ECE 281B - Advanced Computer Vision

3D Reconstruction Report

Parth Kulkarni Pratyush Sinha

I. INTRODUCTION

The recovery of 3D structure from 2D image sequences is a long-standing problem in computer vision with wide-ranging applications in robotics, augmented reality, digital heritage preservation, and autonomous systems. The underlying challenge lies in inferring depth and scene geometry from flat visual data captured from varying viewpoints. This project aims to construct a modular and scalable 3D reconstruction pipeline that transforms unordered RGB images into a 3D point cloud and subsequent visualization using modern rendering techniques.

Leveraging principles from multi-view geometry, feature matching, and Structure-from-Motion (SfM), this pipeline incrementally builds up from basic manual correspondence selection to full automated dense scene reconstruction. The entire system is developed over four structured milestones: (1) Manual Feature Extraction and Two-View Geometry, (2) Automated Feature Detection and Geometric Verification, (3) Camera Pose Estimation and Sparse/Dense Reconstruction, and (4) 3D Scene Visualization using NeRF.

Throughout the process, the Kavli dataset—comprising over 400 high-resolution images of UCSB's Kavli Institute building, was used to evaluate pipeline robustness under varying lighting, scale, and camera motion conditions. Each phase of the project introduces new techniques and tools, from fundamental matrix estimation using manually labeled points to modern descriptor-free matching (LoFTR) and real-time visualization. This project reinforces foundational concepts in 3D vision but also provides hands-on experience with state-of-the-art reconstruction frameworks such as COLMAP and NeRF rendering.

The remainder of this report is organized as follows: Section 2 covers Milestone 1 and its emphasis on geometric understanding through manual correspondences. Section 3 details automated keypoint extraction and match pruning. Section 4 discusses sparse and dense 3D reconstruction with camera pose estimation. Section 5 presents the final visualization using NeRF. Section 6 includes reflections and challenges, and Section 7 concludes with a summary and outlook.

II. Milestone 1: Manual Feature Extraction and Two-View Geometry

The goal of this phase was to develop a fundamental understanding of multi-view geometry. This included computing homographies for planar surfaces, and estimating fundamental matrices for epipolar geometry.

A. Methodology

Initially, manual correspondences across three images were established, where multiple sets of corresponding points were carefully identified. For homography estimation, 15 manually selected points were chosen on clearly identifiable planar surfaces between each image pair (Image 1–2, Image 2–3, and Image 1–3). These points were used as input to compute homography matrices using a DLT (Direct Linear Transform) method. Validation involved transforming test points excluded from the initial estimation and comparing their positions to manually marked ground-truth points in the target images.

Subsequently, to estimate the fundamental matrix, additional points lying outside the planar regions were selected. An 8-point algorithm was used for this computation, generating fundamental matrices between pairs of images. Validation involved drawing epipolar lines corresponding to points from one image onto the paired image, visually assessing how closely these lines passed through manually identified corresponding points. Cross-view consistency would be further validated by computing intersections of epipolar lines derived from two image pairs, but our images were very much related to translation of 2 cameras and hence the Epipolar lines were nearly parallel to each other and did not intersect in the image area.

B. Results

The homography matrices accurately mapped selected test points with pixel errors ranging between approximately 3 and 20 pixels which is very less in proportion to a HD quality image, signifying strong alignment with manually identified correspondences. Fundamental matrices showed precise alignment, with manually marked points closely aligning with corresponding epipolar lines, confirming accurate estimation.

Fig. 1. Manual correspondence points in image 1- Marked in Blue and the points mapped in image 2 - Marked in Red

Fig. 2. Epipolar lines in Image 2 corresponding to the 4 validation points from Image 1

III. Milestone 2: Automated Feature Extraction and Matching

This milestone introduced automated keypoint detection and descriptor matching using modern pipelines to reduce human effort and enhance matching consistency.

A. Methodology

We utilized two automated pipelines:

1. SuperPoint + LightGlue: SuperPoint was used to detect repeatable keypoints that were invariant to rotation, scale, and lighting. LightGlue then performed matching using spatial reasoning and global consistency via graph neural networks.
2. SIFT + AdaLAM: This classical pipeline utilized SIFT's robustness to scale and orientation for detection, and AdaLAM's local affine model verification for accurate match filtering.

RANSAC was applied to all matches to remove geometric outliers and retain only those that satisfied the two-view geometry constraints. Homographies and fundamental matrices were then re-computed using these pruned sets.

B. Results

Visualisations included pre- and post-pruned point matches for all image pairs, as well as epipolar lines overlaid on the target images. The automated matches offered higher correspondence coverage (even post RANSAC) and precision compared to manual selection.

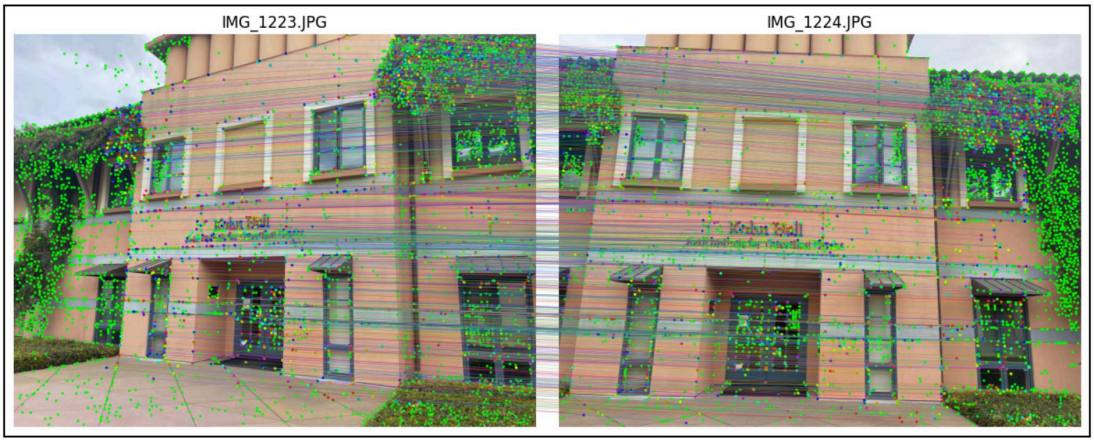

Fig. 3. Pre-pruned mapped points

Fig. 4. RANSAC-pruned and downsampled for visualisation in image 1 with image 2

Fig. 5. Subset of epipolar line matches shown for visualisation in image 1 with image 2

In particular, SuperPoint + LightGlue showed superior performance across all image pairs, especially in cases of wide baselines and motion blur, whereas SIFT + AdaLAM performed faster but occasionally suffered from over-pruning and failed to handle repetitive textures or lighting shifts.

IV. Milestone 3: Sparse and Dense Reconstruction

Estimate camera parameters and reconstruct both sparse and dense 3D scenes using verified point matches.

A. Methodology

Verified correspondences from Milestone 2 were input into the COLMAP pipeline using COLMAP GUI to execute incremental Structure-from-Motion (SfM). The SfM process included estimating camera intrinsics and extrinsics for each image. Feature triangulation was performed across views to obtain a sparse 3D point cloud.

For dense reconstruction, the sparse model and 428-image dataset were processed using the COLMAP GUI. Multi-view stereo and depth fusion generated dense geometry, exported as a .ply file for visualization.

Two matchers were used for comparative evaluation:

1. SuperPoint + LightGlue: Resulted in a dense and well-connected sparse point cloud with more registered images and fewer reconstruction holes.

2. SIFT + AdaLAM: Was faster, but suffered from lower coverage due to insufficient triangulation and fewer surviving inliers.

Bundle adjustment was automatically performed to minimize reprojection error. The reconstructed scene was exported in the form of a .ply file, along with cameras.txt, images.txt, and points3D.txt outputs for analysis.

B. Results

- Sparse reconstruction using SuperPoint + LightGlue produced dense point clouds with high image registration success.
- SIFT + AdaLAM yielded fewer triangulated points due to aggressive pruning.
- Dense reconstruction covered large surface areas with high fidelity and minimal holes in overlapping regions.

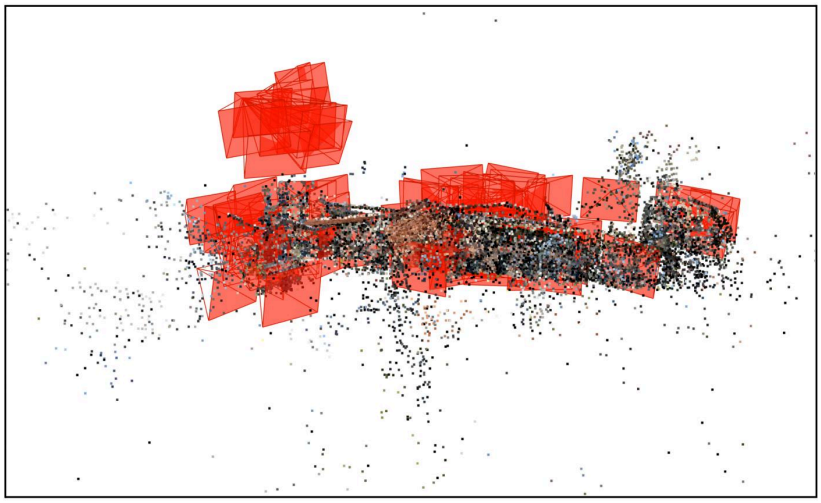

Fig. 6. Sparse reconstruction done using H-Loc pipeline and visualised in COLMAP GUI

Video Link: Milestone3_DenseReconstruction

V. Milestone 4: Neural Radiance Fields Visualisation

To render the reconstructed scene using NeRF techniques and explore visual realism and fidelity under different splat size and transparency parameters.

A. Methodology

The COLMAP outputs (images.txt, cameras.txt, points3D.txt) were parsed to extract image file paths and camera parameters. These were converted into a transforms.json file following NeRF's input format. This included converting the camera intrinsics and camera-to-world extrinsics into the required structure.

We then used the prepared data to train a NeRF model using an open-source NeRF implementation. The training script was run on a GPU-enabled system. During training, the model used the image-pose pairs to optimize a volumetric scene representation. Training was continued until the loss stabilized to below 0.01. Checkpoints were saved periodically for later evaluation.

B. Results

- NeRF rendered smooth and coherent geometry with realistic shading and lighting effects.
- View synthesis was continuous and visually pleasing even for interpolated viewpoints.
- Regions with high image overlap and consistent lighting contributed to the best-quality outputs. Sections of the building which had a lesser number of images in the dataset (dataset of some 250 images coming from model 0 of COLMAP) showed slightly worse rendering quality.

Video Link: Milestone4_NeRF

VI. Observations

Throughout the four milestones, we observed a consistent dependency between the accuracy of earlier geometric computations and the quality of downstream 3D reconstruction and rendering. In Milestone 1, manually selected correspondences were effective for learning purposes but introduced human errors and limited scalability. The resulting homographies and fundamental matrices were reasonably accurate but suffered in non-planar or repetitive-texture scenarios. Transitioning to automated feature extraction in Milestone 2 significantly improved coverage and precision, especially with the SuperPoint + LightGlue pipeline. However, we also noted challenges with over-pruning in AdaLAM and sensitivity to texture and lighting inconsistencies in SIFT-based methods.

In Milestone 3, Structure-from-Motion using SuperPoint + LightGlue provided better image registration and better sparse point clouds compared to SIFT-AdaLAM pipeline. Dense reconstruction in COLMAP using LoFTR further emphasized the importance of feature robustness and image overlap, as well-distributed viewpoints led to cleaner surface continuity. Finally, Milestone 4's NeRF-based rendering effectively converted this geometric output into photorealistic visualizations. Successful NeRF training depended heavily on accurate camera poses and scene coverage, reinforcing the need for precision in earlier stages. Across the pipeline, strong geometric verification, consistent keypoint detection, and high viewpoint overlap proved to be the most critical factors for reliable 3D reconstruction.

VII. Reflections & Challenges

Several potential improvements and challenges were identified during the development and execution of this pipeline. One of the main opportunities lies in adopting better matchers. Transformer-based matchers like MAST3R can offer end-to-end feature detection and matching, increasing robustness under extreme viewpoint changes. Such methods outperform traditional pipelines in wide-baseline scenarios and could potentially replace both SuperPoint and LightGlue.

Another challenge lies in the depth map fusion process during dense reconstruction. Tuning photometric consistency thresholds and parameters such as window_radius, filter_min_num_consistent, and

filter_min_triangulation_angle can lead to improved completeness in areas affected by lighting variations or limited triangulation angles. This would enhance surface continuity and reduce reconstruction noise.

In terms of rendering, while NeRF provided high-quality outputs, further improvements can be achieved by exploring Instant-NGP (Neural Graphics Primitives). This technique allows faster training and finer control over rendering parameters, such as voxel grid resolution, to achieve better scene fidelity with reduced computational overhead. These avenues highlight areas for future exploration to improve efficiency, robustness, and scalability of the reconstruction pipeline.

VIII. Conclusion

This project successfully implemented an end-to-end 3D reconstruction pipeline, beginning with manual correspondence and advancing to automated feature matching, Structure-from-Motion, and NeRF-based rendering. Each stage built upon the previous one, contributing to the final goal of generating a high-quality 3D scene from 2D images.

The integration of robust feature extraction techniques, geometric verification, and accurate camera pose estimation proved essential for both reconstruction accuracy and rendering fidelity. This pipeline lays the groundwork for future enhancements such as real-time NeRF rendering, semantic mapping, or application to more complex and dynamic scenes.