Multi-Sensor SLAM: Low-cost LiDAR and Visual Fusion (LLV-SLAM)

Phurithat Tangsripairoje, Meng Wu, Yikuan Xiong

University of Michigan

Motivation

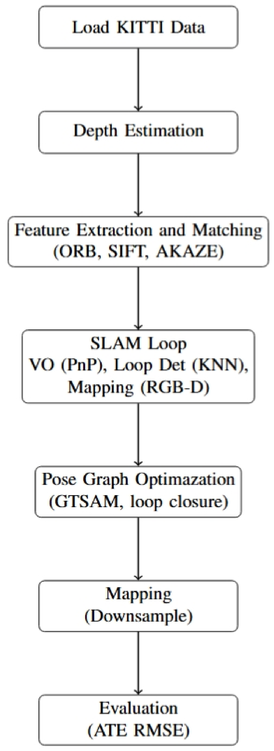
Simultaneous Localization and Mapping (SLAM) is a fundamental technology for autonomous agents, enabling them to localize and build a map of their environment. Visual SLAM offers semantic richness and low cost but struggles in dynamic or low-light scenarios and suffers from scale ambiguity. LiDAR SLAM provides accurate depth and robustness but lacks semantics and is costly.

Multi-modal SLAM leverages the strengths of both sensors, fusing visual appearance with LiDAR geometry for improved localization and mapping in complex environments.

In this work, we propose a **lightweight multi-modal SLAM framework** that fuses RGB images with low-cost **sparse LiDAR** using an adaptive **depth refinement module**. Our method improves depth estimation and tracking robustness for resource-constrained systems and introduces a regression-based strategy to refine stereo depth using sparse LiDAR cues.

SLAM Framework Overview

- **KITTI Loader:** Load stereo image sequences and calibration.
- **Depth Estimation:** Compute disparity-based depth from stereo pairs.
- **Feature Matching:** Detect and match features.
- **SLAM Loop:** Perform VO, loop closure, and mapping.
- **Pose Graph Optimization:** Refine global poses using GTSAM with loop closures.
- **Mapping:** Back-project depth into 3D and apply voxel downsampling.
- **Evaluation:** Compute ATE RMSE against ground truth for trajectory accuracy.



Trajectory Optimization and Mapping

Figure 2. A snippet of the dataset of Seq.0002

We implement an offline stereo visual SLAM system to evaluate trajectory estimation and dense mapping performance under diverse driving conditions. Three KITTI sequences were chosen to represent distinct motion and lighting characteristics:

- 0002: straight path with uniform lighting
- 0011: gentle curves with strong lighting variation
- 0009: sharp turns with moderate illumination changes

The proposed pipeline consists of five key modules: stereo depth estimation, visual odometry using feature matching and PnP, loop closure detection via descriptor matching, pose graph optimization with GTSAM, and dense point cloud reconstruction through RGB-D fusion.

The system supports various feature detectors (ORB, SIFT, AKAZE) and matchers (BF, FLANN), enabling comparative evaluation. To improve robustness in challenging conditions, sparse LiDAR is integrated to refine stereo depth, resulting in better scale consistency and reduced drift.

LiDAR-Assisted Depth Refinement

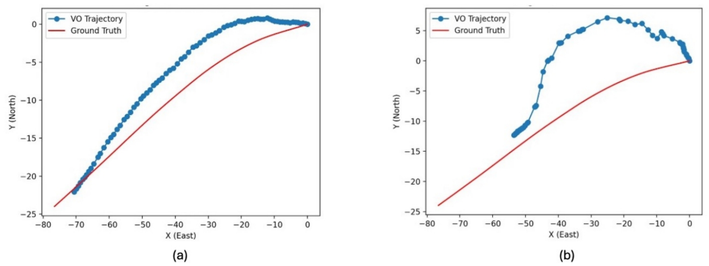


Figure 3. (a): VO of the monocular SLAM with the alignment (b) VO of the monocular SLAM without the alignment

We enhance both stereo and monocular depth predictions using sparse LiDAR measurements to improve robustness in low-texture or poorly lit regions. For stereo inputs, a rule-based fusion strategy selectively replaces or blends depth values when large discrepancies are detected, yielding a refined depth map that combines the density of stereo with the precision of LiDAR.

For monocular inputs, we compute a global scale factor via the median ratio between monocular predictions and valid LiDAR depths. This enables scale-consistent mapping without the need for network retraining.

An optional bilateral filter is applied to smooth sparse areas while preserving edge information. This lightweight fusion pipeline improves depth quality and downstream SLAM performance, allowing for metrically accurate localization and mapping even in visually degraded scenes.

Results and Evaluation

Evaluated on the KITTI dataset using stereo images and sparse LiDAR scans. Compared with baseline stereo SLAM and monocular + learned depth:

- **Improved Trajectory Accuracy:** ATE RMSE significantly reduced post-fusion; best result of **0.54m** on Seq.0002 with AKAZE-BF_crosscheck.
- **Monocular SLAM Correction:** LiDAR-based scale alignment eliminated drift and improved metric consistency.
- **Visual Improvements:** Fused depth maps enhanced point cloud quality, especially in low-texture regions.
- **Robustness:** Fusion preserved tracking under lighting variation (Seq.0011) and sharp turns (Seq.0009).

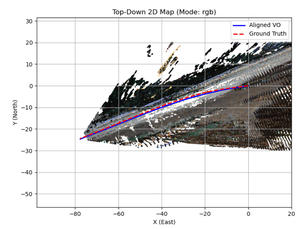


Figure 4. *

(a) Seq.0002 using AKAZE-BF_crosscheck

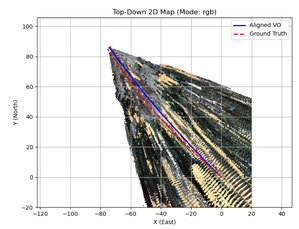


Figure 5. *

(b) Seq.0011 using SIFT-BF_crosscheck

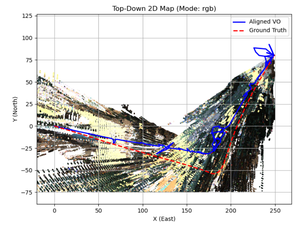


Figure 6. *

(c) Seq.0009 using SIFT-BF_crosscheck

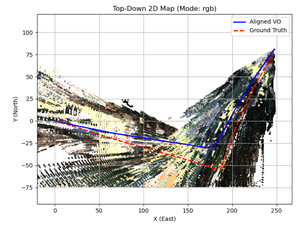


Figure 7. *

(d) Seq.0009 using SIFT-FLANN_KNN

Conclusion and Future Work

This work demonstrates how sparse LiDAR data can significantly enhance depth accuracy in visual SLAM, particularly for monocular systems. Our lightweight, training-free fusion approach improves both trajectory estimation and map consistency without the need for high-end sensors.

Future directions include:

- Learning-based depth–LiDAR fusion with uncertainty modeling
- Online calibration and self-supervised refinement
- Robustness to dynamic objects and environmental changes

References

[1] Claude E. Shannon.
A mathematical theory of communication.
Bell System Technical Journal, 27(3):379–423, 1948.