Project 2 - Buildings built in minutes - SfM RBE/CS549 Computer Vision

Shambhuraj Mane
MS Robotics Engineering
samane@wpi.edu

Swati Shirke
MS Robotics Engineering
svshirke@wpi.edu

***Abstract*—This paper presents a reconstruction of the 3D scene and camera pose estimation of a monocular camera using a computer vision technique called Structure from Motion (SfM). Here, we are creating the entire rigid structure from a set of images with different viewpoints (or equivalently a camera in motion). By extracting and using common features between multiple images, we are obtaining a 3D point cloud of the scene. The methodology includes feature matching, estimation of a fundamental and essential matrix, triangulation, PnPRANSAC and bundle adjustment technique to obtain a 3D point cloud of a scene.**

Index Terms* — *Epipolar geometry, fundamental and essential matrix, triangulation, perspective n-point transform, bundle adjustment

I. Introduction

The entire pipeline of 3D point cloud reconstruction is as follows:

- Feature Matching and Outlier rejection using RANSAC
- Estimating Fundamental Matrix
- Estimating Essential Matrix from Fundamental Matrix
- Estimate Camera Pose from Essential Matrix
- Check for Cheirality Condition using Triangulation
- Perspective-n-Point
- Bundle Adjustment

A. Dataset

The data set we are using consists of 5 images of Unity Hall of Worcester Polytechnic Institute. These images were captured using Samsung S22 Ultra's primary camera at f/1.8 aperture, ISO 50 and 1/500 sec shutter speed. The camera is calibrated after resizing using a Radial-Tangential model with 2 radial parameters and 1 tangential parameter using the MATLAB R2022a's Camera Calibrator Application. It also includes text files containing common features between multiple files.

B. Feature Matching, Fundamental Matrix and RANSAC

Feature matching is done using SIFT. The fundamental matrix, denoted by F, is a 3×3 (rank 2) matrix that relates the corresponding set of points in two images from different views (or stereo images). The approach utilized involves employing the Random Sample Consensus algorithm, commonly known as RANSAC. This method selects 8 random pairs of points, which are then used to compute the fundamental matrix. Feature matching between multiple images after RANSAC implementation is as shown in the figures.

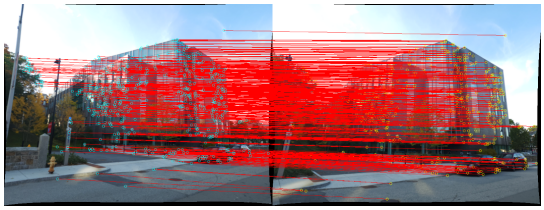

Fig. 1. Feature Matching between image 1 and 2

Fig. 2. Feature Matching between image 1 and 3

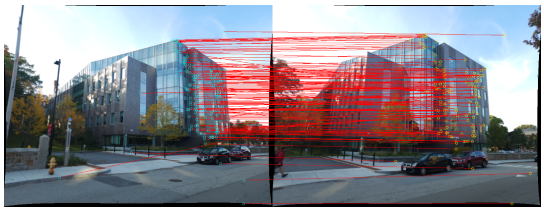

Fig. 3. Feature Matching between image 1 and 4

Fig. 4. Feature Matching between image 1 and 5

An epipolar line is the intersection of an epipolar plane with the image plane. The following figures show epipolar lines between different image pairs.

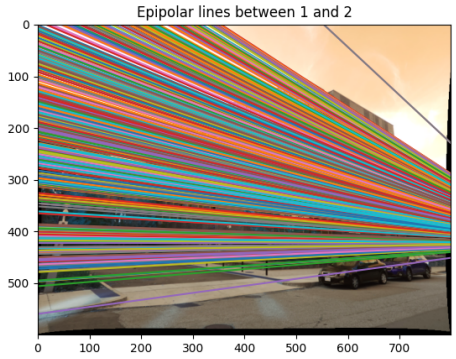

Fig. 5. Epipolar lines between image 1 and 2

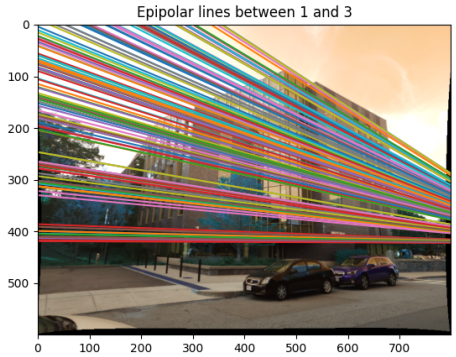

Fig. 6. Epipolar lines between image 1 and 3

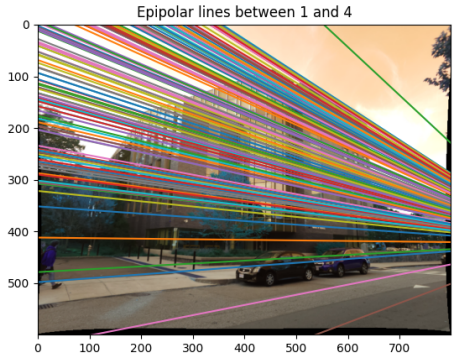

Fig. 7. Epipolar lines between image 1 and 4

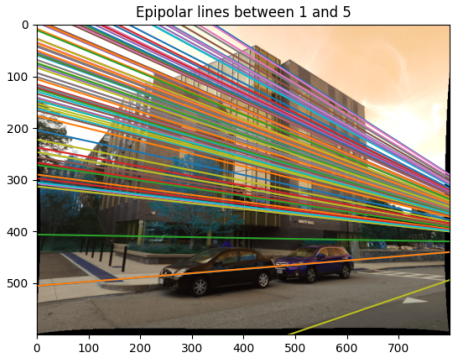

Fig. 8. Epipolar lines between image 1 and 5

C. Essential Matrix

The Essential matrix is calculated from the Fundamental matrix and intrinsic parameter matrix of a camera. The essential matrix relates the corresponding points assuming that the cameras obey the pinhole model (unlike F). More specifically, E = KTFK where K is the camera calibration matrix or camera intrinsic matrix. The camera pose consists of 6 degrees-of-freedom (DOF) Rotation (Roll, Pitch, Yaw) and Translation (X, Y, Z) of the camera with respect to the world.

D. Triangulation and Cheirality condition

In this case, we triangulate 3D points from the perspectives of two camera poses. To determine the accurate camera poses, we must verify the Cheirality condition. This condition guarantees that the selected camera pose corresponds to the one positioned in front of the camera's centre. To assess the Cheirality condition, we triangulate the 3D points given the two camera poses using linear least squares. This enables us to examine the sign of depth Z concerning the camera centre within the camera's coordinate system. After performing linear triangulation, we refined 3D points using non-linear triangulation. The results of both linear and non-linear triangulation are shown in Fig 9 and 10. In Fig. 11, it can be seen that, before optimization, triangulation points are sparse and after optimization, denser points can be seen.

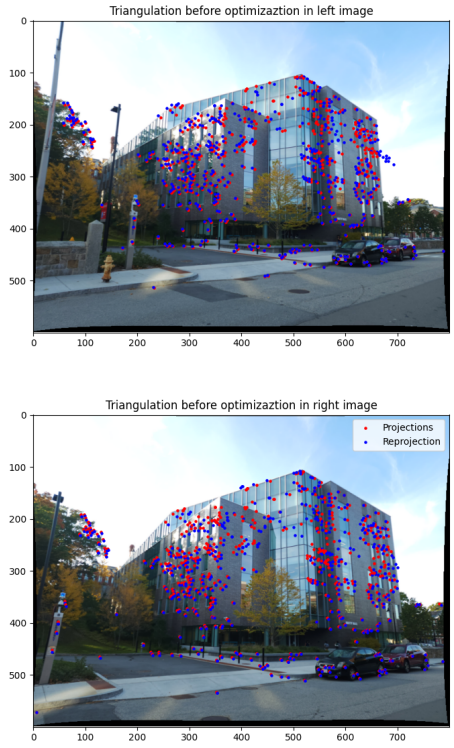

Fig. 9. Triangulation before optimization between image 1 and 2

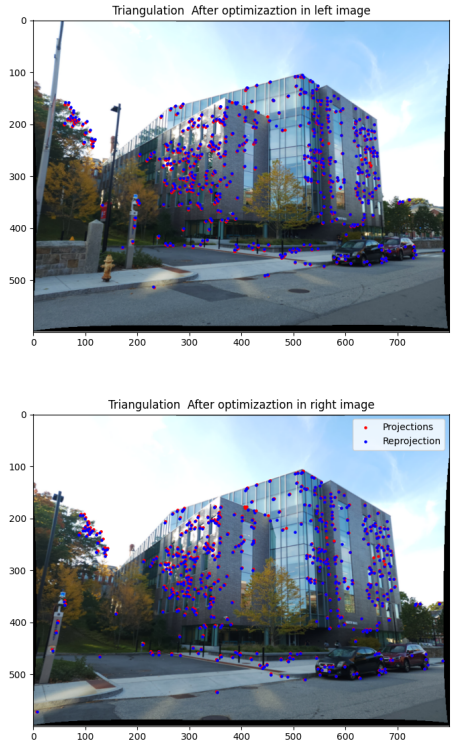

Fig. 10. Triangulation after optimization between image 1 and 2

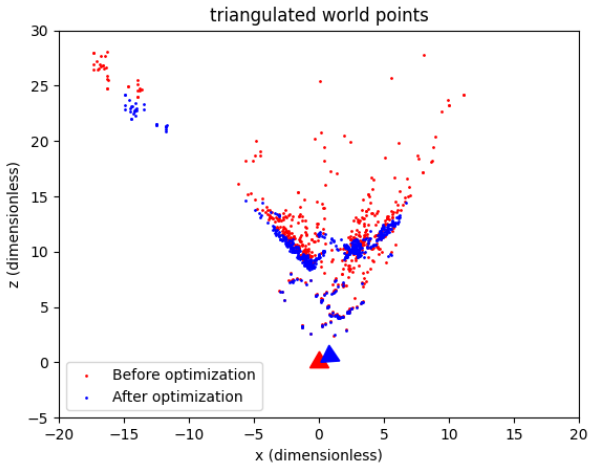

Fig. 11. Triangulated points before and after optimization

E. PnP and PnP-RANSAC

Given a collection of n 3D points in the world, along with their 2D projections in the image and the intrinsic parameters, the 6 degrees of freedom (DOF) camera pose can be inferred through linear least squares estimation. This is in general is known as Perspective-n-Point (PnP) transform. A linear PnP is implemented to find camera rotation matrix and centre: R and C. Pnp-RANSAC is implemented to perform outliers rejection. Further, these poses are refined using non-linear optimization.

Fig. 12. PnP RANSAC results before optimization between image 1 and 3

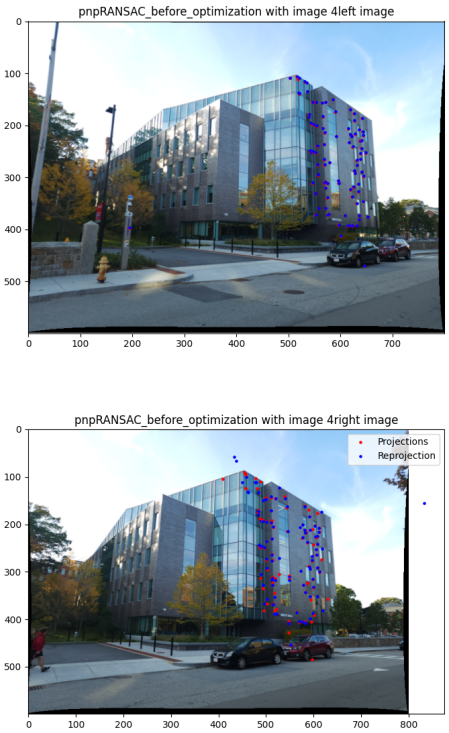

Fig. 13. PnP RANSAC results before optimization between image 1 and 4

Fig. 14. PnP RANSAC results before optimization between image 1 and 5

Fig. 15. PnP RANSAC results after optimization between image 1 and 3

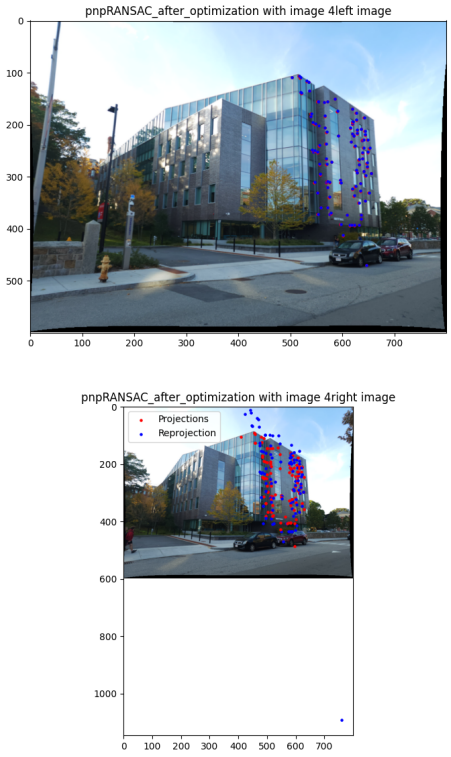

Fig. 16. PnP RANSAC results after optimization between image 1 and 4

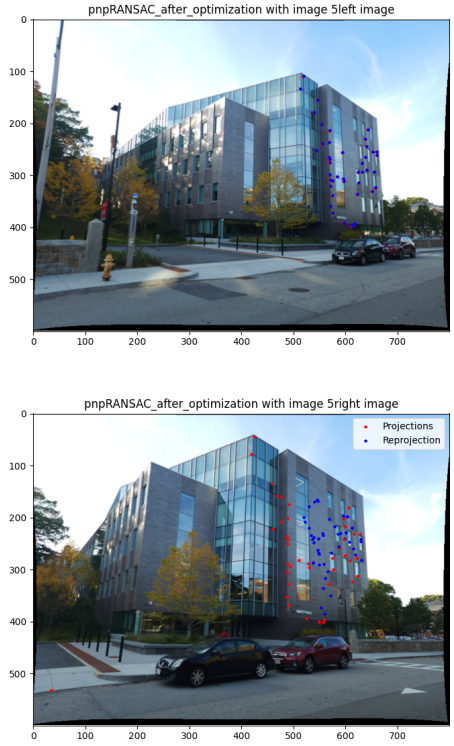

Fig. 17. PnP RANSAC results after optimization between image 1 and 5

F. Bundle Adjustment

After finding 5 camera poses and 3D points, these can be further refined by using bundle adjustment to minimize re-projection error. While implementing this, we create a boolean matrix which states whether a point is visible from a particular camera or not. Results achieved after this step are the final results of the SfM pipeline.

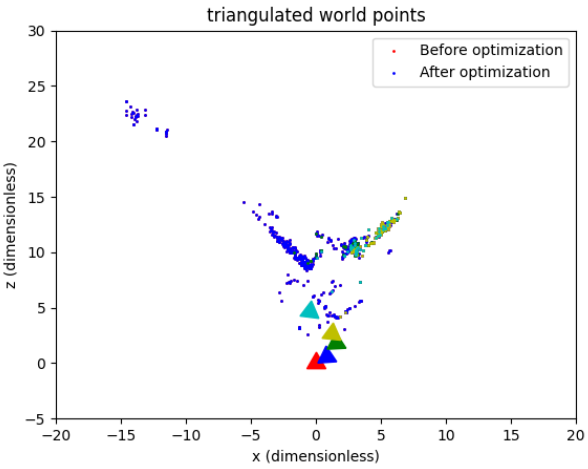

Fig. 18. All Camera Poses Before bundle adjustment

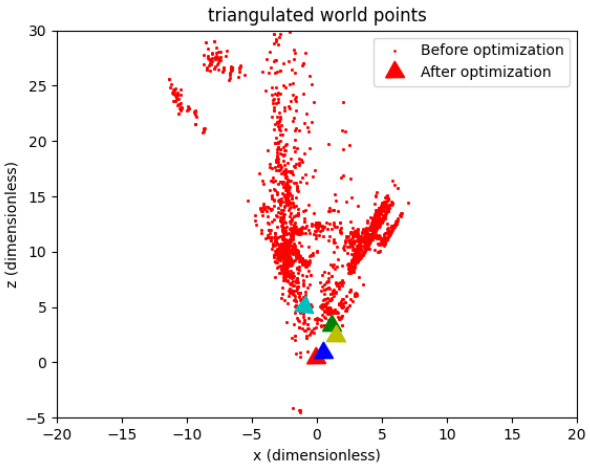

Fig. 19. All Camera Poses Before bundle adjustment

Fig. 20. Camera pose from VisualSFM, a GUI application for 3D reconstruction

G. Putting pipeline together

Structure from Motion (SfM), a computer vision technique involves the extraction of key features and their matching across images, followed by camera pose estimation. Bundle adjustment optimizes the camera poses and 3D points to minimize reprojection errors. The algorithmic process of SfM enables the recovery of geometric information from multiple

images, facilitating a deeper understanding of the underlying scene. We have used the following algorithm to get the results.

```
Data: Image Matches, K
Result: X, C, R
for all possible pair of images do
    // Reject outlier correspondences
    [x1, x2] = GetInliersRANSAC(x1, x2);
end
// For first two images
F = EstimateFundamentalMatrix(x1, x2);
E = EssentialMatrixFromFundamentalMatrix(F, K);
[Cset, Rset] = ExtractCameraPose(E);
// Perform linear triangulation
for i = 1:4 do
    Xseti = LinearTriangulation(K, zeros(3,1), eye(3), Cseti, Rseti, x1, x2);
end
// Check cheirality condition
[C R] = DisambiguateCameraPose(Cset, Rset, Xset);
// Perform Non-linear triangulation
X = NonlinearTriangulation(K, zeros(3,1), eye(3), C, R, x1, x2, X0));
Cset = C, Rset = R;
// Register camera and add 3D points for the rest of images
for i=3:I do
    // Register the i^th image using PnP.
    [Cnew Rnew] = PnPRANSAC(X, x, K);
    [Cnew Rnew] = NonlinearPnP(X, x, K, Cnew, Rnew);
    Cset = Cset∪Cnew, Rset = Rset∪Rnew;
    // Add new 3D points.
    Xnew = LinearTriangulation(K, C0, R0, Cnew, Rnew, x1, x2);
    Xnew = NonlinearTriangulation(K, C0, R0, Cnew, Rnew, x1, x2, X0);
    X = X∪Xnew;
    // Build Visibility Matrix.
    V = BuildVisibilityMatrix(traj);
    // Perform Bundle Adjustment.
    [Cset Rset X] = BundleAdjustment(Cset, Rset, X, K, traj, V);
end
```

Algorithm 3: Alogrithmic overview of SfM.

Fig. 21. End-to-end pipeline of SfM

H. Re-projection Error

Steps	Name	Between	Error Values
Step1	LinearTriangulation	1 and 2	4.5766
Step2	NonLinearTriangulation	1 and 2	1.3904
Step3	linearPnP	1 and 3	0.7251
Step4	NonlinearPnP	1 and 3	0.7251
Step5	linearPnP	1 and 4	0.7350
Step6	NonlinearPnP	1 and 4	0.7350
Step7	linearPnP	1 and 5	0.8025
Step8	NonlinearPnP	1 and 5	0.8025
Step9	BundleAdjustment	1 and 3	450.277
Step10	BundleAdjustment	1 and 4	90.7544
Step11	BundleAdjustment	1 and 5	100.9657

TABLE I
ERROR VALUES AT EACH STEP OF THE RECONSTRUCTION PROCESS

II. ANALYSIS

A. Data Augmentation

Considering the end-to-end pipeline of SfM as shown in Fig 21, feature points common between image 1 and image n are used to find camera poses using PnP transform. However, there could be extra feature points common between image m and n where $m >= 2, n >= 2$. These feature points are not used in camera pose and 3D point cloud estimation. If we use these extra feature points in the calculations, better results can be achieved.

B. Impact of outliers on Non-linear PnP

Results of the linear PnP-RANSAC are satisfactory as shown in Fig 16, 17. However, after performing non-linear optimization, the results obtained are not as satisfactory as before the non-linear optimization. One of the possible reasons could be outliers. Even a few outliers with high values can impact the results of non-linear PnP, resulting in erroneous camera poses, and consequently introducing re-projection error.

C. Bundle Adjustment

The application of bundle adjustment in Structure from Motion (SfM) has yielded unsatisfactory results, contrary to the positive outcomes observed prior to its implementation. The preliminary evaluation suggests that the issue may not solely stem from an inadequate initial guess, as the results obtained before the bundle adjustment process were promising. However, the inconsistency prompts the need for a more in-depth analysis to identify the underlying reasons behind the suboptimal performance of bundle adjustment in this particular context. Further investigation is essential to uncover potential limitations or nuances in the algorithm that may be impeding its effectiveness within the SfM framework. By understanding the specific challenges faced, we can develop targeted solutions to enhance the overall performance of bundle adjustment for improved 3D reconstruction in SfM applications.

III. CONCLUSION

In conclusion, the successful implementation of the Structure from Motion (SfM) pipeline involved several crucial steps, including feature matching and RANSAC-based outlier rejection. The estimation of the Fundamental Matrix and subsequent derivation of the Essential Matrix were key components, leading to the accurate estimation of camera pose. The integration of Cheirality Condition verification through triangulation ensured the robustness of the reconstruction. The Perspective-n-Point algorithm further refined the localization accuracy. Finally, the application of Bundle Adjustment aimed to optimize the entire reconstruction process, although its efficacy requires further investigation for optimal performance in our specific context.

REFERENCES

[1] Agarwal, S., Snavely, N., Simon, I., Seitz, S. M., & Szeliski, R. (2011). Building Rome in a Day. Proceedings of the International Conference on Computer Vision (ICCV), 1792-1799. University of Washington, Cornell University, Microsoft Research.

[2] C.Wu,"VisualSFM A Visual Structure from MotionSystem,"Computer Vision at Stony Brook University,Available:http://ccwu.me/vsfm/

[3] Wikipedia contributors. (2022, February 18). Eight-point algorithm. In Wikipedia. Retrieved February 21, 2024, from https://en.wikipedia.org/wiki/Eight-point_algorithm

[4] SciPy Cookbook. (n.d.). Bundle Adjustment. Retrieved from https://scipycookbook.readthedocs.io/items/bundle_adjustment.html

[5] Prasannanatu. (2022, February 21). SfM and NeRF Implementation. GitHub. https://github.com/Prasannanatu/sfm_and_nerf.git