基于 ROS 机器人激光雷达与视觉的环境感知交互研究

尹江山

武汉软件工程职业学院 湖北 武汉 430205

【摘　要】：针对单一传感器感知缺陷，本文基于 ROS 构建激光雷达与视觉融合系统，通过 EKF 深度融合、YOLOv5s 语义识别与多模态交互提升室内机器人感知性能。实验表明，目标检测 mAP 达 91.2%，定位精度提升 12.8%，可稳定部署于嵌入式平台。

【关键词】：激光雷达；机器视觉；ROS；多传感器融合；环境感知；多模态人机交互

DOI:10.12417/2811-0536.26.10.063

1 引言

单一传感器难以适配复杂室内场景，本文依托 ROS 生态搭建融合感知交互系统，完成多传感器时空对齐、语义目标跟踪与语音手势协同交互，通过真实场景实验验证其工程实用性。

核心工作可归纳为三点：第一，构建多传感器硬件平台与分层 ROS 软件架构，完成激光雷达与 RGB-D 相机的高精度时空同步与外参标定；第二，提出基于 EKF 的松耦合深度融合方法，结合轻量化 YOLOv5s 算法实现室内目标的识别、三维定位与动态跟踪，构建融合语义信息的分层代价地图；第三，设计实现语音与手势协同的多模态人机交互模块，通过多通道置信度决策机制提升交互鲁棒性。最终通过多组对照实验，从目标检测精度、定位性能、交互响应延迟三个维度完成系统综合验证。

2 相关工作

多传感器融合分为数据级、特征级、决策级，不同层级在性能、算力消耗与部署难度上差异显著。数据级融合常利用棋盘靶标开展外参标定，通过靶标角点与激光扫描线几何约束求解传感器旋转、平移参数，奠定标定基础。特征级融合将激光点云投影到图像生成候选检测区域，降低背景误检；PointPainting 把语义信息叠加至激光点云，增强点云特征，改善三维检测性能。决策级融合采用 D-S 证据理论对障碍物检测结果做概率融合，提升复杂光照下检测稳定性。

AVOD、MV3D 等端到端多模态网络自动驾驶场景表现优异，但依赖大量标注数据，算力开销大，不适用于机器人嵌入式平台。ROS 生态下 Cartographer、GMapping 为建图定位提供支撑，darknet_ros 简化 YOLO 算法部署。端侧语音、MediaPipe 手势识别技术不断成熟，多模态交互可将噪声环境指令识别准确率提升 15%以上。

综上，激光-视觉融合与多模态交互已有大量研究，但在统一 ROS 框架下完成二者集成并面向室内机器人开展完整实验验证，仍有进一步研究价值。

3 系统总体设计

3.1 硬件平台

系统以两轮差速移动机器人作为运动载体，底盘配备增量式编码器与六轴 IMU，最大运动速度 0.8m/s，输出里程计与姿态信息。搭载 RPLIDAR A1 二维激光雷达，扫描频率 10Hz，测距 0.15-12m，安装高度 0.8m，减少低矮障碍物盲区；前部安装 Intel RealSense D435i RGB-D 相机，RGB 分辨率 1920×1080，深度分辨率 1280×720，帧率 30fps，镜头向下倾斜 15°，拓展近地面感知范围。选用 NVIDIA Jetson Xavier NX 作为计算单元，AI 算力 21TOPS，搭载 Ubuntu20.04 与 ROS Noetic，保障多传感器同步处理与模型推理；配套 USB 麦克风阵列与扬声器实现语音采集和反馈输出。

3.2 软件架构

系统采用 ROS 分布式节点架构，模块间通过话题通信，整体划分为四层：硬件驱动层完成数据采集与设备控制；感知融合层为核心，实现数据同步、深度融合、目标跟踪、SLAM 及代价地图构建；交互决策层把用户非结构化指令转换为机器人可执行任务；应用执行层完成运动控制与任务执行。依靠 TF 树管理多组坐标系变换，保障感知数据空间一致性。

3.3 坐标系统与外参标定

为实现传感器空间对齐，采用 9×6 角点棋盘格开展分步外参标定，采集 20 组不同位姿样本，提取图像亚像素角点与激光点云簇，通过几何约束求解旋转矩

作者简介：尹江山，男（1976.02），汉，湖北省武汉市，讲师，研究方向：人工智能。

阵 R 和平移向量 t，以静态 TF 变换在 ROS 中发布。经重投影误差检验，激光点投影至图像平均像素误差小于 0.5 像素，满足融合算法精度需求。

4 激光雷达与视觉融合的环境感知方法

4.1 数据时间同步与空间配准

针对激光雷达 10Hz 与相机 30fps 的频率差异，利用 ROS 的 message_filters 工具中 Approximate Time Synchronizer 实现软时间同步，设置 50ms 时间窗口，仅对时间戳差值满足阈值的激光与 RGB-D 图像做配对处理，削弱数据不同步带来的精度损失。空间配准时将二维激光点通过标定得到的外参 R、t 转换至相机坐标系，再结合相机内参矩阵投影到图像像素平面，形成激光扫描线；筛选图像范围内的有效激光点，建立激光测距值与图像像素的映射关系，为后续融合提供深度参考。

4.2 基于 EKF 的深度信息融合

RGB-D 深度存在远距离、反光场景下噪声与缺失问题，而激光雷达测距精度高但点云稀疏，本文采用 EKF 实现两者深度融合。将像素深度作为系统状态 x=d，建立状态方程，分别设置视觉、激光两组观测模型，视觉观测噪声随深度二次增大，激光观测噪声取固定较小值。存在激光投影点的像素采用双观测更新，无激光覆盖区域仅依靠视觉深度。通过 EKF 预测与更新迭代，校正视觉深度偏差，同时依靠视觉稠密信息弥补激光点稀疏缺陷，输出精度与完整性更优的深度图。

4.3 目标检测与三维定位

目标检测选用 YOLOv5s 轻量化模型，借助 Mosaic 增强、CSPDarknet53 主干网络以及 FPN+PAN 结构，面向室内场景重新训练，识别人、椅子、桌子、门、垃圾桶五类目标。通过 TensorRT 做 FP16 量化加速，在 Jetson Xavier NX 上单帧推理可达 45fps。检测完成后剔除框内深度异常点，取有效深度中位数作为目标深度，将检测框中心反投影并经 TF 变换得到目标在机器人底盘坐标系下三维位置。针对人员等动态目标，采用卡尔曼滤波跟踪，维护位置与速度状态向量；连续 5 帧未检测到则删除跟踪对象，节约算力资源。

4.4 融合代价地图构建

在 ROS 原生 costmap_2d 框架下新增语义障碍层插件，把检测目标的三维位置投影至二维栅格，区分标记静态、动态障碍物；结合动态目标运动预测，将未来 1s 可能占据栅格设置为高代价区域。将激光障碍层与语义障碍层融合后输出综合代价地图，交付 move_base 模块，完成机器人路径规划与动态避障。

5 人机交互模块设计

（1）总体架构：双模态同时支持语音与手势交互，流程为：音频采集转文本指令，图像采集转手势指令，经多模态融合决策生成最终任务，规划后输出语音反馈。

（2）语音交互：采用流式语音识别 API，安静环境准确率 95%以上，离线场景内置 PocketSphinx 支持 20 条核心指令，通过关键词槽位提取解析意图，TTS 模块将反馈转为自然语音输出。

（3）手势交互：基于 MediaPipe Hands 检测手部 21 个骨骼点，定义 8 种适配机器人控制的手势指令，采用连续 3 帧识别一致才输出指令的平滑策略，避免单帧误检导致的指令抖动。

（4）多模态指令融合：制定差异化规则：指令一致直接执行，冲突时选高置信度通道并二次确认，语音复杂指令优先级高于手势微调，噪声超 60dB 时自动切换手势优先模式，保证复杂场景交互稳定性。

6 实验与结果分析

6.1 实验环境

所有实验在 8m×6m 的真实室内办公环境中开展，场景内包含桌椅、文件柜等典型静态障碍物，同时引入随机行走的人员作为动态障碍物，环境光照强度约 300lux，覆盖正常办公场景的典型光照条件。实验平台为本文搭建的差速驱动移动机器人，所有算法均在 Jetson Xavier NX 边缘平台上本地运行，无云端算力依赖。实验从目标检测性能、定位建图精度、人机交互性能三个维度开展系统性对照验证。

6.2 目标检测性能

设置三组对照实验：方法 A 为纯视觉 YOLOv5 检测方案，方法 B 为视觉+原生 RGB-D 深度辅助方案，方法 C 为本文提出的视觉+激光-视觉融合深度方案。针对 5 类室内目标各选取 200 张标注测试图像，以 mAP@0.5 作为核心评价指标，实验结果如表 1 所示。

表 1 目标检测 mAP 对比（%）

目标类别	方法 A	方法 B	方法 C
人	89.2	92.5	95.8
椅子	85.6	88.3	91.2
桌子	83.1	86.7	89.5
门	87.4	89.1	91.8

垃圾桶	81.5	84.2	87.6
整体 MAP	85.4	88.2	91.2

注：续表 1。

实验结果表明，本方法的整体 mAP 达到 91.2%，较纯视觉方法提升 5.8 个百分点，较仅使用原生 RGB-D 深度的方法提升 3.0 个百分点。分析可知，激光雷达提供的高精度深度信息，在目标局部遮挡、光照突变的场景下可有效排除虚假检测框，修正检测框的位置偏差。实时性测试结果显示，三种方法的单帧处理时间分别为 22ms、28ms 和 35ms，本文方法的整体处理帧率约 28fps，完全满足移动机器人实时感知的性能要求。

6.3 定位与建图精度

控制机器人沿 30m 预设路径行驶，以 10 个参考点的实际位置对比定位输出，计算 RMSE。实验结果如表 2 所示。

表 2 定位精度对比

评价指标	纯激光雷达	融合方法
x 方向 RMSE(m)	0.082	0.071
y 方向 RMSE(m)	0.095	0.083
平面距离 RMSE(m)	0.125	0.109
航向角 RMSE(°)	2.31	1.96

实验结果表明，融合后机器人的平面定位 RMSE 从 0.125m 降至 0.109m，相对精度提升 12.8%；航向角 RMSE 从 2.31°降至 1.96°，相对精度提升 15.2%。视觉检测输出的门、桌子等语义稳定目标，可作为天然辅助定位标志，有效修正激光雷达在几何特征稀疏场景下的定位漂移。在包含 2~3 名随机行走人员的动态环境中，纯激光雷达方案的定位 RMSE 升至 0.168m，而本文融合方案的定位 RMSE 仅为 0.142m，相对精度提升 15.5%，说明融合语义信息的代价地图可有效区分动态障碍物与静态环境，大幅降低动态目标对 SLAM 建图定位的干扰。在建图质量方面，融合方案不会将短暂停留的动态人员误标记为静态障碍物写入地图，而纯激光雷达方案在人员停留超过 3s 时，极易在地图中生成伪障碍。

6.4 人机交互性能

人机交互实验选取 5 名测试者开展测试，每名测试者朗读 20 条控制指令。安静条件下语音识别准确率为 96.0%，60dB 噪声环境下降至 88.0%，加入语义约束与二次确认机制后，指令理解准确率可达 93%以上。手势交互共采集 400 组有效样本，在 1-3m 交互距离内整体识别准确率 92.5%；握拳、张掌等简单手势识别率高于 95%，OK、挥手手势约 88%，误差来自手部自遮挡和个体手势习惯差异。端到端平均响应延迟 185ms，由语音识别、指令解析规划与底层运动控制三部分构成，该延迟处于用户可接受范围。

6.5 综合场景演示

搭建室内服务综合场景开展全流程测试，用户下达“去会议室”语音指令，机器人确认后启动导航；行进过程遇到行走行人可自动减速避让；抵达目标后接收握拳手势继续运动，最后接收“返回起点”指令自主返航。经过 5 轮重复测试，全部任务顺利完成，未出现碰撞、导航及交互故障，验证系统整体实用性能。

7 结论与展望

本文实现一套基于 ROS 的激光雷达与视觉融合感知交互系统，采用 EKF 松耦合深度融合提升感知性能，利用 YOLOv5s 完成室内目标识别跟踪，构建语义分层代价地图实现避障导航，并开发语音-手势多模态交互模块。实验表明，系统目标检测 mAP 达 91.2%，定位精度提升 12.8%，交互平均延迟 185ms，适合室内服务机器人应用。后续将探究端到端特征级融合网络，改用三维激光雷达完成全三维语义建图，搭建长期语义地图；拓展交互通道，结合大模型优化语言理解，并研究机器人长期自主运行与故障自恢复能力。

参考文献：

[1] 李帅,等.基于 D-S 证据理论的激光视觉融合障碍物检测[J].仪器仪表学报,2019,40(7):123-131.

[2] 陈小雨,李明浩.多模态人机交互在家庭服务机器人中的应用综述[J].计算机工程与应用,2022,58(18):1-12.

[3] Vora S,Lang A H,Helou B,et al.PointPainting:Sequential fusion for 3D object detection[C]//CVPR,2020:4604-4612.

[4] 王坤,刘美华,王耀南。基于 D-S 证据理论的移动机器人多传感器障碍物检测[J].控制与决策,2019,34(8):1665-1672.