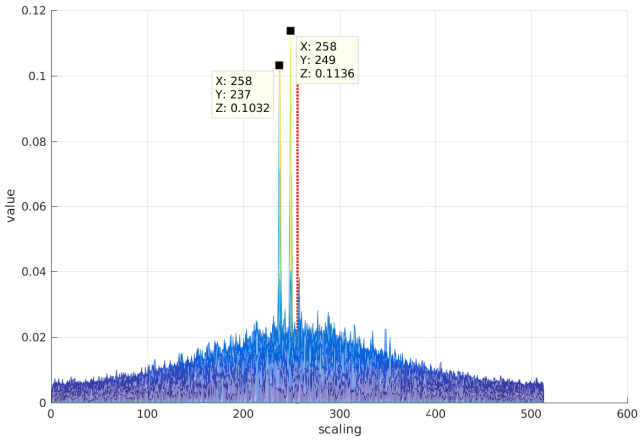
(d) 输入图像 2



(e) 图像 0 和 1 之间的相移图



(f) 图像 1 和 2 之间的相移图



(g) 图像 0 和 2 之间的相移图

图 4.8 多缩放情况下的三帧旋转和缩放相移图示例

Figure 4.8 Three rotation and zoom phase shift diagrams (PSD) with multiple zooms

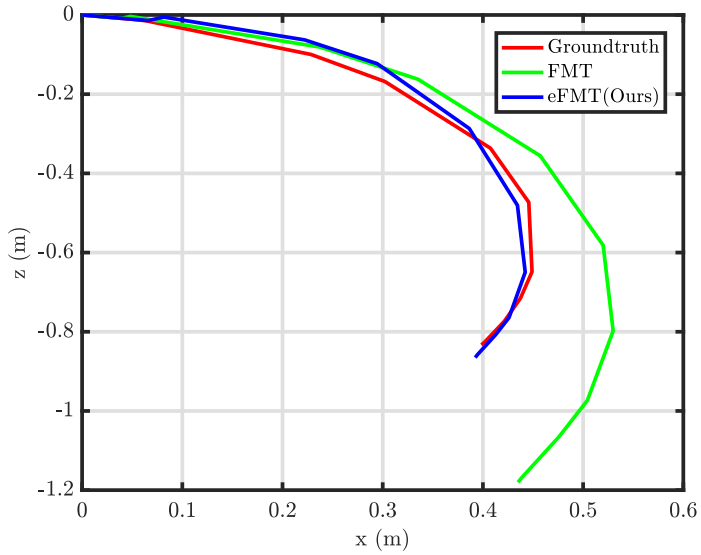


图 4.9 仿真场景中的视觉里程计比较

Figure 4.9 Visual odometry comparison in a simulated scenario

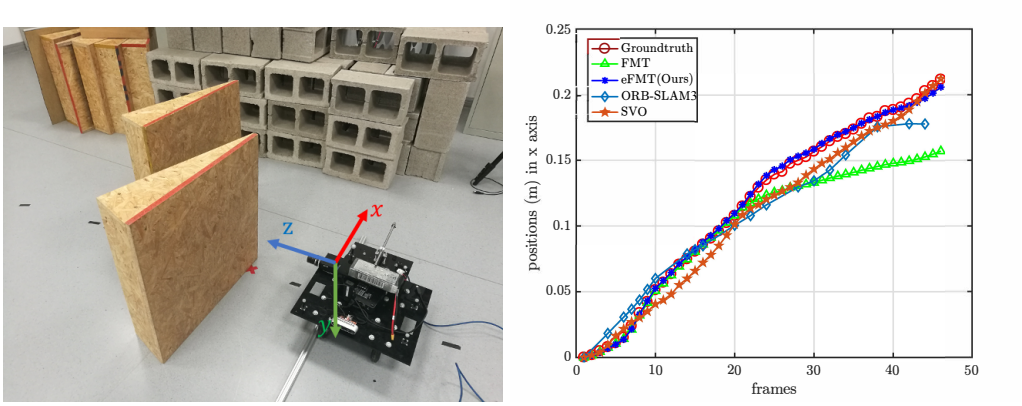
到的这三帧图像之间的缩放，可以看到 eFMT 正确地估计了缩放，因此结果符合缩放闭环，即 $||{}^{1}_{0}z * {}^{2}_{1}z/{}^{2}_{0}z|| \approx 1$；但是 FMT 只跟踪最高的峰，而图4.8g中最高峰和图4.8e、4.8f中的最高峰所对应的平面并不相同，也就是是通过不同平面的缩放估算得到的缩放 ${}^{2}_{0}z$ 和 ${}^{2}_{1}z$，从而导致 $||{}^{1}_{0}z * {}^{2}_{1}z/{}^{2}_{0}z||$ 与 1 相差较大。

4.4.1.2 仿真场景下的视觉里程计

在该仿真场景下，仿真的机械臂在 $x-z$ 平面内移动，从而生成既有平移又有缩放的图像，本实验在该数据集上对比了 eFMT 和 FMT 的 VO 结果。图4.9表明了 eFMT 保持了正确的尺度一致性，而 FMT 从 $z=-0.5m$ 开始发生了尺度不一致，说明在这种多深度的缩放和平移变换下，eFMT 比 FMT 表现更好。eFMT 的良好表现主要得益于4.2节介绍的基于模式匹配的尺度一致因子计算。

4.4.2 真实场景下的实验

在仿真环境下进行了基础实验后，本章在真实环境下进一步比较了 eFMT、FMT 和其他最先进的 VO 算法，用于评估 eFMT 的性能。真实场景下的第一个实验设置和仿真环境下的设置类似，都是在相机前有两个不同深度的木板，如图4.10a所示。该实验下的位姿真值由室内定位系统 OptiTrack 提供。在第二个真实场景的实验中，数据集由无人机上一个朝向向下的相机所采集，无人机从校园上方按指定路线飞行，更多相关细节将在下文中详述。



(a) 真实世界场景（y 轴向下）　　**(b) 视觉里程计结果比较**

图 4.10 真实场景中的视觉里程计比较

Figure 4.10 A visual odometry example in a real-world environment

4.4.2.1　基础实验

本实验评估了不同的 VO 方法在有两种不同深度场景下的表现，其中相机只沿 x 轴进行了平移（见图4.10a）。和仿真场景类似，离相机较近的木板先进入相机视野，然后两个木板都在相机视野中，最后只有较远的木板在相机视野中。

图4.10b展示了不同方法的定位比较结果，包括 FMT（绿色三角形）、eFMT（蓝色星形）、SVO（蓝色三角形）和 ORB-SLAM3（棕色星形），DSO 在该场景中未能成功估计相机位姿，因此其结果没有包含在其中。为了补偿未知的尺度参数，所有算法的估计结果都通过手动放缩的方式和真值（由定位系统提供）对齐。由于在设定中相机只在沿 x 轴方向运动，因此在该图上只展示了 x 轴方向上在不同帧下的位置，但是在表4.2展示的绝对误差中包含了 x 轴和 y 轴的误差，y 轴的运动主要来自于相机移动时的偏差。

从图4.10b中可以看出，FMT 大约从第 20^{th} 帧开始受尺度不一致的影响，因为此时相机视野中占据大部分的平面从较近的平面变成了较远的平面，因此相移图的最高峰对应的平面发生了改变，新的平面距离相机较远，因此对应的像素移动得较慢，因此 FMT 估计的相机运动较之前的帧更小。相较之下，本章所提出的 eFMT 直到最后一帧都一直保持着正确的尺度，主要归功于模式匹配为所有的单位平移向量计算了正确的尺度一致因子。此外，和 SVO、ORB-SLAM3 相比，eFMT 对于每帧位姿估计得更为准确。表4.2展示了不同方法的绝对轨迹误差，包括误差的平均值、最大值和中值，也说明了 eFMT 的误差最小，其次

是 SVO 和 ORB-SLAM3。具体来说，eFMT 的平均误差大约是 ORB-SLAM3 的 1/6、FMT 和 SVO 的 1/8。该实验初步表明在具有挑战性的环境下，得益于频域配准的鲁棒性，eFMT 比当下流行的 VO 算法表现得更好，并且它还成功地弥补了 FMT 在多深度场景下尺度不一致的缺陷。

表 4.2 绝对轨迹误差比较

Table 4.2 Absolute trajectory error comparison

方法	平均值 (mm)	最大值 (mm)	中值 (mm)
FMT[95,96]	17.1	54.7	10.1
eFMT	**2.1**	**6.0**	**1.8**
SVO[74]	17.0	38.0	17.5
ORB-SLAM3[86]	13.0	21.6	13.6
DSO[135]	/	/	/

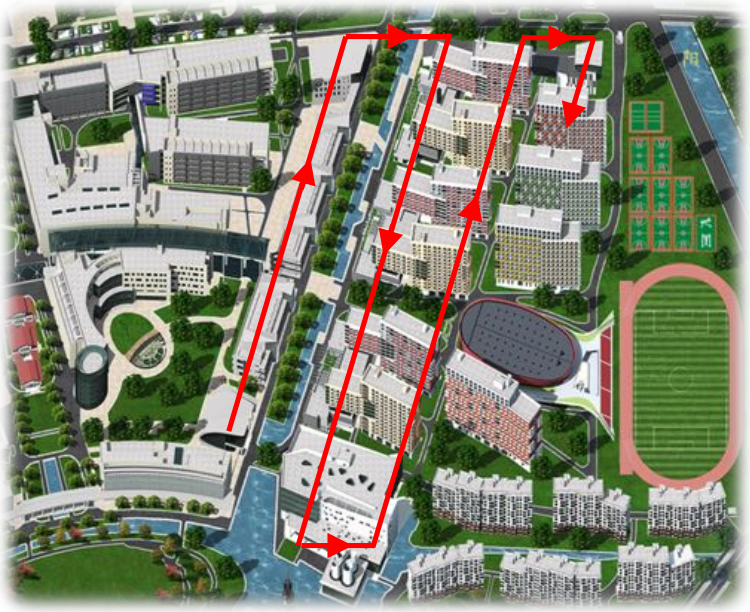


图 4.11 UAV 的飞行路线示意图

Figure 4.11 A UAV’s flying trajectory over a campus

4.4.2.2 UAV 数据集

除了上述的基础实验，本节还将在一个更大的数据集上比较 eFMT、FMT、ORB-SLAM3、SVO 和 DSO。该数据集由装载在 DJI Matrice-300 RTK 的朝向向

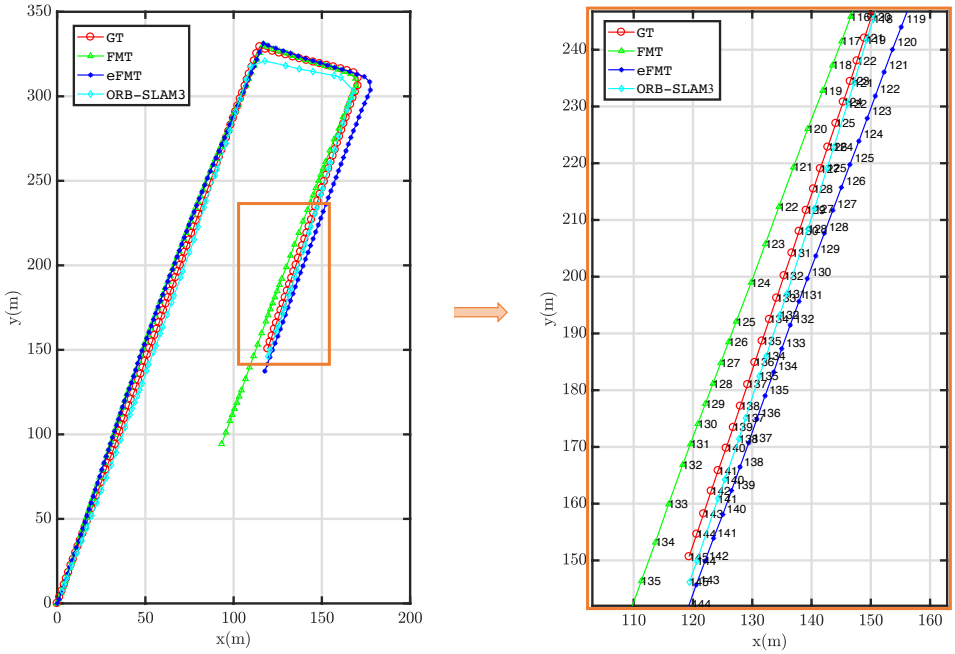


图 4.12 不同算法在 UAV 数据集上的轨迹对比

Figure 4.12 Estimated trajectories on the UAV dataset

下的相机采集得到，无人机的飞行速度设置为 $2m/s$、飞行高度约为 80 米高，比最高楼层高 20 米，图像的采集频率约为 0.5Hz，无人机从校园上空飞行的路线如图4.11所示。该 DJI 无人机的飞行轨迹约为 1,400 米，共采集了 350 帧图像。位置真值由无人机上的厘米级精度的 RTK GPS 提供。正如在实验开头所提到的，该数据集包含各种不同的元素，包括屋顶、路面、小河和草地，这些对不依赖于 FMT 的经典 VO 算法提出了挑战；并且由于多种深度的存在，增加了 FMT 估计位姿的难度。在该实验中，我们将展示 eFMT 不仅保持了 FMT 的鲁棒性，还克服了其单一深度的局限性。

图4.15展示了不同方法估计出来的总体轨迹，估计出来的轨迹和真值通过从第 0 帧和第 80 帧的位姿计算出来的一个旋转系数和一个放缩系数对齐。由于 SVO 和 DSO 在该数据集上未能成功估计相机位姿，因此其估计未包含在图4.15中。此外，从图4.15中可看出 ORB-SLAM3 也在跟踪时失败了几次（见红色星形），每次失败之后，ORB-SLAM3 的路径都需要重新手动对齐。FMT 和 eFMT 在该数据集上一直都能估计相机位姿，尽管存在一些平移的偏移问题。为了评估 FMT、eFMT 和 ORB-SLAM3 的性能，我们专门对比了 ORB-SLAM3 还未

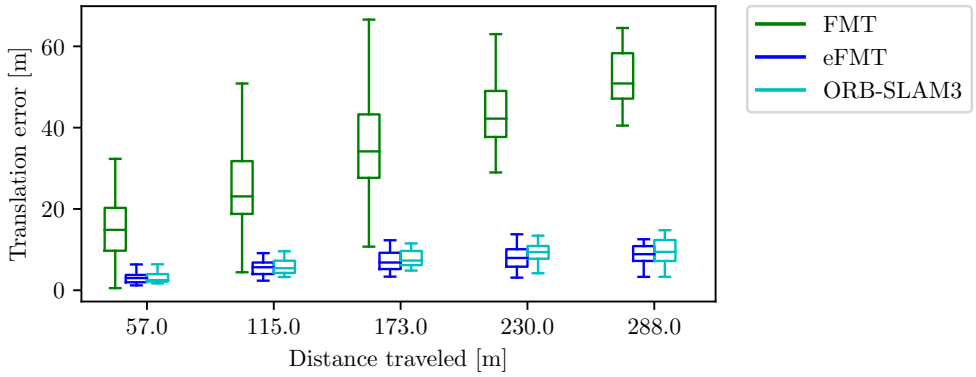


图 4.13 不同算法在 UAV 数据集上的绝对平移误差

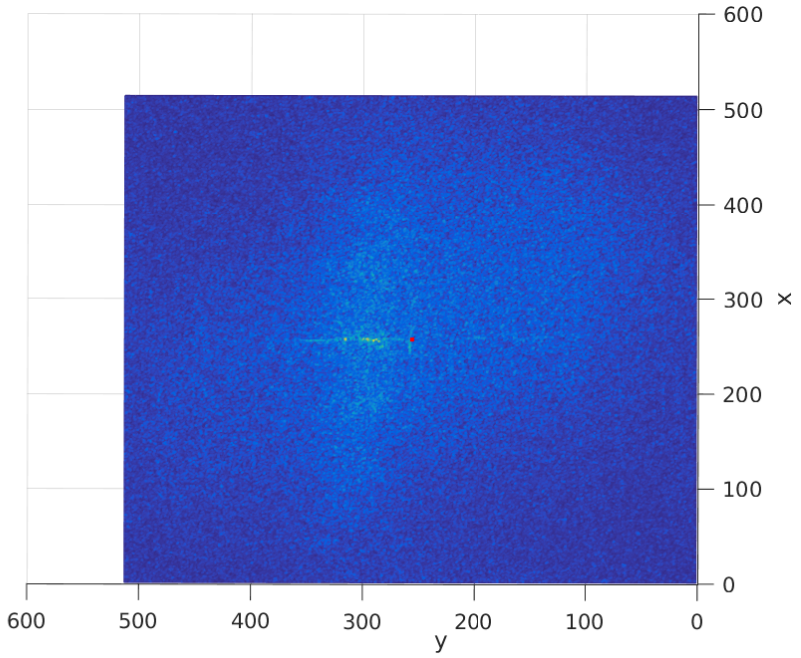
Figure 4.13 Absolute translation error on the UAV dataset

失败之前的一段，结果展示在图4.12中。从右边的局部放大图中可以看到，eFMT 和 ORB-SLAM3 估计得到的两两帧之间的距离几乎是恒定的，说明相机是匀速运动的，和厘米级精度的 RTK GPS 数据真值一致。但是 FMT 估计到的相机速度则是在改变的，例如从 125 帧到 132 帧的速度快于从 132 到 138 帧的速度，这是因为前者对应的图像中占大部分的是地面而后者对应的图像中占大部分的是屋顶。此外，图4.13还展示了不同距离下的绝对平移误差，由文献[137]提供的评估工具生成。如果只比较这三种方法成功跟踪相机位姿时的性能，可以发现 eFMT 和 ORB-SLAM3 的准确度相似，两者都比 FMT 的结果要好，因为 FMT 的性能受多深度的影响较大。

值得关注的是当相机视野中有倾斜平面时，平移的相移图上就会有连续的较高能量值线段。如图4.14所示，受透射投影的影响，在第 143 帧和第 144 帧上建筑物会变成倾斜的平面，这就会导致下方对应的平移相移图上会出现高亮的线段（在红点的左侧），代表能量值较大。在该 UAV 数据集上，这样的倾斜平面很常见，因此在计算尺度一致因子时，采用模式匹配很重要，图4.14中估计的轨迹足以说明 eFMT 可以应对这样的倾斜平面的状况。

综上，本实验表明 eFMT 具有两点优势：1）它成功地将 FMT 扩展到了多深度场景，即不管不同深度是连续的（如倾斜平面）还是离散的（如屋顶和地面），eFMT 都能成功估计相机运动；2）它保持了 FMT 的鲁棒性，即在屋顶等经典 VO 算法无法工作的特征不明显场景下，它还能估计相机运动。然而 eFMT 也还是有缺陷的，如当相机运动了很长一段时间时，该算法的累计误差较大。在未来的工

(a) 第 143 帧　　**(b) 第 144 帧**



(c) 平移相移图

图 4.14 平移相移图上较高的连续能量值示例

Figure 4.14 Example line segment in the translation phase shift diagram (PSD)

作中将引入位姿优化来克服该缺点。

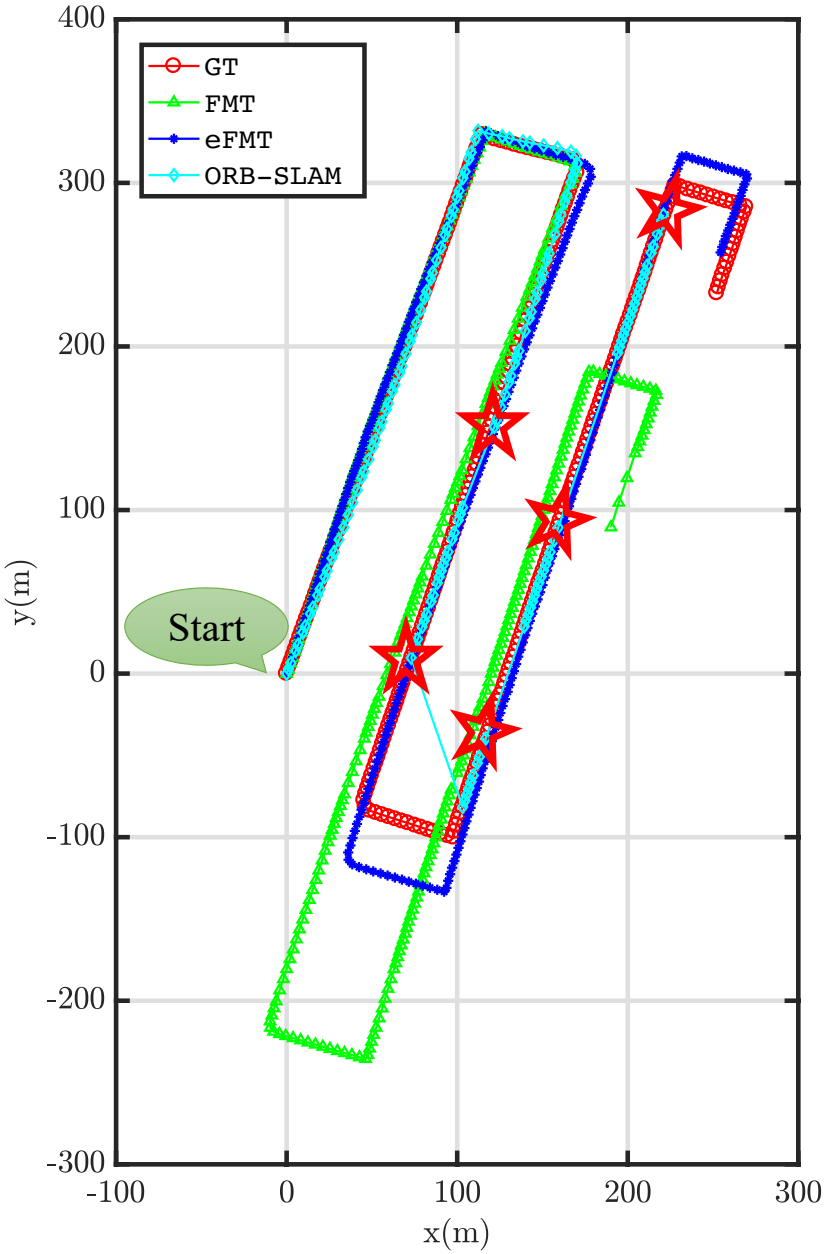


图 4.15 不同方法在 UAV 数据集上的总体轨迹图

Figure 4.15 Overall trajectories of different methods on the UAV dataset

4.4.3 计算分析

文献[120]指出图像的分辨率对 FMT 的计算速度有较大的影响，并且图像降采样不会影响 FMT 的性能。通过初步测试，发现该结论同样适用于 eFMT。当

图像分别率为 512×512 时，eFMT 每帧的平均处理时间约为 0.7 秒；当图像分辨率为 256×256 时，其平均处理时间约为 0.2 秒。总体而言，eFMT 的处理时间约为 FMT 的两倍，该统计主要是基于 C++ 单线程实现来考虑的。eFMT 较慢的一大原因是因为对于多个缩放值，需要多次计算平移能量向量。然而这些计算都是相互独立的，因此可以通过并行计算来进行加速，这样的话 eFMT 理论上可以和 FMT 运行得一样快。

4.5　小结

本章扩展了经典的 FMT，使其能够处理多场景下的缩放和平移。本章详细展示了问题描述，并提出了相关算法，实验表明相比于 FMT，eFMT 在多深度场景下计算 VO 的尺度一致因子更有优势，eFMT 不仅继承了 FMT 的优势，还扩展了它的应用场景。而且在本章中所展示的实验场景下，eFMT 比其他流行的 VO 方法表现得更好，包括 ORB-SLAM3、DSO 和 SVO。由于 FMT 已经被应用于各种不同的场景，因此在未来的工作中也可以考虑将 eFMT 应用到这些场景中，如遥感、图像匹配和定位等。此外，由于 eFMT 比经典的 VO 算法在某些环境下更为鲁棒，因此它还可以应用于其他多深度但特征不明朗的环境下，如水下、多雾等场景。并且如果考虑融合 IMU，eFMT 将可以应用于 6DoF 的运动估计中，辅之以闭环、不确定性检测和凸优化等手段，将可以实现一个完全的 SLAM 系统。

第 5 章　总结与展望

5.1　全文总结

随着 VO/VSLAM 相关技术框架的开源，越来越多的机器人应用在不同场景下使用视觉技术进行自定位。然而当场景难度较大时，现有的 VO/VSLAM 技术受到了一些挑战，如在特征点较少、图像模糊、水下等能见度较低的场景下，现有的 VO/VSLAM 技术有时因无法提取有效信息而不能提供准确的位姿估计。本文利用 FMT 技术进行图像之间的运动估计，从而提高 VO 的准确度与鲁棒性。但是 FMT 只能估计针孔相机的二维运动，且要求场景是单一深度的，为此本文还提出了扩展 FMT 的方法，使其可以应用于多深度场景。总的来说，本文主要研究了 FMT 在视觉机器人定位中的应用，包含以下几个方面。

首先，第 2 章利用 FMT 进行了全向图像的特征匹配，并基于此实现了一个简单的 VO。由于 FMT 只能估计单一深度的场景，本章中将全向图像先转换为全景图像，然后利用递归划分子图策略从全景图像中提取单一深度的子图，其中是否为单一深度可以由 FMT 配准过程中相移图的信噪比判断。通过实验证明，该递归划分子图策略比固定划分子图策略对全向相机的定位更有帮助。而且，通过与其他特征点法的对比，发现利用 FMT 进行匹配点对的提取更为鲁棒，因此可以为全向相机的位姿估计提供更为准确的结果，尤其是在图像模糊、动态场景或草坪等特征类似情况下。

其次，第 3 章中在利用 FMT 计算全景图像的子图运动估计的基础上，提出了一种新颖的全向相机姿态估计方法，将全向相机的旋转估计建模成正弦曲线的拟合问题。首先将全景图像沿水平方向划分为若干子图，然后利用 FMT 计算子图之间的旋转和平移，其次通过数学推导证明这些旋转和平移符合正弦曲线规律，因此只需要对旋转和平移数据进行拟合，得到相关参数即可估算出全向相机的姿态。通过与几何法对比，展示了该基于正弦曲线拟合的全向相机姿态估计算法的准确性和鲁棒性都很高。本章中虽然对全向相机的平移也进行了正弦曲线建模，但是要求场景中的物体深度相同，由于在全向相机的测试场景中，该要求较难满足，因此本章并未进行全向相机的平移估计。此外，多深度场景也会对 FMT 的性能产生影响，本章中通过实验表明正弦曲线拟合的鲁棒性会弥补 FMT

在多深度场景中性能不佳的劣势。

然后，针对前两章中都存在的 FMT 无法应用于多深度场景的缺点，第 4 章提出了扩展傅里叶梅林算法 (eFMT)，使其可以在多深度场景中仍能正常工作。和 FMT 总是搜索相移图中的最高峰不同，eFMT 提取了相移图中能量最大的线段：能量向量，因此保留了多个深度的运动信息。此外，eFMT 还通过能量向量之间的图样匹配，保证了多深度场景下平移的尺度一致性。最后，本章实现了一个基于 eFMT 的 VO 算法，并与当下主流的 VO 算法在多深度场景下进行了对比，发现在测试数据集上，基于 eFMT 的 VO 算法的准确度和 ORB-SLAM3 接近，但是比 ORB-SLAM3 还有其他 VO 算法更加鲁棒。不过，eFMT 算法也存在较为明显的弊端，它和 FMT 一样，只能估计 4DoF 的运动。

5.2 工作展望

正如总结中所提到的，本文所提出的算法还存在弊端：如基于正弦曲线拟合的相机姿态估计算法无法估计相机的平移，eFMT 只能用于 4DoF 的运动估计等。针对这些弊端，本文将在未来的工作中进行相关的研究。

- 第一点是考虑基于正弦曲线拟合的全向相机位姿估计问题。首先 eFMT 的引入可以使子图之间的运动估计更加鲁棒准确，不再受限于单一深度场景；其次可以考虑将全向相机的平移建模成一个多正弦曲线拟合问题，因为场景中不同的深度会对应不同的正弦曲线，但是对 $x-y$ 平面上的运动而言，这些不同深度对应的正弦曲线只有幅度不同，对沿 z 轴的运动而言，不同深度对应的正弦曲线只有偏移量不同。
- 关于 eFMT 只能用于 4DoF 的运动估计，未来的工作考虑从两方面入手，一是考虑借鉴现有的改进 FMT 的算法，其中部分算法利用分数傅里叶变换估计了图像的仿射变换，可以基于此推测相机的倾斜角；第二种方案则是考虑与其他传感器融合，如 IMU、加速度计等，这些传感器可以提供关于相机倾斜角的信息，但这并不意味着辅之以此类传感器，eFMT 就能估算出 6DoF 的位姿，因为当相机倾斜角很大时，进行倾斜补偿后的图像不一定能保留足够多的重叠信息，用于 eFMT 的运动估计。
- 第三点则是考虑基于 eFMT 的建图，在 eFMT 的分析中提到，不同的深度会导致相移图上不同的能量位置，但是这些能量较高的位置是在同一条线上的，

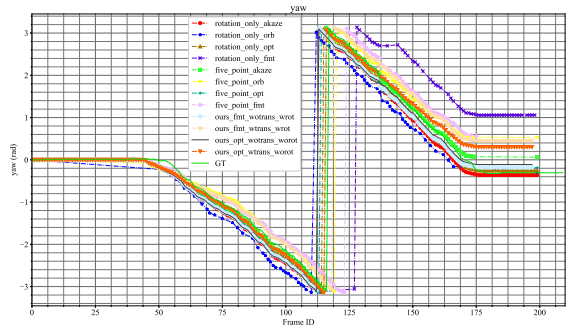
因此可以认为 eFMT 的能量向量中蕴含着深度信息，可以尝试利用多组能量向量来恢复场景的深度图，从而构建环境的三维地图。

- 此外，未来的工作中还将在大雾、水下等场景下进一步评估 eFMT 以及相关拓展算法的鲁棒性。

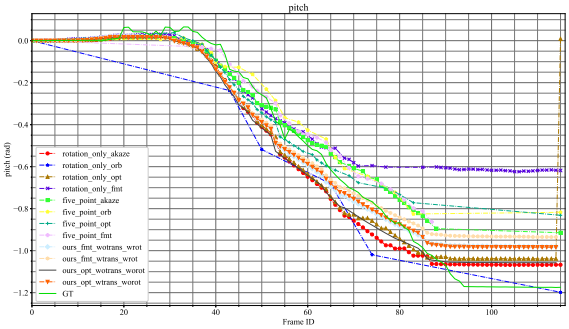
附录 A　中英文术语与缩写对照表

英文缩写	英文全称	中文全称
VO	Visual Odometry	视觉里程计
VSLAM	Visual Simultaneous Localization and Mapping	视觉即时定位与建图
SFM	Structure from Motion	从运动中恢复结构
FMT	Fourier-Mellin Transform	傅里叶梅林变换
eFMT	extended Fourier-Mellin Transform	扩展傅里叶梅林变换
RANSAC	Random Sample Consensus	随机抽样一致算法
PSD	Phase Shift Diagram	相移图
DoF	Degree of Freedom	自由度
RMSE	Root Mean Square Error	均方误差
RMSAE	Root Mean Square Angle Error	均方角度误差
DVL	Doppler Velocity logs	多普勒测距仪
LRF	Laser Range Finder	激光测距仪
GPS	Global Positioning System	全球定位系统
IMU	Inertial Measurement Units	惯性测量单元
BA	Bundle Adjustment	集束调整
FFT	Fast Fourier Transform	快速傅里叶变换
E	Essential Matrix	本质矩阵
PTAM	Parallel Tracking and Mapping	
DTAM	Dense Tracking and Mapping	
SVO	Semi-direct Visual Odometry	
DSO	Direct Sparse Odometry	
FMS	Fourier-Mellin-SOFT	

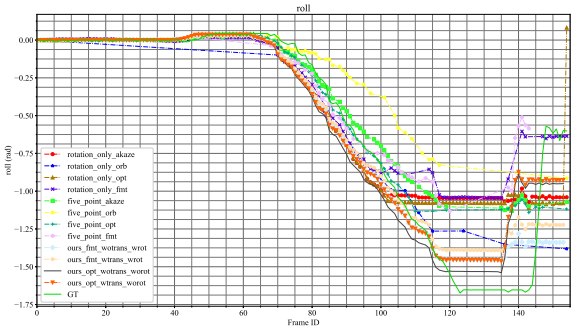
附录 B　正弦曲线拟合方法与几何法在不同数据集上的结果对比



(a) 偏航角



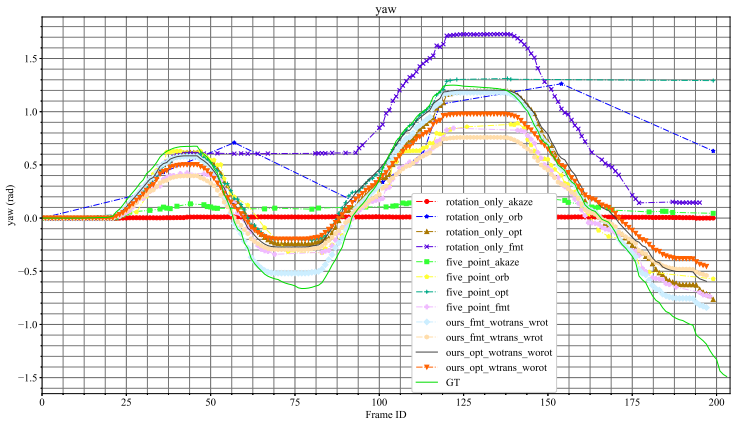
(b) 俯仰角



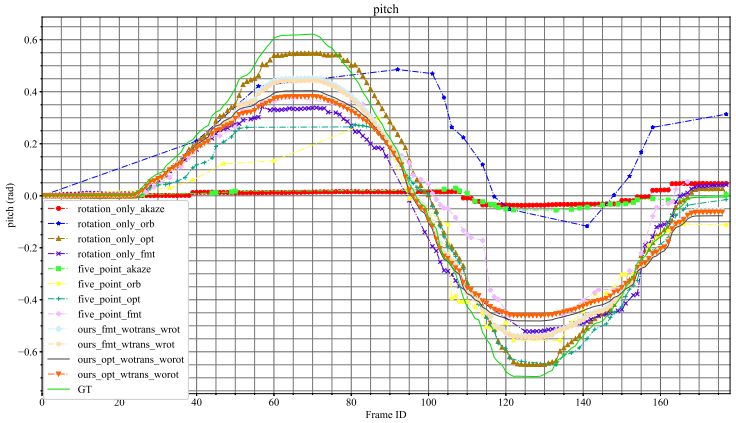
(c) 横滚角

图 B.1 在 indoor_single_yaw、indoor_single_pitch 和 indoor_single_roll 数据集上进行单一旋转估计的定性对比结果

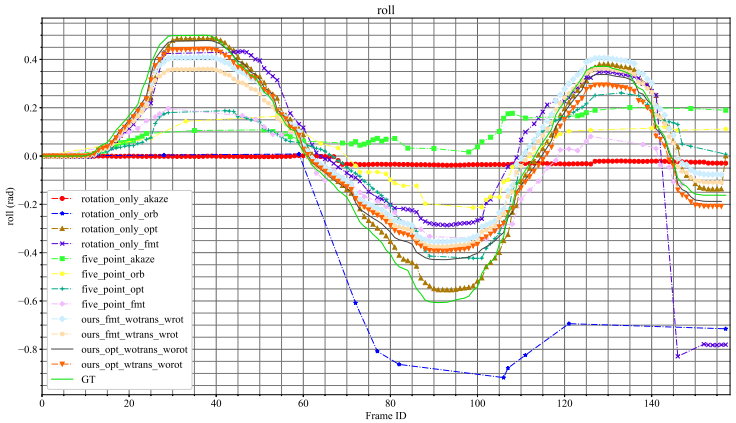
Figure B.1 Qualitative results for single rotation estimation on indoor_single_yaw, indoor_single_pitch and indoor_single_roll datasets



(a) 偏航角



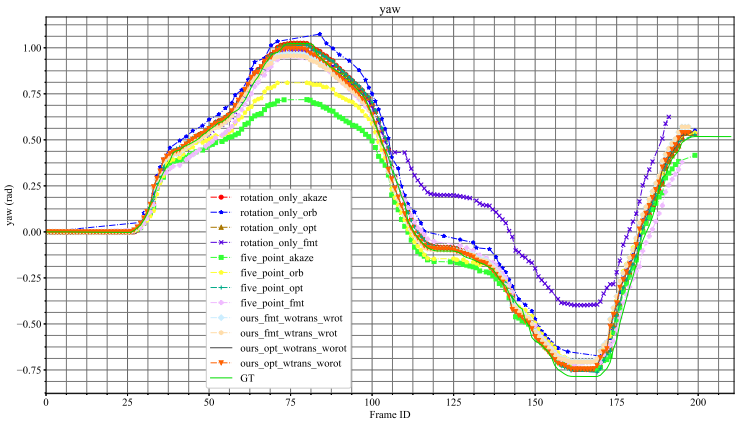
(b) 俯仰角



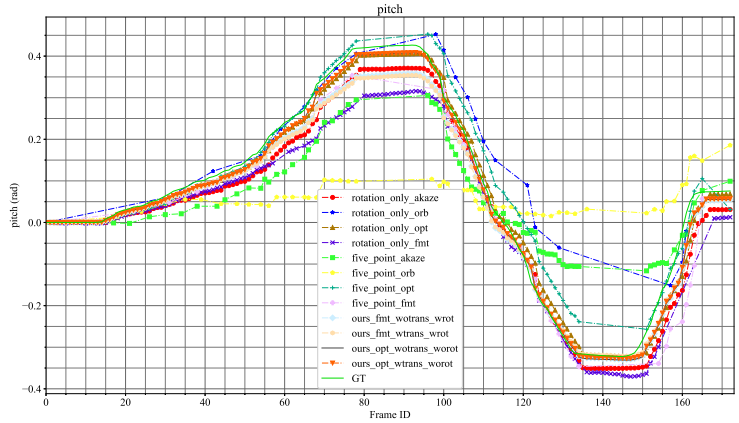
(c) 横滚角

图 B.2 在 grass_single_yaw、grass_single_pitch 和 grass_single_roll 数据集上进行单一旋转估计的定性对比结果

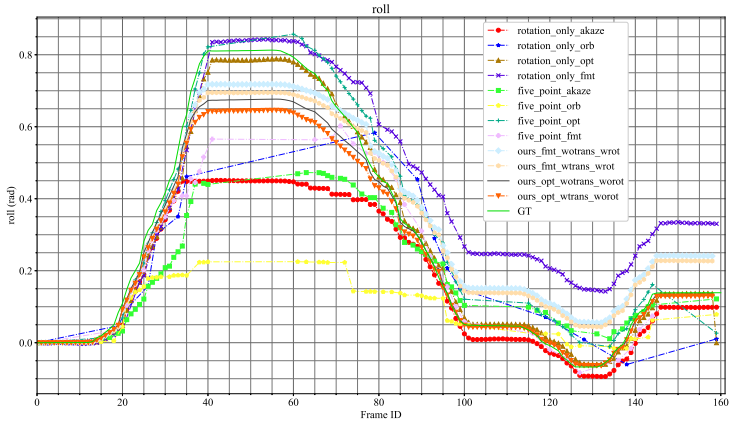
Figure B.2 Qualitative results for single rotation estimation on grass_single_yaw, grass_single_pitch and grass_single_roll datasets



(a) 偏航角



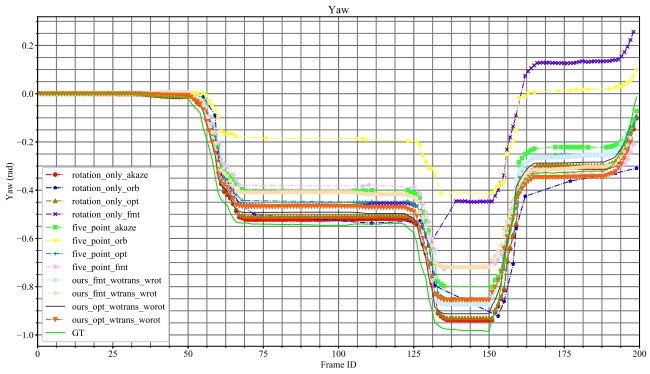
(b) 俯仰角



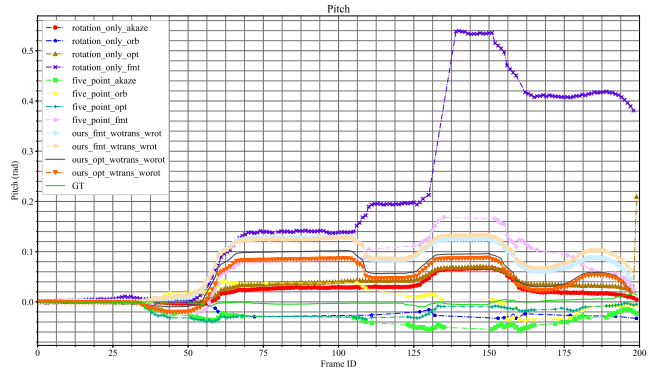
(c) 横滚角

图 B.3 在 street_single_yaw、street_single_pitch 和 street_single_roll 数据集上进行单一旋转估计的定性对比结果

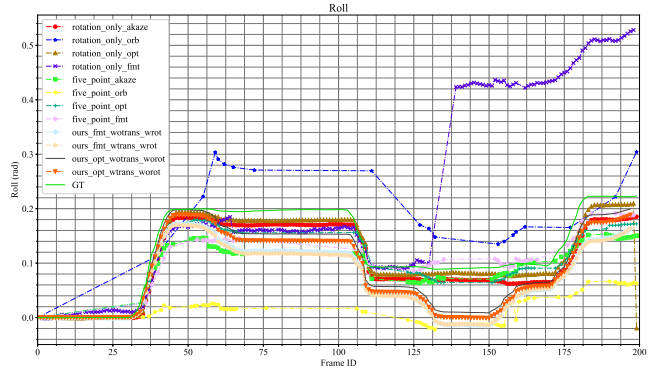
Figure B.3 Qualitative results for single rotation estimation on street_single_yaw, street_single_pitch and street_single_roll datasets



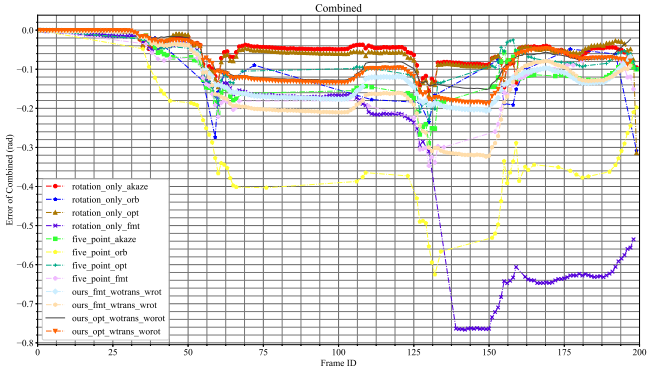
(a) 偏航角



(b) 俯仰角



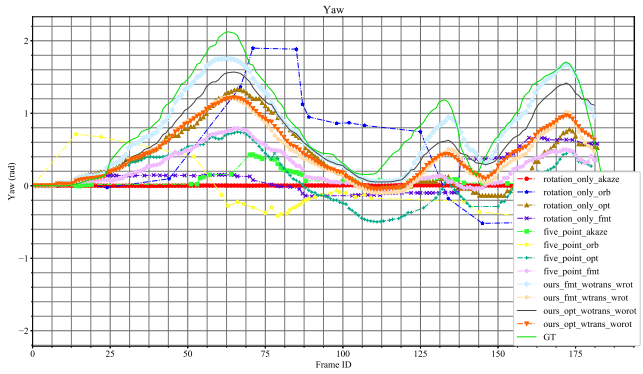
(c) 横滚角



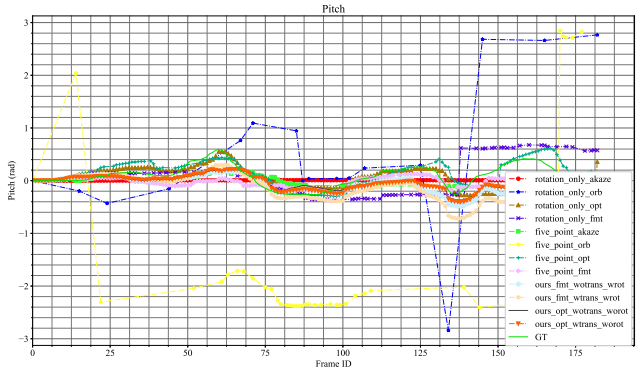
(d) 混合旋转

图 B.4 在 indoor_rpy 数据集上进行混合旋转估计的定性对比结果

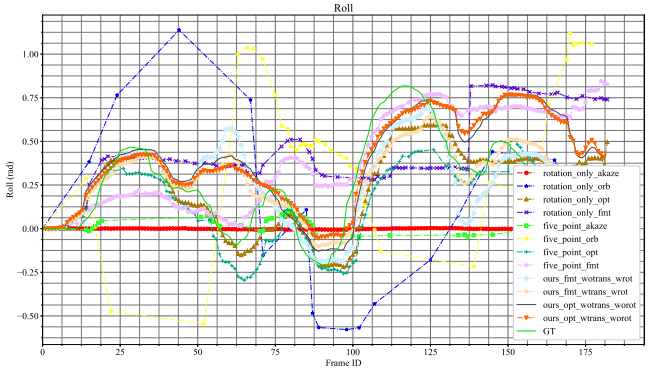
Figure B.4 Hybrid rotation estimation experiments on our datasets: indoor_rpy



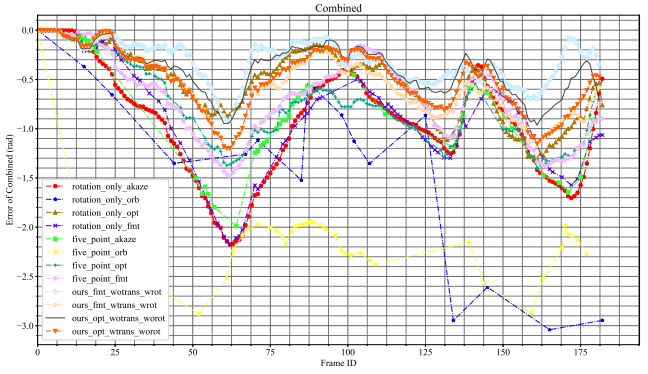
(a) 偏航角



(b) 俯仰角



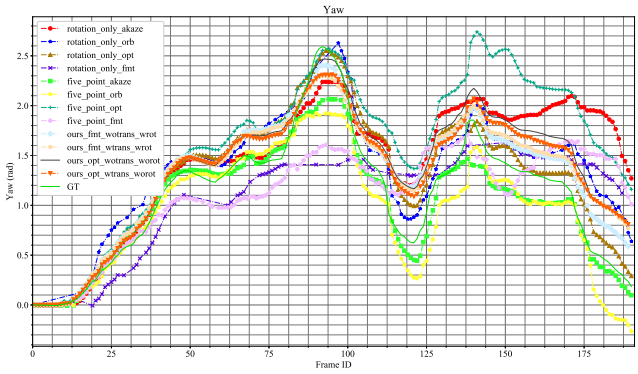
(c) 横滚角



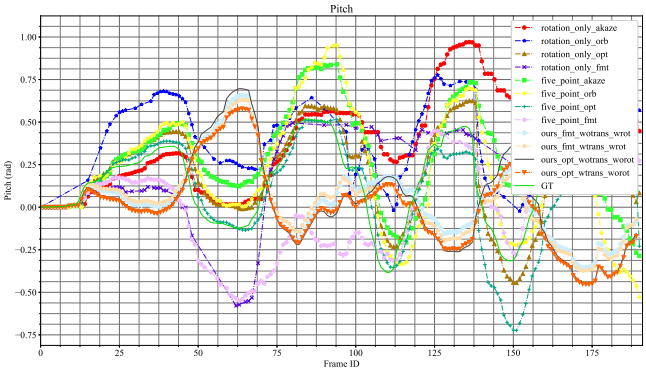
(d) 混合旋转

图 B.5 在 grass_rpy 数据集上进行混合旋转估计的对比结果

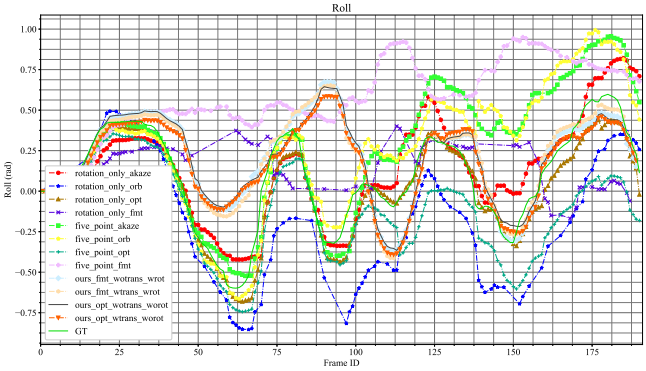
Figure B.5 Hybrid rotation estimation experiments on our datasets: grass_rpy



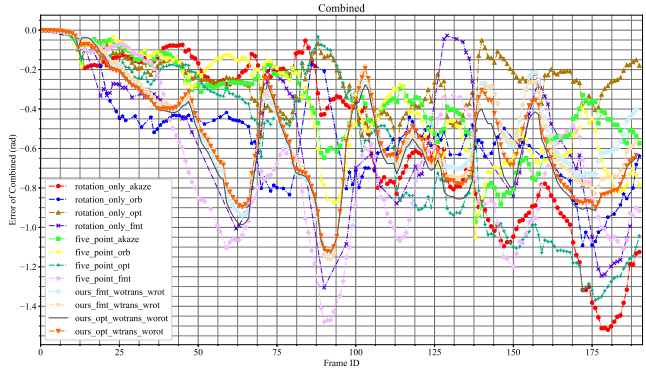
(a) 偏航角



(b) 俯仰角



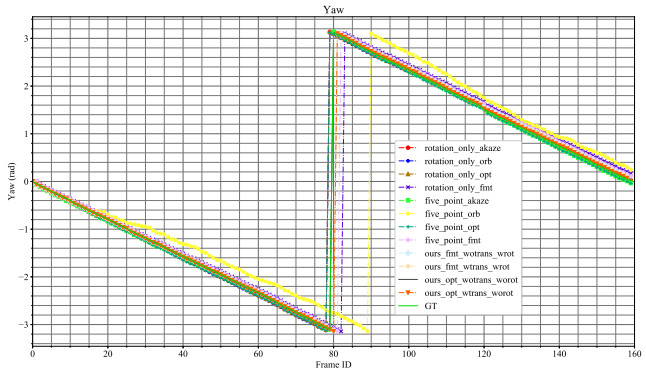
(c) 横滚角



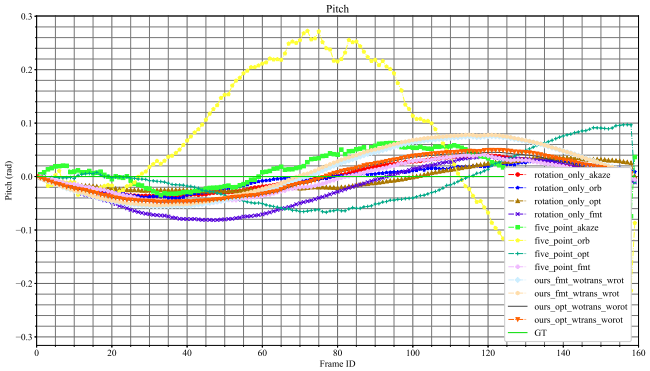
(d) 混合旋转

图 B.6 在 street_rpy 数据集上进行混合旋转估计的对比结果

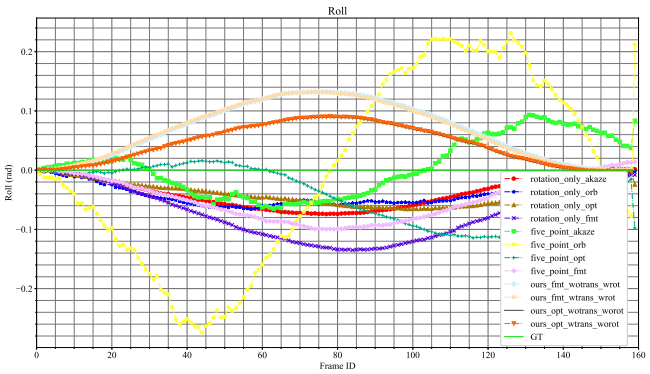
Figure B.6 Hybrid rotation estimation experiments on our datasets: street_rpy



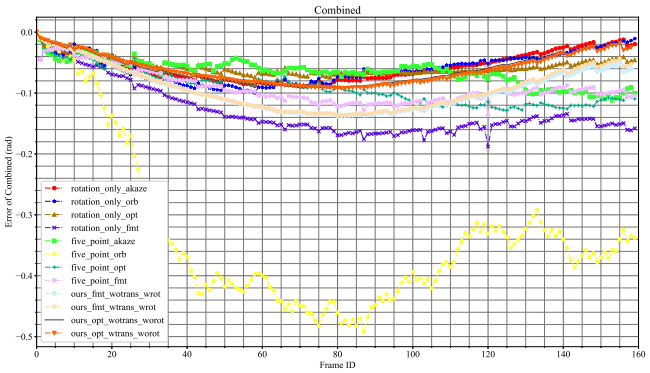
(a) 偏航角



(b) 俯仰角



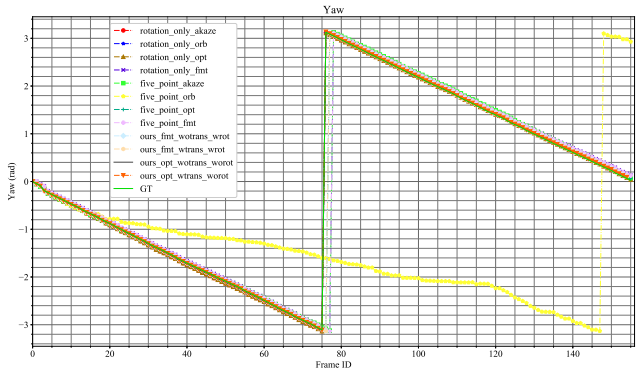
(c) 横滚角



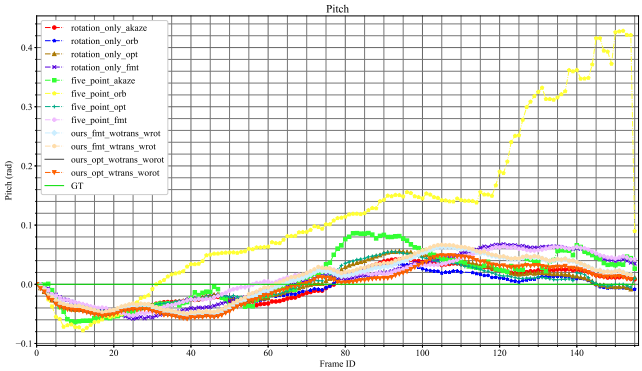
(d) 混合旋转

图 B.7 在 *OVMIS_1* 数据集上进行混合旋转估计的对比结果

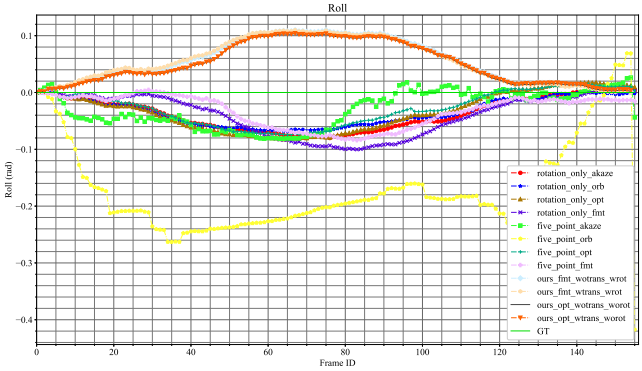
Figure B.7 Hybrid rotation estimation experiments on public datasets: *OVMIS_1*



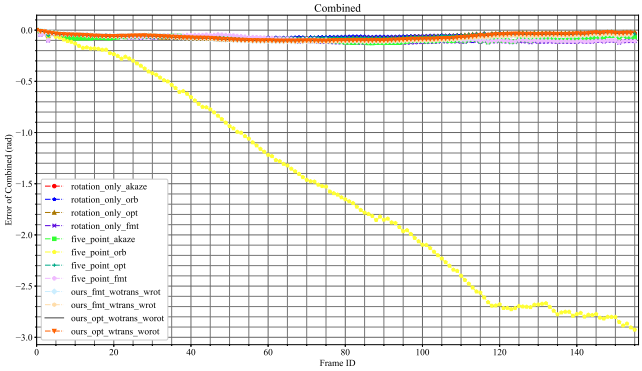
(a) 偏航角



(b) 俯仰角



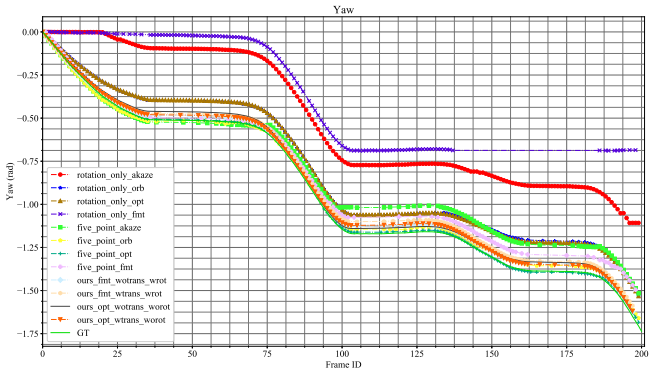
(c) 横滚角



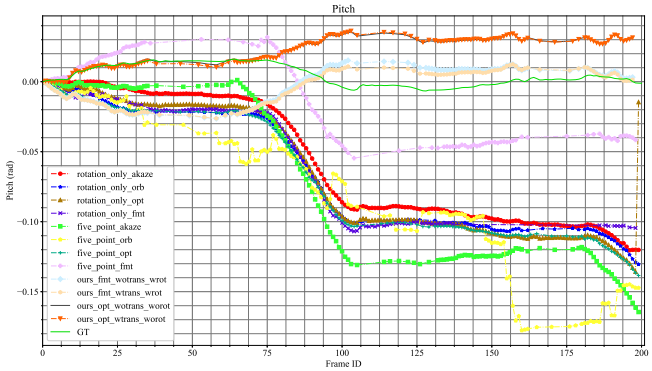
(d) 混合旋转

图 B.8 在 *OVMIS_2* 数据集上进行混合旋转估计的对比结果

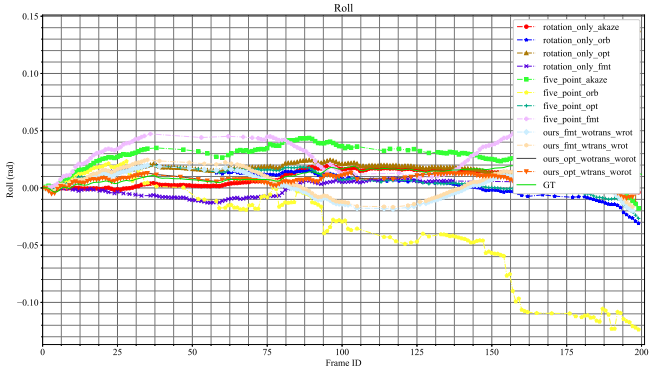
Figure B.8 Hybrid rotation estimation experiments on public datasets: *OVMIS_2*



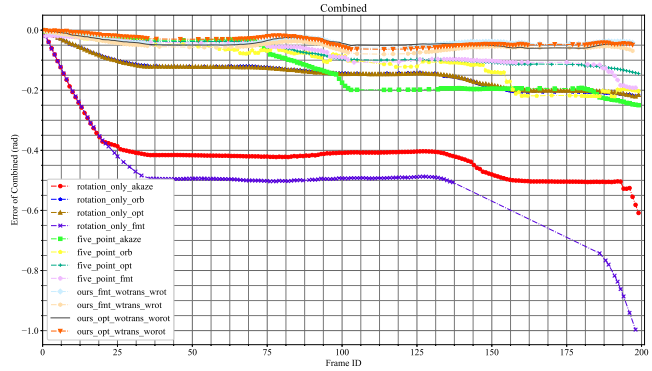
(a) 偏航角



(b) 俯仰角



(c) 横滚角



(d) 混合旋转

图 B.9 在 *CVLIBS* 数据集上进行混合旋转估计的对比结果

Figure B.9 Hybrid rotation estimation experiments on public datasets: *CVLIBS*

参考文献

[1] Scaramuzza D, Martinelli A, Siegwart R. A flexible technique for accurate omnidirectional camera calibration and structure from motion [C]//Computer Vision Systems, 2006 ICVS'06. IEEE International Conference on. IEEE, 2006: 45-45.

[2] Gao X S, Hou X R, Tang J, et al. Complete solution classification for the perspective-three-point problem [J]. IEEE transactions on pattern analysis and machine intelligence, 2003, 25 (8): 930-943.

[3] Schönbein M, Strauss T, Geiger A. Calibrating and centering quasi-central catadioptric cameras [C]//International Conference on Robotics and Automation (ICRA). 2014.

[4] Schönbein M, Geiger A. Omnidirectional 3d reconstruction in augmented manhattan worlds [C]//International Conference on Intelligent Robots and Systems (IROS). 2014.

[5] Morbidi F, Caron G. Phase correlation for dense visual compass from omnidirectional camera-robot images [J]. IEEE Robotics and Automation Letters, 2017, 2(2): 688-695.

[6] Goel R, Gupta P. Robotics and industry 4.0 [M]//A Roadmap to Industry 4.0: Smart Production, Sharp Business and Sustainable Development. Springer, 2020: 157-169.

[7] Dhanabalan T, Sathish A. Transforming indian industries through artificial intelligence and robotics in industry 4.0 [J]. International Journal of Mechanical Engineering and Technology, 2018, 9(10): 835-845.

[8] Ahmadi A, Nardi L, Chebrolu N, et al. Visual servoing-based navigation for monitoring row-crop fields [C]//2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020: 4920-4926.

[9] Wu X, Aravecchia S, Lottes P, et al. Robotic weed control using automated weed and crop classification [J]. Journal of Field Robotics, 2020, 37(2): 322-340.

[10] Stojkoska B L R, Trivodaliev K V. A review of internet of things for smart home: Challenges and solutions [J]. Journal of Cleaner Production, 2017, 140: 1454-1464.

[11] Ahmadi M, Stone P. A multi-robot system for continuous area sweeping tasks [C]// Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006. IEEE, 2006: 1724-1729.

[12] Romero-Trigueros C, Nortes P A, Alarcón J J, et al. Effects of saline reclaimed waters and deficit irrigation on citrus physiology assessed by uav remote sensing [J]. Agricultural water management, 2017, 183: 60-69.

[13] Romero M, Luo Y, Su B, et al. Vineyard water status estimation using multispectral imagery

from an uav platform and machine learning algorithms for irrigation scheduling management [J]. Computers and electronics in agriculture, 2018, 147: 109-117.

[14] Maurer M, Christian Gerdes J, Lenz B, et al. Autonomous driving: technical, legal and social aspects [M]. Springer Nature, 2016.

[15] Scaramuzza D, Fraundorfer F. Visual odometry [tutorial] [J]. IEEE robotics & automation magazine, 2011, 18(4): 80-92.

[16] Scaramuzza D. Omnidirectional camera [J]. Computer Vision: A Reference Guide, 2014: 552-560.

[17] Derpanis K G. The harris corner detector [J]. York University, 2004, 2.

[18] Shi J, et al. Good features to track [C]//1994 Proceedings of IEEE conference on computer vision and pattern recognition. IEEE, 1994: 593-600.

[19] Smith S M, Brady J M. Susan - a new approach to low level image processing [J/OL]. International Journal of Computer Vision, 1997, 23(1): 45-78. http://dx.doi.org/10.1023/A:1007963824710. DOI: 10.1023/a:1007963824710.

[20] Rosten E, Drummond T. Machine learning for high-speed corner detection [C]//European conference on computer vision. Springer, 2006: 430-443.

[21] Lowe D G. Object recognition from local scale-invariant features [C/OL]//Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on: volume 2. 1999: 1150-1157 vol.2. DOI: 10.1109/iccv.1999.790410.

[22] Lowe D. Local feature view clustering for 3d object recognition [C/OL]//Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on: volume 1. 2001: I-682-I-688 vol.1. 10.1109/CVPR.2001.990541.

[23] Bay H, Tuytelaars T, Van Gool L. Surf: Speeded up robust features [C/OL]//Leonardis A, Bischof H, Pinz A. Lecture Notes in Computer Science: volume 3951 European Conference on Computer Vision (ECCV). Springer Berlin / Heidelberg, 2006: 404-417. http://dx.doi.org/10.1007/11744023_32.

[24] Bay H, Ess A, Tuytelaars T, et al. Speeded-up robust features (surf) [J/OL]. Computer Vision and Image Understanding, 2008, 110(3): 346-359. http://www.sciencedirect.com/science/article/pii/S1077314207001555. DOI: 10.1016/j.cviu.2007.09.014.

[25] Rublee E, Rabaud V, Konolige K, et al. Orb: An efficient alternative to sift or surf [C]// Computer Vision (ICCV), 2011 IEEE international conference on. IEEE, 2011: 2564-2571.

[26] Alcantarilla P F, Bartoli A, Davison A J. Kaze features [C]//European Conference on Computer Vision. Springer, 2012: 214-227.

[27] Alcantarilla P F, Nuevo J, Bartoli A. Fast explicit diffusion for accelerated features in nonlinear scale spaces [J]. IEEE Trans. Patt. Anal. Mach. Intell, 2011, 34(7): 1281-1298.

[28] Horn B K, Schunck B G. Determining optical flow [J]. Artificial intelligence, 1981, 17(1-3): 185-203.

[29] Muja M, Lowe D G. Fast approximate nearest neighbors with automatic algorithm configuration. [J]. VISAPP (1), 2009, 2(331-340): 2.

[30] Nister D. An efficient solution to the five-point relative pose problem [C/OL]//2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.: volume 2. 2003: II-195. DOI: 10.1109/CVPR.2003.1211470.

[31] Hartley R I. In defense of the eight-point algorithm [J]. IEEE Transactions on pattern analysis and machine intelligence, 1997, 19(6): 580-593.

[32] Fischler M A, Bolles R C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography [J]. Communications of the ACM, 1981, 24(6): 381-395.

[33] Lepetit V, Moreno-Noguer F, Fua P. Epnp: An accurate o (n) solution to the pnp problem [J]. International journal of computer vision, 2009, 81(2): 155.

[34] Kneip L, Scaramuzza D, Siegwart R. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation [C]//CVPR 2011. IEEE, 2011: 2969-2976.

[35] Sivic J, Zisserman A. Video google: A text retrieval approach to object matching in videos [C]//Computer Vision, IEEE International Conference on: volume 3. IEEE Computer Society, 2003: 1470-1470.

[36] Grisetti G, Kümmerle R, Strasdat H, et al. g2o: A general framework for (hyper) graph optimization [C]//Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Shanghai, China. 2011: 9-13.

[37] Agarwal S, Mierle K, Others. Ceres solver [EB/OL]. http://ceres-solver.org.

[38] Sturm P, Ramalingam S. Camera models and fundamental concepts used in geometric computer vision [M]. Now Publishers Inc, 2011.

[39] Ikeuchi K. Computer vision: A reference guide [M]. Springer Publishing Company, Incorporated, 2014.

[40] Dunne A K, Mallon J, Whelan P F. A comparison of new generic camera calibration with the standard parametric approach [J]. 2007.

[41] Geyer C, Daniilidis K. A unifying theory for central panoramic systems and practical implications [C]//European conference on computer vision. Springer, 2000: 445-461.

[42] Barreto J P. A unifying geometric representation for central projection systems [J]. Computer Vision and Image Understanding, 2006, 103(3): 208-217.

[43] Ying X, Hu Z. Can we consider central catadioptric cameras and fisheye cameras within

a unified imaging model [C]//European Conference on Computer Vision. Springer, 2004: 442-455.

[44] Micusik B, Pajdla T. Structure from motion with wide circular field of view cameras [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(7): 1135-1149.

[45] Dunne A K, Mallon J, Whelan P F. Efficient generic calibration method for general cameras with single centre of projection [C]//2007 IEEE 11th International Conference on Computer Vision. IEEE, 2007: 1-8.

[46] Mei C, Rives P. Single view point omnidirectional camera calibration from planar grids [C]// Proceedings 2007 IEEE International Conference on Robotics and Automation. IEEE, 2007: 3945-3950.

[47] Scaramuzza D, Martinelli A, Siegwart R. A toolbox for easily calibrating omnidirectional cameras [C]//Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on. IEEE, 2006: 5695-5701.

[48] Urban S, Leitloff J, Hinz S. Improved wide-angle, fisheye and omnidirectional camera calibration [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2015, 108: 72-79.

[49] Sturm P, Ramalingam S. A generic concept for camera calibration [C]//European Conference on Computer Vision. Springer, 2004: 1-13.

[50] Ramalingam S, Sturm P. A unifying model for camera calibration [J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(7): 1309-1319.

[51] Nister D, Naroditsky O, Bergen J. Visual odometry [C/OL]//Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.: volume 1. 2004: I-I. DOI: 10.1109/CVPR.2004.1315094.

[52] Scaramuzza D, Fraundorfer F, Siegwart R. Real-time monocular visual odometry for on-road vehicles with 1-point ransac [C]//2009 IEEE International conference on robotics and automation. Ieee, 2009: 4293-4299.

[53] Maimone M, Cheng Y, Matthies L. Two years of visual odometry on the mars exploration rovers [J]. Journal of Field Robotics, 2007, 24(3): 169-186.

[54] Badino H, Yamamoto A, Kanade T. Visual odometry by multi-frame feature integration [C]// Proceedings of the IEEE International Conference on Computer Vision Workshops. 2013: 222-229.

[55] Song S, Chandraker M, Guest C C. Parallel, real-time monocular visual odometry [C]//2013 ieee international conference on robotics and automation. IEEE, 2013: 4698-4705.

[56] Fischler M A, Bolles R C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography: volume 24 [M/OL]. Association for Computing Machinery, 1981: 381–395. https://doi.org/10.1145/358669.358692.

[57] Nistér D, Naroditsky O, Bergen J. Visual odometry for ground vehicle applications [J]. Journal of Field Robotics, 2006, 23(1): 3-20.

[58] Davison A J, Reid I D, Molton N D, et al. Monoslam: Real-time single camera slam [J]. IEEE transactions on pattern analysis and machine intelligence, 2007, 29(6): 1052-1067.

[59] Peng X, Cui J, Kneip L. Articulated multi-perspective cameras and their application to truck motion estimation [C]//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019: 2052-2059.

[60] Goecke R, Asthana A, Pettersson N, et al. Visual vehicle egomotion estimation using the fourier-mellin transform [C]//2007 IEEE Intelligent Vehicles Symposium. IEEE, 2007: 450-455.

[61] Milford M J, Wyeth G F. Single camera vision-only slam on a suburban road network [C]// 2008 IEEE International Conference on Robotics and Automation. IEEE, 2008: 3684-3689.

[62] Bülow H, Birk A, Unnithan V. Online generation of an underwater photo map with improved fourier mellin based registration [C]//IEEE OCEANS. IEEE Press, 2009.

[63] Scaramuzza D, Siegwart R. Appearance-guided monocular omnidirectional visual odometry for outdoor ground vehicles [J]. IEEE transactions on robotics, 2008, 24(5): 1015-1026.

[64] Fraundorfer F, Scaramuzza D, Pollefeys M. A constricted bundle adjustment parameterization for relative scale estimation in visual odometry [C]//2010 IEEE International Conference on Robotics and Automation. IEEE, 2010: 1899-1904.

[65] Sünderhauf N, Konolige K, Lacroix S, et al. Visual odometry using sparse bundle adjustment on an autonomous outdoor vehicle [M]//Autonome Mobile Systeme 2005. Springer, 2006: 157-163.

[66] Konolige K, Agrawal M, Sola J. Large-scale visual odometry for rough terrain [M]//Robotics research. Springer, 2010: 201-212.

[67] Klein G, Murray D. Parallel tracking and mapping for small ar workspaces [C]//2007 6th IEEE and ACM international symposium on mixed and augmented reality. IEEE, 2007: 225-234.

[68] Leutenegger S, Lynen S, Bosse M, et al. Keyframe-based visual–inertial odometry using nonlinear optimization [J]. The International Journal of Robotics Research, 2015, 34(3): 314-334.

[69] Mur-Artal R, Montiel J M M, Tardos J D. Orb-slam: a versatile and accurate monocular slam system [J]. IEEE transactions on robotics, 2015, 31(5): 1147-1163.

[70] Tan W, Liu H, Dong Z, et al. Robust monocular slam in dynamic environments [C]//2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE, 2013: 209-218.

[71] Newcombe R A, Lovegrove S J, Davison A J. Dtam: Dense tracking and mapping in real-time [C]//2011 international conference on computer vision. IEEE, 2011: 2320-2327.

[72] Engel J, Schöps T, Cremers D. Lsd-slam: Large-scale direct monocular slam [C]//European Conference on Computer Vision. Springer, 2014: 834-849.

[73] Engel J, Koltun V, Cremers D. Direct sparse odometry [J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(3): 611-625.

[74] Forster C, Pizzoli M, Scaramuzza D. Svo: Fast semi-direct monocular visual odometry [C]// 2014 IEEE international conference on robotics and automation (ICRA). IEEE, 2014: 15-22.

[75] Argyros A A, Bekris K E, Orphanoudakis S C, et al. Robot homing by exploiting panoramic vision [J]. Autonomous Robots, 2005, 19(1): 7-25.

[76] Lhuillier M. Automatic structure and motion using a catadioptric camera [C]//Proceedings of the 6th Workshop on Omnidirectional Vision, Camera Networks and Non-Classical Cameras. 2005.

[77] Benosman R, Kang S, Faugeras O. Panoramic vision [M]. Springer-Verlag New York, Berlin, Heidelberg, 2000.

[78] Corke P, Strelow D, Singh S. Omnidirectional visual odometry for a planetary rover [C]// 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566): volume 4. IEEE, 2004: 4007-4012.

[79] Rituerto A, Puig L, Guerrero J J. Visual slam with an omnidirectional camera [C]//2010 20th International Conference on Pattern Recognition. IEEE, 2010: 348-351.

[80] Tardif J P, Pavlidis Y, Daniilidis K. Monocular visual odometry in urban environments using an omnidirectional camera [C]//2008 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2008: 2531-2538.

[81] Scaramuzza D, Fraundorfer F. Visual odometry - part i: The first 30 years and fundamentals [J/OL]. Robotics and Automation Magazine (RAM), IEEE, 2011, 18(4): 80-92. DOI: 10.1109/mra.2011.943233.

[82] Lemaire T, Lacroix S. Slam with panoramic vision [J]. Journal of Field Robotics, 2007, 24 (1-2): 91-111.

[83] Hansen P, Corke P, Boles W. Wide-angle visual feature matching for outdoor localization [J]. The International Journal of Robotics Research, 2010, 29(2-3): 267-297.

[84] Arican Z, Frossard P. Omnisift: Scale invariant features in omnidirectional images [C]//2010 IEEE International Conference on Image Processing. IEEE, 2010: 3505-3508.

[85] Payá L, Reinoso O, Berenguer Y, et al. Using omnidirectional vision to create a model of the environment: A comparative evaluation of global-appearance descriptors [J]. Journal of Sensors, 2016, 2016: 1-21.

[86] Campos C, Elvira R, Gómez J J, et al. ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM [J]. arXiv preprint arXiv:2007.11898, 2020.

[87] Urban S, Wursthorn S, Leitloff J, et al. MultiCol Bundle Adjustment: A Generic Method for Pose Estimation, Simultaneous Self-Calibration and Reconstruction for Arbitrary Multi-Camera Systems [J]. International Journal of Computer Vision, 2016: 1-19.

[88] Forster C, Zhang Z, Gassner M, et al. Svo: Semidirect visual odometry for monocular and multicamera systems [J/OL]. IEEE Transactions on Robotics, 2017, 33(2): 249-265. DOI: 10.1109/TRO.2016.2623335.

[89] Matsuki H, von Stumberg L, Usenko V, et al. Omnidirectional dso: Direct sparse odometry with fisheye cameras [J]. IEEE Robotics and Automation Letters, 2018, 3(4): 3693-3700.

[90] Seok H, Lim J. Rovo: Robust omnidirectional visual odometry for wide-baseline wide-fov camera systems [C]//2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019: 6344-6350.

[91] Reddy B, Chatterji B. An fft-based technique for translation, rotation, and scale-invariant image registration [J]. Image Processing, IEEE Transactions on, 1996, 5(8): 1266-1271.

[92] Cooley J W, Tukey J W. An algorithm for the machine calculation of complex fourier series [J/OL]. Mathematics of Computation, 1965, 19(90): 297-301. www.jstor.org/stable/2003354. DOI: 10.2307/2003354.

[93] Anuta P E. Spatial registration of multispectral and multitemporal digital imagery using fast fourier transform techniques [J/OL]. IEEE Transactions on Geoscience Electronics, 1970, 8 (4): 353-368. DOI: 10.1109/TGE.1970.271435.

[94] Chen Q s, Defrise M, Deconinck F. Symmetric phase-only matched filtering of fourier-mellin transforms for image registration and recognition [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1994(12): 1156-1168.

[95] Bülow H, Birk A. Fast and robust photomapping with an unmanned aerial vehicle (uav) [C]// 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2009: 3368-3373.

[96] Bülow H, Birk A, Unnithan V. Online generation of an underwater photo map with improved fourier mellin based registration [C]//OCEANS 2009-EUROPE. IEEE, 2009: 1-6.

[97] Tong X, Ye Z, Xu Y, et al. Image registration with fourier-based image correlation: A comprehensive review of developments and applications [J/OL]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2019, 12(10): 4062-4081. DOI: 10.1109/JSTARS.2019.2937690.

[98] Xu Q, Chavez A G, Bülow H, et al. Improved fourier mellin invariant for robust rotation

estimation with omni-cameras [C]//2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019: 320-324.

[99] Pfingsthorn M, Bülow H, Birk A, et al. Large-scale mosaicking with spectral registration based simultaneous localization and mapping (ifmi-slam) in the ligurian sea [C]//IEEE Oceans. IEEE Press, 2013.

[100] Guo X, Xu Z, Lu Y, et al. An application of fourier-mellin transform in image registration [C]//The Fifth International Conference on Computer and Information Technology (CIT'05). IEEE, 2005: 619-623.

[101] Abdelfattah R, Nicolas J. Insar image co-registration using the fourier–mellin transform [J]. International Journal of Remote Sensing, 2005, 26(13): 2865-2876.

[102] Abdullahi S M, Wang H. Fourier-mellin transform and fractal coding for secure and robust fingerprint image hashing [C]//2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2018: 1-7.

[103] Cassinis R, Duina D, Inelli S, et al. Unsupervised matching of visual landmarks for robotic homing using fourier–mellin transform [J]. Robotics and Autonomous Systems, 2002, 40 (2-3): 131-138.

[104] Bülow H, Birk A. Scale-free registrations in 3d: 7 degrees of freedom with fourier-mellin-soft transforms [J/OL]. International Journal of Computer Vision (IJCV), 2018, 126(7): 731-750. DOI: 10.1007/s11263-018-1067-5., 2018.

[105] Bülow H, Mueller C A, Chavez A G, et al. A divide and conquer method for 3d registration of inhomogeneous, partially overlapping scans with fourier mellin soft (fms) [C]//2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020: 8594-8601.

[106] Ye Z, Xu Y, Hoegner L, et al. Precise disparity estimation for narrow baseline stereo based on multiscale superpixels and phase correlation. [J]. International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, 2019.

[107] Ordonez A, Argüello F, Heras D B. Fourier–mellin registration of two hyperspectral images [J]. International Journal of Remote Sensing, 2017, 38(11): 3253-3273.

[108] Kazik T, Göktoğan A H. Visual odometry based on the fourier-mellin transform for a rover using a monocular ground-facing camera [C]//2011 IEEE International Conference on Mechatronics. IEEE, 2011: 469-474.

[109] Checchin P, Gérossier F, Blanc C, et al. Radar scan matching slam using the fourier-mellin transform [C]//Field and Service Robotics. Springer, 2010: 151-161.

[110] Lucchese L. Estimating affine transformations in the frequency domain [C]//Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205): volume 2. IEEE, 2001: 909-912.

[111] Bülow H, Birk A. Large-scale fourier-mellin invariant registration using a dirichlet based phase filter [J/OL]. under review, 2019. http://robotics.jacobs-university.de/TMP/Dirichlet-FMI.pdf.

[112] Ho H T, Goecke R. Optical flow estimation using fourier mellin transform [C]//2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008: 1-8.

[113] Pathak K, Pfingsthorn M, Bülow H, et al. Robust estimation of camera-tilt for ifmi based underwater photo-mapping using a calibrated monocular camera [C]//Robotics and Automation (ICRA), 2013 IEEE International Conference on. IEEE, 2013: 5111-5116.

[114] Kuang H, Xu Q, Long X, et al. Pose estimation for omni-directional cameras using sinusoid fitting [C/OL]//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2019: 900-906. DOI: 10.1109/IROS40897.2019.8968087.

[115] Hurtos N, Ribas D, Cufí X, et al. Fourier-based registration for robust forward-looking sonar mosaicing in low-visibility underwater environments [J/OL]. Journal of Field Robotics, 2015, 32(1): 123-151. https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21516. DOI: doi:10.1002/rob.21516.

[116] Hurtos N, Nagappa S, Palomeras N, et al. Real-time mosaicing with two-dimensional forward-looking sonar [C/OL]//IEEE International Conference on Robotics and Automation (ICRA). 2014: 601-606. DOI: 10.1109/icra.2014.6906916.

[117] Hurtos N, Nagappa S, Cufí X, et al. Evaluation of registration methods on two-dimensional forward-looking sonar imagery [C/OL]//MTS/IEEE OCEANS. 2013: 1-8. DOI: 10.1109/OCEANS-Bergen.2013.6608124.

[118] Birk A, Wiggerich B, Bülow H, et al. Safety, security, and rescue missions with an unmanned aerial vehicle (uav): Aerial mosaicking and autonomous flight at the 2009 european land robots trials (elrob) and the 2010 response robot evaluation exercises (rree) [J]. Journal of Intelligent and Robotic Systems, 2011, 64(1): 57-76.

[119] Bülow H, Birk A. Fast and robust photomapping with an unmanned aerial vehicle (uav) [C]// International Conference on Intelligent Robots and Systems (IROS). IEEE Press, 2009.

[120] Schwertfeger S, Bülow H, Birk A. On the effects of sampling resolution in improved fourier mellin based registration for underwater mapping [J]. IFAC Proceedings Volumes, 2010, 43 (16): 617-622.

[121] Nister D. An efficient solution to the five-point relative pose problem [J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(6): 756-770. DOI: 10.1109/TPAMI.2004.17.

[122] Hartley R, Zisserman A. Multiple view geometry in computer vision [M]. Cambridge university press, 2003.

[123] Kneip L, Furgale P. Opengv: A unified and generalized approach to real-time calibrated geometric vision [C]//Robotics and Automation (ICRA), 2014 IEEE International Conference on. IEEE, 2014: 1-8.

[124] Stewénius H, Nistér D, Kahl F, et al. A minimal solution for relative pose with unknown focal length [C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05): volume 2. IEEE, 2005: 789-794.

[125] Nister D, Stewenius H. Using algebraic geometry for solving polynomial problems in computer vision [C]//tutorial, IEEE International Conference on Computer Vision (ICCV). 2005: 283.

[126] Tareen S A K, Saleem Z. A comparative analysis of sift, surf, kaze, akaze, orb, and brisk [C]//Computing, Mathematics and Engineering Technologies (iCoMET), 2018 International Conference on. IEEE, 2018: 1-10.

[127] Ordóñez Á, Argüello F, Heras D B. Gpu accelerated fft-based registration of hyperspectral scenes [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017, 10(11): 4869-4878.

[128] Lindoso A, Entrena L, López-Ongil C, et al. Correlation-based fingerprint matching using fpgas [C]//Proceedings. 2005 IEEE International Conference on Field-Programmable Technology, 2005. IEEE, 2005: 87-94.

[129] Labrosse F. The visual compass: Performance and limitations of an appearance-based method [J]. Journal of Field Robotics, 2006, 23(10): 913-941.

[130] Barron J L, Fleet D J, Beauchemin S S. Performance of optical flow techniques [J]. International journal of computer vision, 1994, 12(1): 43-77.

[131] Marquardt D W. An algorithm for least-squares estimation of nonlinear parameters [J]. Journal of the society for Industrial and Applied Mathematics, 1963, 11(2): 431-441.

[132] Huber P J. Robust estimation of a location parameter [M]//Breakthroughs in statistics. Springer, 1992: 492-518.

[133] Charbonnier P, Blanc-Féraud L, Aubert G, et al. Deterministic edge-preserving regularization in computed imaging [J]. IEEE Transactions on image processing, 1997, 6(2): 298-311.

[134] He M, Zhu C, Huang Q, et al. A review of monocular visual odometry [J]. The Visual Computer, 2020, 36(5): 1053-1065.

[135] Engel J, Koltun V, Cremers D. Direct sparse odometry [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018.

[136] Reddy B S, Chatterji B N. An fft-based technique for translation, rotation, and scale-invariant image registration [J]. IEEE transactions on image processing, 1996, 5(8): 1266-1271.

[137] Zhang Z, Scaramuzza D. A tutorial on quantitative trajectory evaluation for visual(-inertial) odometry [C]//IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS). 2018.

作者简历及攻读学位期间发表的学术论文与研究成果

作者简历

徐晴雯，女，江苏省南通人，1993 年 11 月出生，中国科学院上海微系统与信息技术研究所博士研究生。

通讯地址：上海市浦东新区华夏中路 393 号，上海科技大学

邮编：201210

E-mail: xuqw@shanghaitech.edu.cn

教育经历

- 2011 年 8 月至 2015 年 6 月，东南大学，本科，专业：信息工程
- 2015 年 9 月至 2021 年 6 月，中国科学院上海微系统与信息技术研究所，硕博连读研究生，专业：通信与信息系统

研究兴趣

移动机器人，视觉 SLAM

已发表 (或正式接受) 的学术论文:

†: equal contribution

⋆: corresponding author

1. **Q. Xu**, H. Kuang, L. Kneip, and S. Schwertfeger Rethinking the Fourier-Mellin Transform: Multiple Depths in the Camera’s View. *Remote Sensing* 13, no. 5 (2021): 1000.
2. **Q. Xu***, A. G. Chavez, H. Bülow, A. Birk, and S.Schwertfeger Improved Fourier Mellin Invariant for Robust Rotation Estimation with Omni-cameras[C]. *IEEE International Conference on Image Processing (ICIP)*, Taiwan, China. 2019: 320-324.
3. **Q. Xu**†, Z. He†, Z. Chen, and Y. Jiang An Optical Flow Based Multi-Object Tracking Approach Using Sequential Convex Programming[C]. *International Conference on Control, Automation, Robotics and Vision (ICARCV)*. 2020: 1216-

1221.

4. X. Long†, **Q. Xu**†, Y. Yuan, Z. He, and S. Schwertfeger Improved Visual-Inertial Localization for Low-cost Rescue Robots[C]. *21st World Congress of the International Federation of Automatic Control (IFAC)*. 2020.
5. H. Kuang, **Q. Xu**⋆, X. Long, and S. Schwertfeger Pose Estimation for Omni-directional Cameras using Sinusoid Fitting[C]. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Macau, China. 2019: 900-906.
6. A. G. Chavez, **Q. Xu**, C. A. Mueller, S. Schwertfeger, and A. Birk Adaptive Navigation Scheme for Optimal Deep-Sea Localization Using Multimodal Perception Cues[C]. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Macau, China. 2019: 7211-7218.
7. Y.Yuan, **Q. Xu**, and S. Schwertfeger Configuration-Space Flipper Planning on 3D Terrain[C]. *IEEE International Symposium on Safety, Security, Rescue Robotics (SSRR)*. 2020: 318-325.

在投论文:

†: equal contribution

1. **Q. Xu**, H. Bülow, S.Schwertfeger, and A. Birk 3D Visual Odometry based on 2.5D Spectral Registration of Omnidirectional 2D Images. *Robotics and Autonomous Systems*, under review.

2. **Q. Xu**, X. Long, H. Kuang, and S. Schwertfeger Rotation Estimation for Omni-directional Cameras using Sinusoid Fitting. *Journal of Intelligent & Robotic Systems*, minor revision.

3. **Q. Xu**†, Z. Chu†, Y, Jiang†, B. Houska, C.N. Jones, and S. Schwertfeger Optimal Experiment Design Based Self-Reflective Multi-Sensor Calibration. *The IEEE Control Systems Letters*, under review.

申请或已获得的专利:

1. 师泽仁, **徐晴雯**, 龙肖灵, 旷皓飞. 一种基于正弦曲线拟合的全景视觉罗盘估计方法及其应用. CN202010131131.9[P]. 2020-02-28. (实质审查)

2. **徐晴雯**, 师泽仁. 一种应用于多深度场景的扩展傅里叶梅林定位算法. CN202010794306.4[P]. 2020-08-10. (实质审查)

获奖情况:

1. 2016-2017 学年上海科技大学“优秀学生助教”
2. 2016 年全国研究生数学建模三等奖

致 谢

时光飞逝，转眼就到了读博生涯的第六个年头。这六年里，我学会了很多，也成长了很多，这将成为我人生中最珍贵最难忘的旅程之一。在完成这篇博士论文之际，我特别想感谢那些在这六年里给了我帮助与鼓励的人。首先，我特别想感谢我的导师 Sören Schwertfeger 教授。于我而言，Sören Schwertfeger 教授亦师亦友，他不仅在科研道路上给了我很大的帮助，经常帮我答疑解惑，还对我的价值观塑造产生了很大的影响，让我学会更好地待人接物，在遇到挫折时保持良好的心态。很幸运能够成为 Sören Schwertfeger 教授的第一届博士生，并在他的指导下完成了这篇博士论文。

在读博期间，我还要特别感谢 Sören Schwertfeger 教授给我提供了出国交流的机会。在雅各布大学交流的期间，我有幸得到 Andreas Birk 教授和 Heiko Bülow 博士的指导，感谢他们在科研上对我悉心的指导和启发式的建议。也正是在这次交流中，我确定了我的博士研究方向。此外，我也非常感谢在交流期间遇到的小伙伴 Arturo Gomez Chavez、Christian A. Mueller 和 Tobias Doernbach，感谢他们让我在人生地不熟的不莱梅快速适应下来，特别是 Arturo 同学在生活和科研上都给了我很大的帮助。

此外，我还要特别感谢上科大的 Laurent Kneip 教授，他可以算得上是我的 SLAM 启蒙老师。虽然和他的交流不多，但是每次讨论总能让我醍醐灌顶。我还要感谢实验室的小伙伴们。首先感谢一下隔壁组的姜育宁师兄，虽然我们的研究方向不同，但是姜师兄经常在科研上给我新的启发，让我对自己的研究有了新的认识。在生活上，姜师兄也时常带着我们实验室的这群小伙伴们一起寻找新美食。接着，我想感谢一下我们课题组的所有小伙伴，尤其是旷皓飞、龙肖灵和何振鹏师弟，感谢他们的帮助，让我可以尽快地完成各项实验，并在讨论中发现新的思路。此外，我还想感谢一下实验室的其他同学，尤其是王丹丹、王贺、吴旭阳、冯煦晖、石嘉禾、曲志海等，感谢他们为实验室营造了轻松的氛围，并经常和他们交流生活、学习上的点滴。然后，我还想感谢一下我在上科大的室友们：彭欣、许珂和高欣，感谢他们在生活上给我的照顾，尤其是我和彭欣同学的科研方向类似，又在同一个中心，也经常一起交流学习上的难题。

除了在读博期间结识的新朋友，我也特别感谢我之前的朋友们。虽然不经常见面，但是在遇到生活上的困难时，他们总会安慰我、开解我，我也经常和大学室友一起在群里分享生活趣事，和闺蜜谢宗云一起聊聊最近的生活状态，闺蜜还不时给我寄来好吃的。此外，我还要感谢在实习期间遇到的良师益友，感谢他们在我初入社会时给我了帮助，并让我开拓了视野，也对我今后的职业规划产生了启发。

最后，我特别想感谢我的家人们，感谢他们对我的帮助和支持，感谢父母对我的养育之恩，感谢他们在我遇到困难时作为我最坚强的后盾，在我开心时和我一起分享喜悦。同时，也非常感谢我的男朋友徐乃阳，感谢他的理解与包容，愿意一直支持鼓励我。