3D Scene Reconstruction From Multi-View RGB Images Using a Monocular Camera

1st Hariharan Sureshkumar
Computer Science
Northeastern University
Boston, MA, USA
lnu.harih@northeastern.edu

2nd Bruce Maxwell
Computer Science
Northeastern University
Seattle, WA, USA
b.maxwell@northeastern.edu

***Abstract*—This project explores the problem of reconstructing a 3D point cloud of a real-world object using only RGB images captured with a monocular smartphone camera. Without relying on depth sensors or external SLAM systems, I implemented a complete 3D reconstruction pipeline from scratch in C++. The pipeline begins with intrinsic calibration of the camera, followed by feature extraction using SIFT, matching using FLANN, and multi-view pose estimation through fundamental and essential matrix computations. The system chains poses= across views and perform triangulation to obtain a sparse point cloud. To improve the quality of reconstruction, I implement global pose optimization using bundle adjustment through the Ceres Solver. Finally, we experimented with depth map fusion using monocular depth estimation (MiDaS) to generate dense point clouds. I attempt to show both qualitative and quantitative results, demonstrating the difficulty and progress in building a custom 3D reconstruction system from first principles.**

***Index Terms*—Structure from Motion, 3D Reconstruction, Bundle Adjustment, Pose Estimation, Monocular Depth, Multi-view Geometry,bCeres Solver.**

I. Introduction

3D scene reconstruction is the act of establishing a spatial representation of an object or environment from multiple 2D images. Classical 3D reconstruction pipelines utilize LiDAR or stereo camera systems; however, reconstructing detailed 3D models from monocular RGB images is especially interesting due to its ubiquity. This project explores the viability of running this type of pipeline using only an ordinary smartphone camera. The pipeline emulates the functionality of Structure from Motion (SfM) and Multi-View Stereo (MVS) algorithms. It was our intention to build every component of the pipeline from the ground up and progressively enhance the pipeline with theoretical insights and optimization techniques like bundle adjustment. This also gives us an appreciation of what it takes to build systems like COLMAP or OpenSfM from the ground up.

II. Related Work

- "Structure-from-Motion Revisited" – Johannes L. Schönberger and Jan-Michael Frahm, CVPR 2016 This paper revisits core components of SfM and proposes improvements for robustness and scalability. We used this paper as a reference for the design of pose chaining and incremental reconstruction steps.
- "Multi-view Stereo Revisited" – Yao et al., CVPR 2019 This work demonstrates how better depth map fusion and patch matching strategies can generate highly detailed dense reconstructions. This inspired our later attempt to switch from sparse triangulation to dense depth estimation using MiDaS.
- "A Method of 3D Reconstruction from Image Sequence" – Hata and Fukunaga, ACCV 2016 This paper outlines a method of reconstructing 3D point clouds by chaining pose estimations and triangulations from multiple images. We followed this paper closely for designing our global pose chaining strategy.

III. Methods

The 3D reconstruction pipeline developed in this project is fully modular and implemented from scratch in C++. The pipeline is composed of the following stages:

A. Image Acquisition and Calibration

Images of the object were captured using an iPhone from various angles. To ensure accurate geometric reconstruction, camera calibration was performed using a checkerboard pattern. OpenCV was used to estimate the camera intrinsic matrix $\mathbf{K}$ and distortion coefficients. All input images were undistorted prior to further processing using these parameters. This step ensures that the radial and tangential distortions introduced by the lens do not interfere with later geometry estimation.

B. Feature Detection and Matching

We employed the SIFT (Scale-Invariant Feature Transform) algorithm for robust detection of keypoints and computation of descriptors. Keypoints are selected based on scale-space extrema in the image, and each keypoint is described using a 128-dimensional feature vector. To establish correspondences between views, we used the FLANN (Fast Library for Approximate Nearest Neighbors) matcher with Lowe's ratio test to filter out ambiguous matches. This step ensures reliable point correspondences across images.

C. Camera Pose Estimation

Given the matched feature pairs, we estimated the relative pose between consecutive images. First, the fundamental matrix was computed using the 8-point algorithm with RANSAC to handle outliers. Using the known intrinsic matrix $\mathbf{K}$, we computed the essential matrix and decomposed it into the relative rotation $\mathbf{R}$ and translation $\mathbf{t}$ using OpenCV's `recoverPose()`. This provided the transformation from one camera frame to the next.

D. Pose Chaining

Instead of treating each pose estimation independently, we implemented pose chaining to maintain a consistent global coordinate system. Starting from the identity pose, each subsequent pose was chained relative to the previous one using matrix multiplication. Specifically, the global rotation and translation were updated as:

$$\mathbf{R}_w^{(i+1)} = \mathbf{R}_{\text{rel}}^{(i)} \cdot \mathbf{R}_w^{(i)}, \quad \mathbf{t}_w^{(i+1)} = \mathbf{R}_{\text{rel}}^{(i)} \cdot \mathbf{t}_w^{(i)} + \mathbf{t}_{\text{rel}}^{(i)}$$

This chaining improved the global consistency of estimated camera positions.

E. Triangulation

Using the projection matrices constructed from the chained poses and intrinsic matrix $\mathbf{K}$, we triangulated the 3D location of matched keypoints using OpenCV's `triangulatePoints()`. The resulting homogeneous coordinates were converted to Euclidean space. We filtered points based on depth and outlier thresholds to reduce noise and remove invalid geometry.

F. Bundle Adjustment

To improve the accuracy of the estimated 3D structure and camera poses, we implemented global bundle adjustment using the Ceres Solver. A custom reprojection error cost function was defined, and we jointly optimized all camera poses and 3D point coordinates. This step significantly reduced reconstruction drift and enhanced point cloud compactness. The optimization minimized reprojection errors over all views:

$$\min \sum_{i=1}^{N} \|\mathbf{p}_i - \pi(\mathbf{R}_i, \mathbf{t}_i, \mathbf{X}_j)\|^2$$

where π is the projection function defined by the camera intrinsics.

G. Depth Map Fusion (Experimental - Still in progress)

As an experimental extension, we attempted dense reconstruction using the MiDaS model for monocular depth estimation. Each RGB image was passed through MiDaS to obtain a per-pixel depth map. These depth maps were unprojected to 3D using the intrinsic parameters and combined into a global point cloud using Open3D. While this method provided visually interesting results for certain structures like a car's bonnet, it was sensitive to lighting conditions and object textures, and it failed for flat or textureless surfaces like building facades.

IV. Experiments and Results

The proposed 3D reconstruction pipeline was tested on multiple image sequences captured using an iPhone camera. The experiments included both indoor scenes with a miniature cityscape object and outdoor scenes of real urban settings involving parked vehicles and buildings captured at night. The objective was to evaluate how well a sparse reconstruction pipeline, built entirely from scratch, could generalize across different types of visual complexity and structure.

Fig. 1. Example of the two images used for this experiment.

A. Sparse Reconstruction with Pose Chaining

Initially, the 3D reconstruction pipeline utilized relative pose estimation between consecutive image pairs. While this allowed us to estimate camera motion and triangulate 3D points, it quickly became clear that utilizing only pairwise transformations caused huge geometric inconsistencies. As seen in Fig. 2, each new pose was only defined relative to its immediate predecessor, without being anchored to a global frame. Therefore, even tiny errors in translation or rotation began to add up as we got more image pairs.

This compounding error manifested as "structural drift" in the rebuilt point cloud. Triangulated points warped along curved paths, resulting in slanted lines rather than uniform shapes. The surfaces of the objects appeared smeared, and the scene lacked any discernible depth or structure. This made recovery of the actual geometry of the objects challenging, particularly in scenes with low parallax or uniform textures, where feature matching is already a challenge to begin with. This accumulation of error is known as drift and can severely degrade the geometry of reconstructed scenes.

To resolve this, we implemented *pose chaining*, which accumulates the rotation and translation transformations across views with respect to a fixed world origin. This change alone yielded a notable improvement in geometric consistency. Fig. 3 shows how the structure began to compact into denser, object-aligned clusters rather than elongated forms.

B. Bundle Adjustment Refinement

I also attempted to apply the global *bundle adjustment* using the Ceres Solver. This step jointly refines the camera

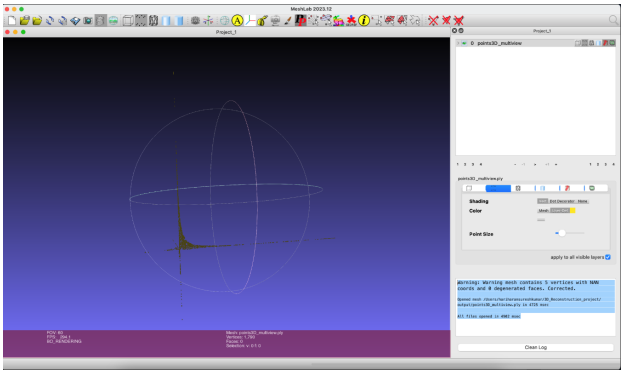

Fig. 2. This highlights the impact of uncorrected relative pose estimation.

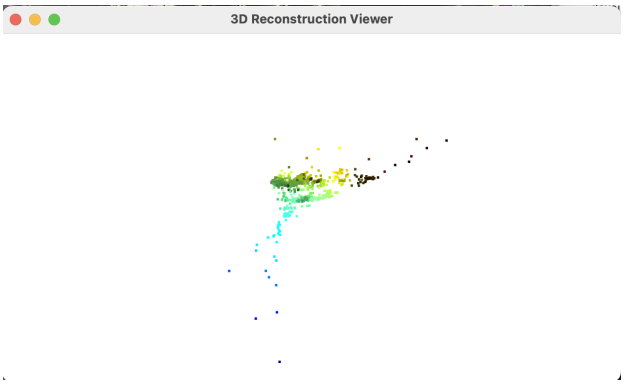

Fig. 3. Improved structure with pose chaining. The 3D points form a much denser and visually plausible cluster.

extrinsics and the 3D point cloud by minimizing a custom reprojection error function. The optimization resolved scale ambiguity issues and improved alignment across views. The improvement is evident when comparing the pre and post-bundle adjusted outputs, as seen in Fig. 4.

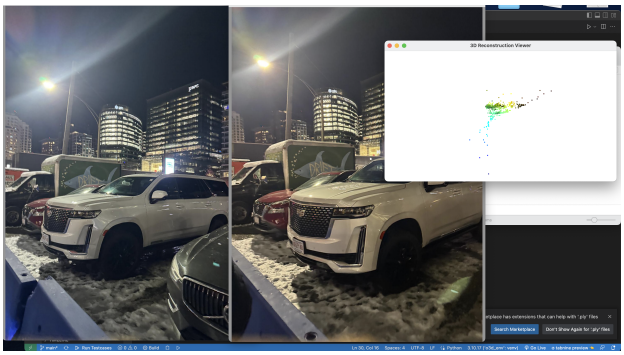

Fig. 4. Post-bundle-adjustment result showing surface-like mesh formation on car bonnet. The surface consistency is due to bundle adjustment refinement.

C. Real-World Outdoor Scene: Cars and Buildings

A visually compelling test case was the night-time capture of parked cars with urban high-rises in the background. Feature matching (Fig. 5) was rich due to the texture on the cars but weaker for the buildings, which had repetitive patterns (glass windows) and low parallax. Despite the sparse reconstruction, Fig. 3 shows that the bonnet and windshield of the white SUV began forming surface-aligned point clusters. However, the buildings were largely reconstructed as vertical point planes due to a lack of depth cues and consistent textures.

Fig. 5. Matched feature points across a car and building sequence using SIFT + FLANN.

D. Failure Case: Metallic Miniature Cityscape

We also tested our pipeline on a complex indoor object: a metallic miniature cityscape featuring intricate geometry and highly reflective surfaces. Although the initial feature detection and matching stages produced dense correspondences between views (as seen in Fig. 6), the final 3D reconstruction was noticeably poor and fragmented. The output point cloud exhibited scattered and incoherent points with no clear indication of the object's underlying shape or structure. This failure was not due to a lack of keypoints, but rather the difficulty in preserving consistent features across views for such a reflective and geometrically complex object (as seen in Fig. 7)

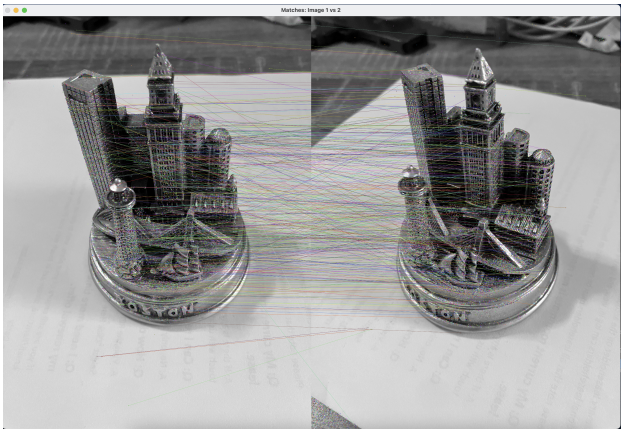

Fig. 6. Dense feature matches across the miniature cityscape object using SIFT.

The problems can be attributed to several compounding factors. First, the reflective surfaces introduced specular highlights that varied significantly with viewpoint, confusing the

SIFT detector and resulting in unstable keypoints. Second, the capture setup had limited parallax, meaning the camera movement between shots was minimal, leading to poor triangulation baselines. Third, the repetitive and fine-scale architecture of the cityscape model likely caused mismatches, as many parts looked similar but occupied different positions in space. These challenges highlight the limitations of purely RGB-based, feature-driven SfM pipelines when dealing with scenes that lack photometric and geometric stability. For such scenarios, robust reconstruction may require incorporating additional modalities like active depth sensing or multispectral imaging.

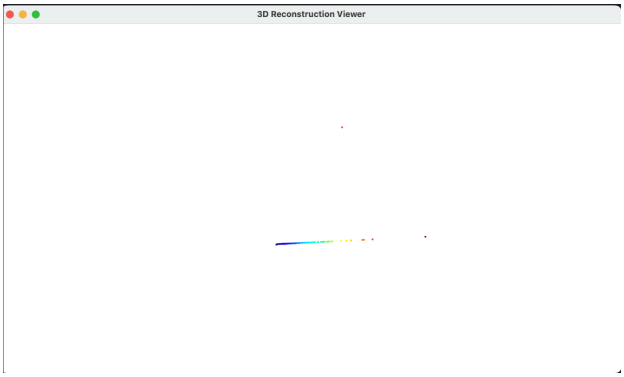

Fig. 7. Miniature cityscape object used in reconstruction tests. Highly metallic and textured surface caused poor results.

E. Comparison with COLMAP and Observations

Unlike COLMAP, which often requires 3 images in minimum to produce a high-quality sparse reconstruction, our pipeline could generate coarse structures with 10–15 well-framed images. However, this compact dataset size comes at the cost of fine detail, especially in flat or repetitive regions. COLMAP also integrates dense multi-view stereo, whereas our pipeline only attempts depth fusion via MiDaS experimentally.

Overall, our results demonstrate that pose chaining and bundle adjustment are essential for any practical pipeline, and even a sparse approach can begin to reveal surface geometry under favorable conditions.

F. Lessons and Limitations

- Curved and feature-rich objects yield higher reconstruction fidelity.
- Buildings or flat textures require more views for robust matching.
- Pose chaining is essential to maintain global consistency.
- Bundle adjustment substantially improves the realism and compactness of reconstructed models.

Unlike COLMAP, which typically needs less no. of images to recover large structures, our goal was to recover a sparse yet structurally meaningful point cloud from only 10–15 input images. The results highlight both the difficulty and feasibility of achieving this goal using a fully custom pipeline from scratch.

V. Discussion and Summary

Developing a complete 3D reconstruction pipeline from scratch without any pre-existing tools like COLMAP was a monumental task that demanded a good understanding of computer vision concepts and meticulous attention to implementation details. Each module such as camera calibration, feature detection, pose estimation, triangulation, and bundle adjustment had to be separately tested. The project was an all-course learning experience, illustrating the complexity and coupling of structure-from-motion systems.

Early reconstruction efforts focused on the importance of pose chaining to maintain global consistency. Initially, relative poses between images in pairs were treated separately, resulting in huge drift and misalignment of the final point cloud. The triangulated structures appeared elongated along arbitrary directions, so any useful geometry was difficult to interpret. With the inclusion of pose chaining, the system began building meaningful 3D shapes by incrementally projecting camera poses onto a global frame.

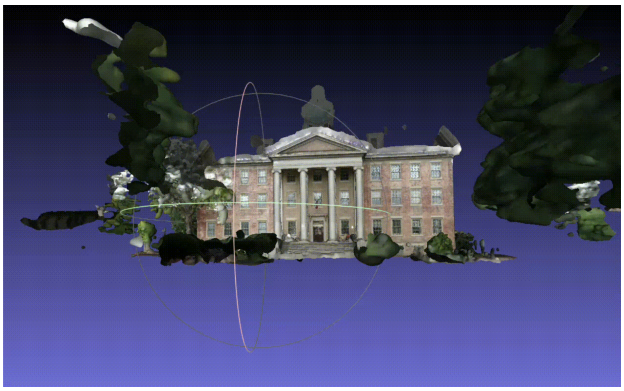

Fig. 8. Colmap sample

Bundle adjustment with the Ceres Solver significantly enhanced the reconstruction. Simultaneous optimization of camera poses and 3D point positions assisted us in resolving scale ambiguities, reducing reprojection errors, and tightening structure. Nevertheless, this optimization process developed its own drawbacks, particularly with respect to data structuring, convergence behavior, and computational cost.

Scene arrangement and visual features had a dramatic influence on system performance. Those objects with textured and curved geometry, such as cars, generated clean point clouds. Planar, repetitive, or highly specular surfaces such as windows or metallic miniatures proved difficult with sparsely or ambiguously located keypoints. The night-time scenes also made it even harder for feature detection using the introduction of blur, inhomogeneous exposure, and noise.

In comparison, COLMAP always provided more precise and denser reconstructions on various datasets. As shown in Fig. 8 However, the speed and elegance of COLMAP are achieved at the cost of significant internal optimizations and heuristics that mask a good proportion of the low-level processing. In contrast to this, implementing each step from scratch took

us fine-tuning, debugging, and proofs of correctness. Whereas the final outputs from our pipeline were less dense and more visually rich, they were rich in important interpretability and transparency making it ideal for research and educational uses.

Overall, this project illustrates the promise and difficulty of building a 3D reconstruction pipeline from scratch. With limited resources and insurmountable challenges, the system was successful in generating sparse 3D maps of real scenes, and it represents an important milestone on the path towards understanding larger, more complex vision systems.

ACKNOWLEDGMENT

I would like to extend my appreciation to Professor Bruce Maxwell for his guidance and encouragement during the entire course and project duration. His comments and recommendations shaped the course of the study and ensured clarity in the implementation. Additional thanks to the Northeastern University Robotics Program for the learning environment and equipment necessary to explore computer vision and 3D reconstruction. The author also thanks the OpenCV, Open3D, and Ceres Solver open-source communities whose libraries were of immense help to this project.

REFERENCES

[1] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 8, pp. 1362–1376, 2010.

[2] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2003.

[3] C. Wu, "Towards linear-time incremental structure from motion," in Proc. Int. Conf. 3D Vision (3DV), 2013, pp. 127–134.

[4] H. Schoenberger and J.-M. Frahm, "Structure-from-Motion Revisited," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4104–4113.

[5] S. Agarwal, K. Mierle, and Others, "Ceres Solver," [Online]. Available: http://ceres-solver.org

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016.

[8] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.