Atiye Nazari

4041342245

Stereo Depth Estimation and Stereo Visual Odometry on KITTI

Introduction

In this project, I implemented a complete classical stereo vision pipeline including:

- Dense depth estimation from stereo image pairs
- Stereo visual odometry (camera motion estimation over time)

The system was developed and evaluated using the KITTI datasets:

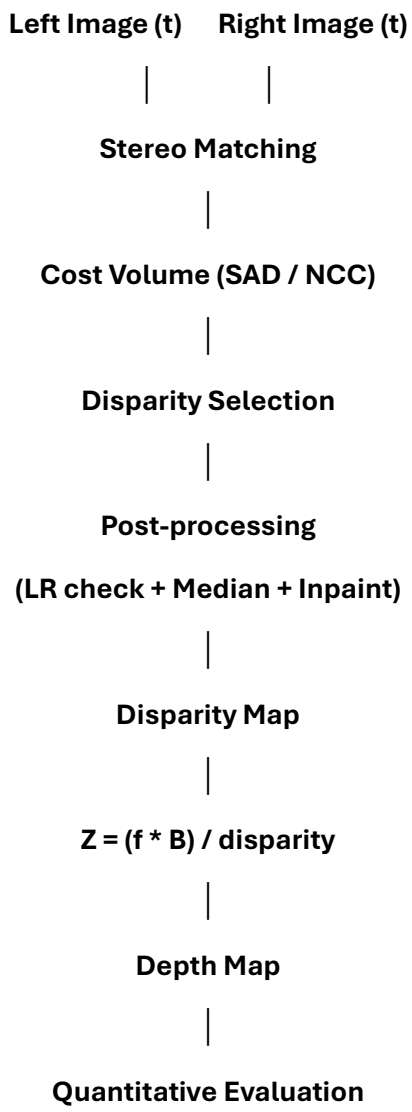
- **KITTI Stereo 2015 (training split)** for depth evaluation
- **KITTI Odometry sequences 00–10** for visual odometry evaluation

All components were implemented using classical computer vision and geometry methods. No deep learning methods were used

Part A) Stereo Depth Estimation

Depth Estimation Pipeline Overview

Pipeline Diagram



Stereo Matching

Matching Cost Functions

I implemented two classical matching cost functions:

SAD – Sum of Absolute Differences

For each pixel and disparity value:

$$SAD = \sum | L(x,y) - R(x-d,y) |$$

- Simple
- Fast
- Sensitive to illumination changes

NCC – Normalized Cross Correlation

$$NCC = \frac{\sum(L-\bar{L})(R-\bar{R})}{\sqrt{\sum(L-\bar{L})^2\sum(R-\bar{R})^2}}$$

- More robust to illumination changes
- Slightly slower
- Performed better for larger windows

Window Sizes (Ablation Study)

I compared two window sizes:

- 7×7
- 11×11

Larger windows:

- Reduce noise
- Improve matching stability
- Slightly reduce fine detail

Post-processing

To improve disparity quality, I implemented:

Left-Right Consistency Check

Invalidates pixels where:

$$| d_L(x) - d_R(x-d_L) | > threshold$$

This removes occlusion mismatches.

Median Filtering

Reduces salt-and-pepper noise.

Hole Filling (Inpainting)

Invalid pixels were interpolated using OpenCV inpainting.

Depth Computation

Depth was computed using:

$$Z = \frac{f \cdot B}{d}$$

Where:

- f= focal length (from KITTI calibration)
- B= stereo baseline
- d= disparity

Both $f$and Bwere extracted from calib.txt

Depth Evaluation (KITTI Stereo 2015)

Metrics used:

MAE (Mean Absolute Error)

$$MAE = \frac{1}{N}\sum | d_{est} - d_{gt} |$$

Bad Pixel Rate (> 3 px)

Percentage of pixels where disparity error > 3 pixels.

Depth Ablation Results

I compared:

Cost	Window	Performance
SAD	7	baseline
SAD	11	improved smoothness
NCC	7	better illumination robustness
NCC	11	best overall stability

Conclusion:

- NCC + larger window performed best.
- SAD is faster but less robust

Failure Cases (Depth)

Observed failure scenarios:

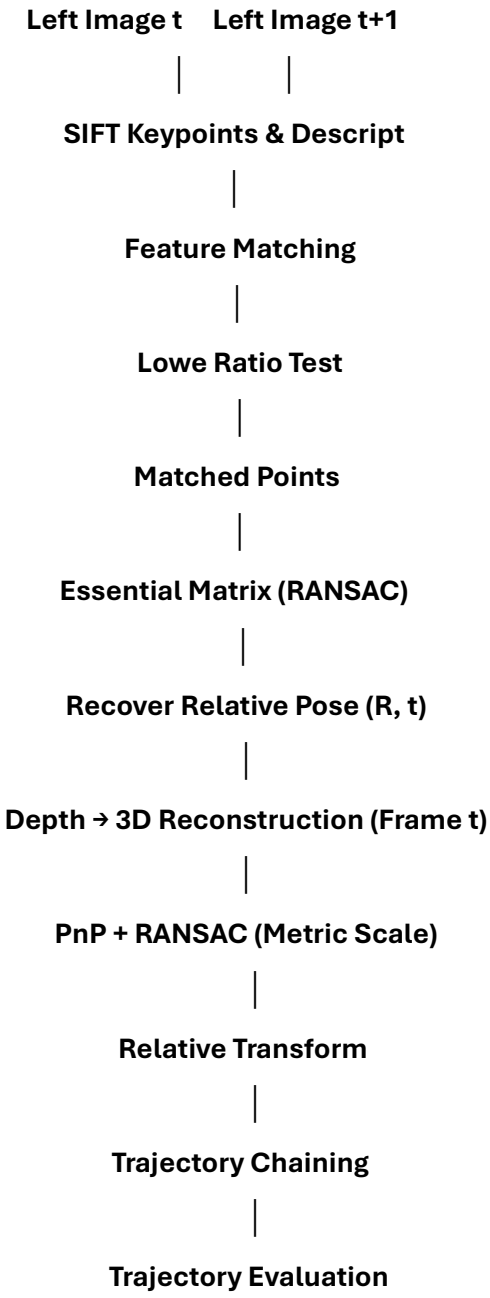
- Repeated textures (building facades)
- Textureless regions (road)
- Occlusions
- Strong lighting differences

These cause ambiguous or incorrect disparity.

Part B) Stereo Visual Odometry

VO Pipeline Overview

Pipeline Diagram



Feature Detection and Matching

- Detector: SIFT
- Descriptor: SIFT (128-dimensional)

- Matcher: BFMatcher (L2 norm)
- Filtering: Lowe ratio test (0.75)

Each frame pair had ~1200 matches after filtering.

Robust Geometry Estimation

I estimated the **Essential matrix (E)** using:

cv2.findEssentialMat(..., method=RANSAC)

RANSAC removes outlier matches.

Relative Pose Recovery

From E, I recovered:

- Rotation R
- Translation direction t(unit scale)

Using:

cv2.recoverPose()
Translation from Essential matrix has unknown scale.

Metric Scale Recovery (Required)

Monocular VO gives translation only up to scale.

To recover real scale:

1. Used stereo depth to convert matched 2D points in frame t into 3D points.
2. Used PnP + RANSAC:

cv2.solvePnPRansac()

This estimates:

- Metric rotation
- Metric translation (real meters)

This step removes scale ambiguity.

Trajectory Chaining

Relative transforms were chained:

$$T_{global} = T_{prev} \cdot T_{rel}^{-1}$$

This produced the full camera trajectory.

VO Evaluation (KITTI Odometry)

Metrics:

ATE (Absolute Trajectory Error)

$$ATE = \sqrt{\frac{1}{N}\sum || p_{est} - p_{gt} ||^2}$$

RPE (Relative Pose Error)

Evaluated at:

- step = 1
- step = 5
- step = 10

VO Results (Example seq 00)

Example output:

- ATE RMSE ≈ 0.18 m
- RPE step=1 ≈ 0.13 m
- RPE step=5 ≈ 0.63 m

The estimated trajectory closely follows the ground truth in short sequences.

VO Failure Cases

- Rapid rotations
- Low texture frames
- Large forward motion (low parallax)
- Incorrect depth in distant regions

Drift accumulates over longer sequences.

Ablation Study (VO)

Compared:

Method	Description
Essential + RANSAC	unit scale
Essential without RANSAC	unstable
PnP + RANSAC (metric)	best performance

Conclusion:
PnP with stereo depth significantly improves trajectory accuracy.

Conclusion

In this project, I implemented a fully classical stereo vision system including:

- Dense depth estimation
- Stereo visual odometry
- Robust geometry with RANSAC
- Metric scale recovery
- Full trajectory reconstruction

All evaluation was quantitative and performed on KITTI datasets.

The system performs well on structured outdoor environments, but struggles in low-texture or highly repetitive scenes.