RBE/CS 549 Computer Vision Project 2 Phase 1 Structure from Motion

Puneet Shetty
MS in Robotics
Worcester Polytechnic Institute
Email: ppshetty@wpi.edu

Using 1 Late Day

Edwin Clement
MS in Robotics
Worcester Polytechnic Institute
Email: eclement@wpi.edu

Using 1 Late Day

***Abstract*—In this project, the Structure from Motion (SfM) method of computer vision techniques is implemented for simultaneous camera pose estimation and 3D scene reconstruction. SfM is able to construct point cloud-based 3D models that are similar to those made by LiDAR technology by analyzing a set of 2D photos. In order to determine the relative 3D poses of objects using stereo pairs, the method depends on the concepts of stereoscopic photogrammetry, triangulation, perspective-n-points, RANSAC, Epipolar Geometry, and Bundle adjustment. The application of SfM in 3D reconstruction is demonstrated in this study, along with its potential for use in conjunction with other deep learning techniques like Neural Radiance Fields (NeRF).**

***Index Terms*— RANSAC, Triangulation, Perspective-n-Points, Bundle Adjustment, Visibility Matrix, Structure from Motion**

I. PHASE 1: CLASSICAL STRUCTURE FROM MOTION

In this phase, we used traditional techniques to recreate a three-dimensional scene using only the camera's inherent properties and its photos. Structure from Motion, or SfM, is the algorithm used for this purpose, and its steps are as follows:

1) Feature Mapping and RANSAC using the Fundamental Matrix.
2) Estimating the Essential Matrix.
3) Camera Pose Estimation and Cheriality Condition
4) Linear & Non-Linear Triangulation
5) Linear & Non-Linear Perspective-*n*-Points
6) Building the Visibility Matrix
7) Perform Bundle Adjustment

Below are the sections that provide an overview of this strategy that were used.

A. Dataset & Feature Extraction

The data provided consists of five images of Unity Hall at WPI. The Samsung S22 Ultra's primary camera took the five pictures at f/1.8, ISO 50, and 1/500 sec. This camera is calibrated using the Ran-Tan Model, which has two radial parameters and one tangential parameter. The photos have been resized to 800 x 600 pixels and have undergone distortion correction. A good feature is still necessary for the proper operation of a computer vision algorithm. A powerful feature descriptor for structure of motion problems is SIFT. The data provided: Images of the matching.txt files for each of the five Unity Hall photos. There are a total of five pictures and four ”txt” files. Features: (the number of feature points in the i^{th} image; the specification of matches between images in the following row depending on the feature location of an i^{th} image; and (J^{th} feature matches as a percentage) in every row ($u_{current}$ image), ($v_{current}$ image), (image id), and (u_{image} id image) (v_{image} id image) stand for the values of the feature coordinates. These values need to be extracted from the ”.txt” file.

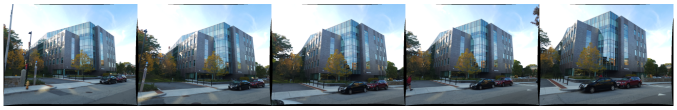

Fig. 1. Images of Unity Hall given

B. Fundamental matrix based feature filtering.

Data becomes noisy after the SIFT feature descriptor, so RANSAC is used with the Fundamental matrix that contains as many Inliers as possible. We use the normalized 8-points approach to get the fundamental matrix. Since correspondence centers of points and epipolar lines don't always match, we normalize it. We compute the fundamental matrix using these normalized points, and then we retrieve the original fundamental matrix. F can have full rank, or 3, due to noise in correspondances, but we have to decrease it to rank 2 by setting the final diagonal element's value to zero, which is how we get the epipoles. Nevertheless, to understand what a fundamental matrix is, we must first understand epipolar geometry. The epipolar geometry is the intrinsic projective geometry that divides two points of view. It simply depends on the relative position and the internal characteristics (K matrix) of the cameras, not the scene structure.

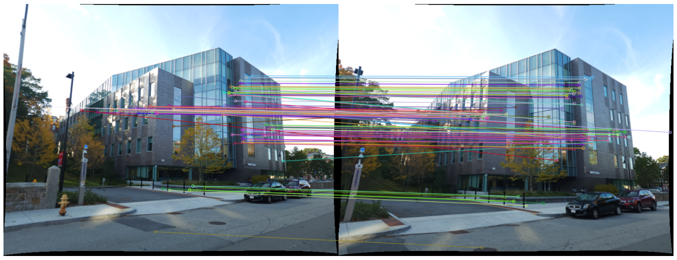

Fig. 2. Inliers between Images 1 & 2

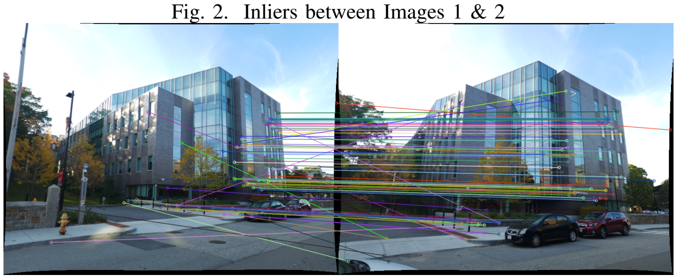

Fig. 3. Inliers between Images 1 & 3

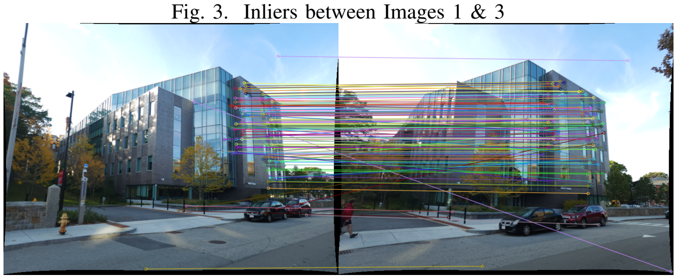

Fig. 4. Inliers between Images 1 & 4

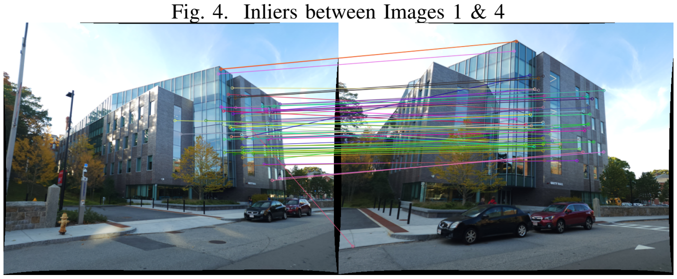

Fig. 5. Inliers between Images 1 & 5

Fig. 6. Inliers after RANSAC

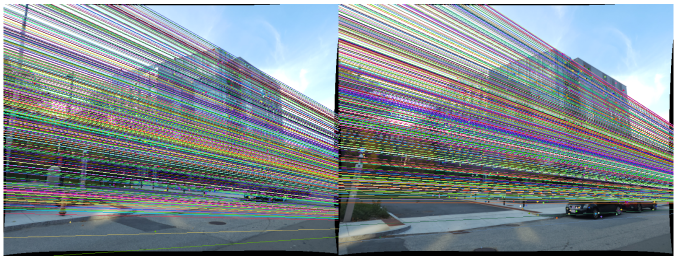

Fig. 7. Epilines for images 1 & 2

C. Estimation of the Essential Matrix

Comparative lens Poses between two images must be determined using the K matrix, which contains the intrinsic values of the camera, and the Fundamental matrix that was previously computed. SVD is used to construct and decompose the essential matrix. Due to this, its diagonal elements are once more enforced to 1,1,0. The relative camera stances between the two perspectives are therefore provided to us. The key matrix is another 3*3 matrix with some extra features that connects the right spots, assuming the cameras follow the pinhole paradigm (unlike F).

$$E = K^T F K$$

D. Camera Pose Estimation & Cheriality Condition

Using SVD, the E matrix is broken down into three rotational, three translation, and six DOF as the camera posture. Given:

$$E = UDV^T \quad \text{and} \quad W = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The four configurations can be written as:

$$\begin{aligned} C1 &= U(:,3) \quad &\text{and} \quad R1 &= UWV^T \\ C2 &= -U(:,3) \quad &\text{and} \quad R2 &= UWV^T \\ C3 &= U(:,3) \quad &\text{and} \quad R3 &= UWV^T V^T \\ C4 &= -U(:,3) \quad &\text{and} \quad R4 &= UWV^T V^T \end{aligned}$$

In this instance, the camera's center is C, and its rotation is R. By taking two camera postures and employing point correspondence, we use linear triangulation to locate the X (3D-point) in the world. This is what we perform for every camera pose to identify the X (3D) point with a positive Z value in front of the camera. We refer to these as depth positivity limitations. By eliminating the disambiguity, our goal is to determine the unique camera stance out of 4. To do this, use the cheirality criteria, which state that the reconstruction points should be in front of the cameras and that $r_3(X - C) > 0$ where r_3 is the third row of the rotation matrix (z-axis of the camera).

E. Triangulation

To obtain four sets of 3D world points, the matching points can be triangulated using each of the four camera postures. The final camera location and orientation are determined by choosing the pose that has the greatest number of points in front of the camera as the appropriate orientation. We attempt to reduce the re-projection inaccuracy of the 3D point position between actual points and re-projected points after obtaining linearized triangulated points. Minimizing algebraic error is the goal of linear triangulation; in non-linear triangulation, the more significant goal is to minimize geometric error, also known as re-projection error. Thus, we fine-tune the 3D point locations in an effort to reduce the re-projection inaccuracy. The linear triangulation provides us with an initial

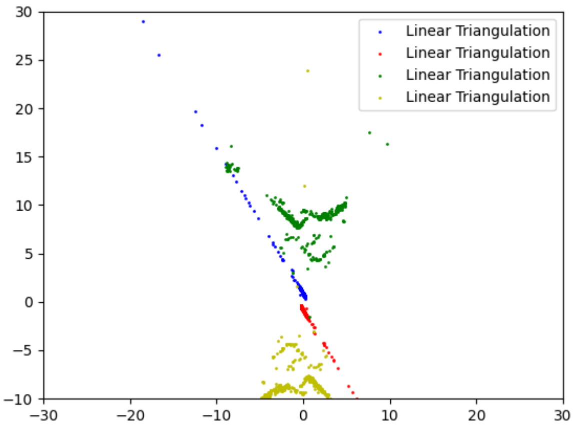

Fig. 8. Initial Triangulation

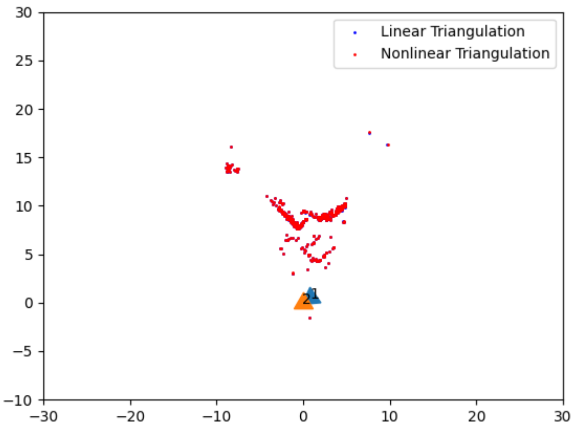

Fig. 9. Linear & Non-Linear Triangulation Comparision

estimate. We use function scipy.optimize.leastsquares and use trust region field as the optimization method.

The mean squared error after non-linear triangulation reaches 2.649

$$\min_{x} \sum_{j=1,2} \left(u^j - \frac{P_1^{jT}\tilde{X}}{P_3^{jT}\tilde{X}} \right)^2 + \left(v^j - \frac{P_1^{jT}\tilde{X}}{P_3^{jT}\tilde{X}} \right)^2$$

F. Perspective-n-Points

We may estimate the 6 DOF camera posture using linear least squares given a set of n real-world 3D points, their 2D image projections, and intrinsic parameters. The Perspective-n-Point issue (PnP) is the name given to this. Using nonlinear optimization, we can register a new image once we obtain 2D-3D correspondences (X-x). The goal of the perspective-n-points, or PnP, issue is to identify the poses of new cameras based on supplementary scene photos that have enough world points visible to the initial two cameras. Since it is expected that there are outlier matching points between the world and the new images, RANSAC is used once more to choose the pose with the least amount of error across all points by examining the reprojection error between the known world points and the new image points. Additionally, a nonlinear estimator uses this as a starting condition to further hone the newly estimated camera posture. It should be emphasized that the camera orientations are represented as quaternions for this process in order to improve the estimator's convergence because quaternions have mathematical features that prevent discontinuity, unlike the Euler angle representation.

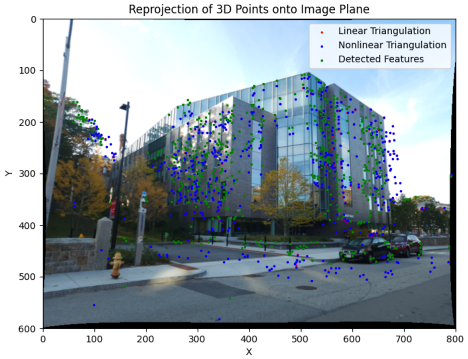

Fig. 10. Reprojected Points

The mean squared error after non-linear PnP reaches 10.52

$$\min_{C,q} \sum_{j=1,J} \left(u^j - \frac{P_1^{jT}\tilde{X}_j}{P_3^{jT}\tilde{X}_j} \right)^2 + \left(v^j - \frac{P_1^{jT}\tilde{X}_j}{P_3^{jT}\tilde{X}_j} \right)^2$$

G. Building Visibility Matrix

Bundle adjustment is done using the visibility matrix, which is a matrix of Booleans indicating whether or not a point is visible to a camera in its picture plane. Although this matrix is often diagonal in form and may be optimally iterated over for the necessary computations, it can be enormous for numerous cameras and points. The visibility matrix Vij connects every camera i to the three-dimensional point j.

H. Bundle Adjustment

Bundle adjustment is a technique used in photogrammetry and computer vision to simultaneously improve the camera poses and 3D point placements calculated from the pipeline's earlier steps. Finding the settings that minimize the gap between the observed picture points and the corresponding points projected from the estimated 3D locations and camera positions is fundamentally an optimization issue.

The objective of bundle adjustment is to identify the parameters that minimize the discrepancy between the projected points from the predicted 3D locations and camera poses and the observed image points. The least squares approach of the trust region reflective algorithm, which is more resilient to

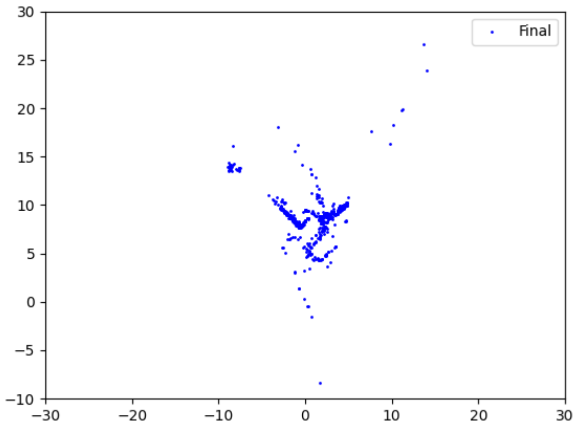

Fig. 11. Point Cloud after Bundle Adjustment

sparse difficulties, can be used for this.
Higher precision and more ideal values are obtained by refining the 3D point positions and camera postures after the bundle adjustment is finished. To assess the technique's efficacy, the refinement can be compared before and after bundle correction. And thus the pipeline is completed with the refinement.

I. *Final Analysis*

In the pursuit of reconstructing a three-dimensional scene from a set of images, the classical structure from motion (SfM) algorithm was employed. This phase involved several key steps, each integral to the accurate reconstruction of the scene. However, upon closer examination, several observations and insights emerged.

1) Feature Matching and Refinement:: One of the foundational steps in the SfM pipeline is feature matching, which lays the groundwork for subsequent calculations. As highlighted, the quality of feature matching directly impacts the accuracy of the reconstruction. It's crucial to acknowledge that the presence of bad data can significantly impede the matching process, leading to erroneous results. Therefore, robust techniques for feature refinement are essential to mitigate such issues and ensure reliable matches. The accuracy of the fundamental matrix (F matrix) calculation hinges on the quality of feature matching, making it imperative to address any discrepancies in the data.

2) Optimization Challenges:: The SfM algorithm heavily relies on optimization, particularly through the process of least squares. However, it's noted that this optimization process can be computationally intensive, resulting in slower performance. As discussed, there exist alternative methods for implementing non-linear optimization, offering potential improvements in both accuracy and speed. Exploring these advanced optimization techniques could lead to more efficient SfM pipelines, enhancing the overall reconstruction process.

3) Conclusion:: In conclusion, while the classical SfM algorithm serves as a fundamental framework for reconstructing three-dimensional scenes, its effectiveness is contingent upon several factors. Addressing challenges related to feature matching and optimization is crucial for achieving accurate and efficient reconstructions. Moving forward, incorporating advanced techniques for feature refinement and optimization could significantly enhance the performance of the SfM pipeline, paving the way for more robust reconstruction solutions in computer vision applications.