Structure-from-Motion (SfM) and Neural Radiance Field (NeRF) - RBE549 Project 3

USING 5 LATE DAYS

Ajith Kumar Jayamoorthy
Robotics Engineering Department
Worcester Polytechnic Institute
Worcester, MA, U.S.A.
ajayamoorthy@wpi.edu

Shiva Kumar Tekumatla
Robotics Engineering Department
Worcester Polytechnic Institute
Worcester, MA, U.S.A.
stekumatla@wpi.edu

***Abstract*—This document consist of project implementation of 3D reconstruction of a scene from multiple 2D image. Two different approaches are considered: First one is the traditional Structure from Motion (SfM) method. Second is an deep learning approach using the Neural Radiance Fields (NeRF). The results and observations for each method as been recorded in this document.**

I. Introduction

The main objective of the project is to generate a 3D model of a particular scene from multiple 2D images captured from different perspectives. Two methods have been studied and implemented in this project. In the first method, called the Structure from Motion (SfM), 6 images have been given to start with, a text file describing the 2D image point correspondences between all possible image pairs and the calibration matrix of the camera used for capturing the images. In case of the second method, a new deep learning approach, called the Neural Radiance Field (NeRF) has been used to train a model that helps in the evaluation of coordinates for reconstruction of the scene in 3D. There sets of lego images have been given, namely, train, test and val for training, testing and validation respectively. Each Camera parameters for each image have been further given by .json files to used for the training purpose.

II. Traditional Approach - Structure from Motion (SfM)

The traditional method consists of following steps:

1) Feature Matching and Outlier rejection using RANSAC
2) Estimating Fundamental Matrix (F)
3) Estimating Essential Matrix (E)
4) Estimate Camera Pose from Essential Matrix
5) Check for Cheirality Condition using Triangulation
6) Perspective-n-Point
7) Bundle Adjustment

A. Feature Matching and Outlier rejection using RANSAC

Initially, there are 6 images provided which are from different camera perspectives of the same building with fixed camera parameters. Each pair of images have lot of pixels common. The figure 1 has one of the sample input images.

Fig. 1. Camera Perspective 1 (Unity Building, WPI) [1]

Feature matching are given in text files and they are predetermined by feature descriptors. The given data tends to be noisy and contains outliers and therefore we will remove the outliers using RANSAC. Given a pair of images long with the matches, we use RANSAC with the 8-point algorithm to estimate the fundamental matrix between them.

B. Estimating Fundamental Matrix (F)

In this method we randomly sample 8 points and then we estimate the fundamental matrix F. Then we count the number of points that satisfy the epipolar constraint (x'Fx ≈ 0) and finally select the fundamental matrix that results in the largest number of inlier correspondences. The Fundamental matrix estimated using the above method is as follows:

$$F = \begin{bmatrix} -6.4121e-07 & -2.4365e-05 & 1.3846e- \\ 2.8041e-05 & 1.6928e-06 & -3.5683e-02 \\ -1.5631e-02 & 3.4325e-02 & 1.0 \end{bmatrix}$$

C. Estimating Essential Matrix (E)

We estimate the essential matrix by from the fundamental matrix by using the equation E= K^TFK, where K is the camera calibration matrix or camera intrinsic matrix. As in the case of F matrix computation, the singular values of E are not necessarily (1,1,0) due to the noise in K. This has been corrected by reconstructing it with (1,1,0) singular values by using SVD, as suggested in the reference website [1]. The Essential Matrix is evaluated and is given below:

$$E = \begin{bmatrix} -0.01404208 & -0.45120026 & 0.21074795 \\ 0.51477257 & 0.0415464 & -0.82473738 \\ -0.24096739 & 0.8640367 & 0.01652605 \end{bmatrix}$$

D. Estimate Camera Pose from Essential Matrix

The first camera is considered to be the origin of the global/world coordinate system. The second camera data is taken into consideration and based on the essential matrix, four different configurations of the second camera from the first camera are calculated in the form of translation vector (C) and Rotation matrix (R). They are evaluated using singular value decomposition as follows:

$$\mathbf{E} = UDV^T$$

$$W = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The four configuration are calculated as follows:

$$C1 = U(:,3) \;\; R1 = UWV^T$$

$$C2 = U(:,3) \;\; R2 = UWV^T$$

$$C3 = U(:,3) \;\; R3 = UW^TV^T$$

$$C4 = U(:,3) \;\; R4 = UW^TV^T$$

E. Cheirality Condition using Triangulation

*1) **Linear Triangulation**:* Using the 4 estimated camera poses, we found out the 3D world points corresponding to the matches between the two images. This was done by performing SVD on a system of linear equations to minimize the L1 distance between a projected 3D point and the actual 2D image point.

*2) **Cheirality condition checking**:* After that using the generated 3D work points and the camera poses, the cheirality check was performed to get the best camera pose by using the condition $r_3(\mathbf{X} - \mathbf{C}) > 0$, where r_3 is the third row of the Rotation matrix (Z-axis of the camera), X is the 3D global co-ordinates and C is the translation vector of the camera centre in global co-ordinates. The configuration which satisfies the Linear triangulation condition is visualized as shown in figure 3

*3) **Non-Linear Triangulation**:* Given two camera poses and linearly triangulated points X, we can optimize the re-projection error of the 3D points in the image plane. We accomplish this task by using the optimize function from the scipy library given as follows **scipy.optimize.least_squares**

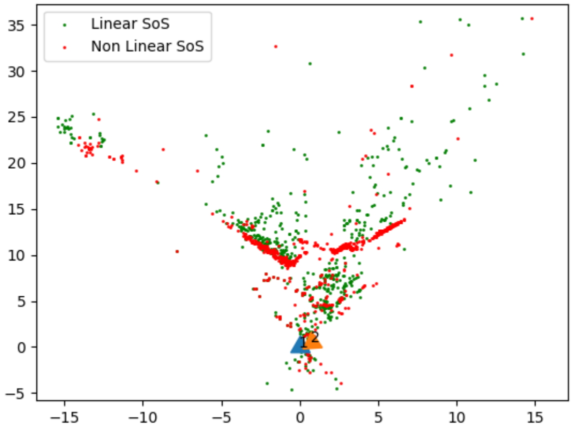

Fig. 2. Linear and Non-Linear Triangulation for first two camera poses [1]

F. Perspective-n-Points (PnP)

Since we have a set of n 3D points in the world, their 2D projections in the image and the intrinsic parameter; the 6 DOF camera pose can be estimated using linear least squares. This fundamental problem, in general is known as Perspective-*n*-Point (P*n*P). For a solution to exist, the number sets of 3D points used should be greater than or equal to three i.e. n≥3

*1) **Linear Camera Pose Estimation (Linear PnP**:* Given the correspondence between the image points(x) and the world points(X) along with the intrinsic parameters of the camera (K), we first calculated the inverse of the intrinsic parameter matrix to normalise the image points. Then we solve the system of equations by using the Ax = 0 to to get linear least squares solution with SVD.

*2) **PnP RANSAC**:* Linear PnP is prone to error as there are outliers in the given set of point correspondences. To overcome this error, we again use RANSAC to make our camera pose more robust to outliers.

*3) **Non-Linear PnP**:* In the Non-Linear PnP we were trying to minimize the projection error by optimizing the rotation matrix and translation vector using the correspondences given before. The optimization is done same as in case of Non-Linear triangulation using the scipy.optimize.least_squares function. The re-projections of the triangulated 3D world points after are as shown in figure 4

G. Bundle Adjustment

1) Visibility Matrix: We create a matrix based on the camera poses and the visibility of the given world point from that particular camera pose. The rows of the matrix represent the camera pose index and the columns represent the index of the real world co-ordinates. For every world point index, 0 if marked if there is no visibility from a particular camera index and 1 is assigned if the point is visible from that particular camera index.

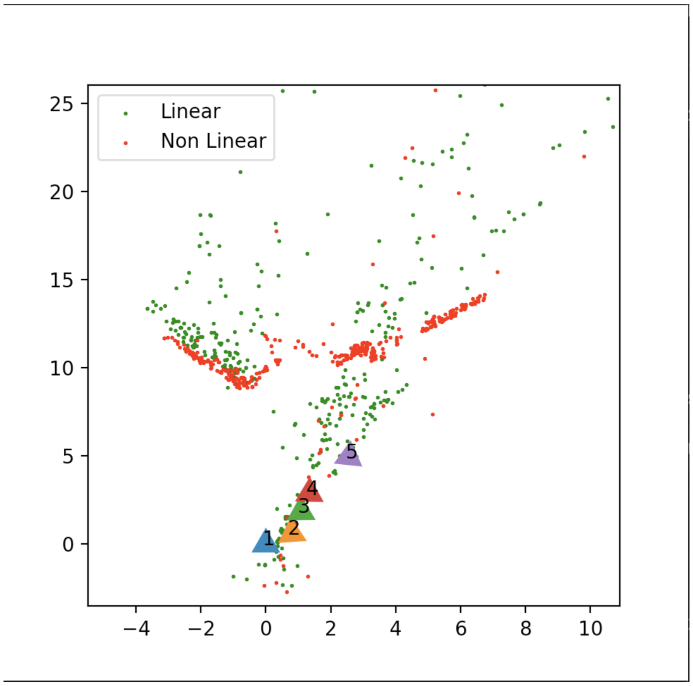

Fig. 3. PnP results before Linear and Non-Linear Triangulation [1]

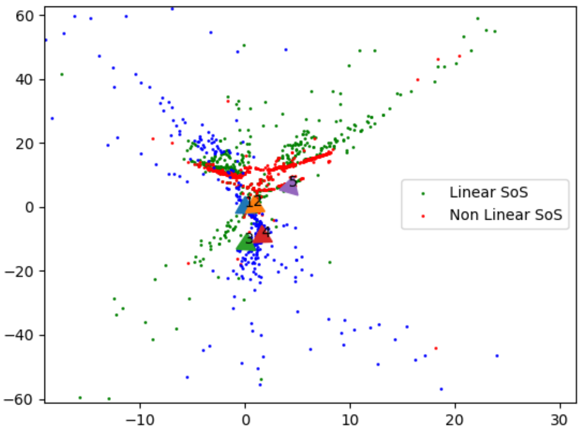

Fig. 4. Camera Pose estimation after PnP optimization and Triangulation [1]

2) Bundle Adjustment Implementation: After the Non-Linear PnP, the new Rotation matrix and translation vector obtained are further optimized. The input the bundle adjustment function are the World coordinates (X), the new camera co-ordinate x, the camera parameter, rotation matrix (R), translation vector (C) and the visibility matrix (V). Further optimization of the rotation matrix (R) and translation vector (C) are performed by using the least_squares method from the scipy.optimize library. The error function considered is same as the re-projection error used in the previous optimization problems. Based on this error the optimization process is done and new values of R and C are obtained. Along with this parameters, new global co-ordinates set (X) is also passed as the output.

III. Deep Learning Approach - Neural Radiance Field (NeRF)

A. *Data*

The data for NeRF is given from the original authors link. [2] The input data consist of three files, namely, train, test and validation with images consisting of a lego model of a object. A sample image is shown in figure 6

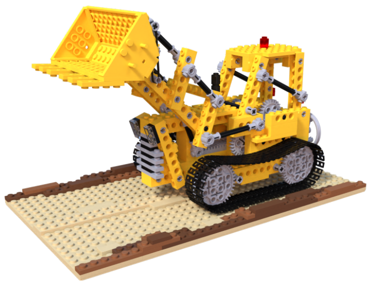

Fig. 5. Example train image used by the NeRF network [1]

B. *NeRF*

The Neural Radiance Field or simply NeRF is the state-of-the-art method that generates complex novel view of an scene by optimizing the underlying volumetric scene function using a sparse set of input images. The input is provided as a 5D function with each images, where each pixel point co-ordinate in 3D space is provided as (x,y,z) co-ordinates followed by the direction of each scene with respect to the global co-ordinates as (θ; ϕ). [4]

C. *Architecture*

NeRF takes a set of input images of a scene and renders the complete scene by interpolating between the scenes. The output is a volume whose color and density are dependent on the direction of view and emitted light radiance at that point. For each ray, we get an output volume and all these volumes make up the complex scene. [5]

The following are the steps to render this NeRF [5]:

1) march camera rays through the scene to sample 3D points
2) use points from step 1 and their corresponding 2D viewing directions(θ, ϕ) as input to the MLP to produce an output set of colors (c = (r, g, b)) and densities σ,
3) use volume rendering techniques to accumulate those colors and densities into a 2D image.

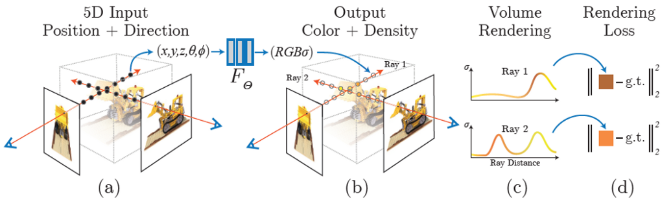

Fig. 6. (a) Sampling 5D coordinates (location and viewing direction) along camera rays; (b) feeding those locations into an MLP to produce a color and volume density; c)using volume rendering techniques to composite these values into an image; (d) optimize the scene representation by minimizing the residual between synthesized and ground truth observed images [3]

References

[1] https://rbe549.github.io/fall2022/proj/p3/

[2] https://drive.google.com/drive/folders/1lrDkQanWtTznf48FCaW5lX9ToRdNDF1a

[3] Mildenhall,B., Srinivasan,P., Tancik, M., Barron, J.T., Ramamoorthi,R., & Ng,R. (2020, August 3). Nerf: Representing scenes as neural radiance fields for view synthesis. arXiv.org. Retrieved November 10, 2022, from https://arxiv.org/abs/2003.08934v2

[4] https://medium.com/swlh/nerf-neural-radiance-fields-79531da37734

[5] https://www.analyticsvidhya.com/blog/2021/04/introduction-to-neural-radiance-field-or-nerf/