Structures From Motion

Venkata Sai Charan Paladugu
120392948
University of Maryland, College Park
vpaladug@umd.edu

Hoang Pham
120230301
University of Maryland, College Park
hmp61595@umd.edu

Fazil Mammadli
120227561
University of Maryland, College Park
mammadli@umd.edu

Rachana Devarapalli
120334733
University of Maryland, College Park
rdeverap@umd.edu

Abstract

In order to reconstruct three dimensions from a sequence of 13 photographs, this project explores the use of Structure from Motion (SfM) methods. Through the use of two-dimensional image sequences and computer vision techniques, SfM is able to estimate three-dimensional structures while capturing the motion and geometry of the surroundings. Our goal is to find and recreate objects in the provided dataset by utilising a variety of perspectives. The research combines triangulation, matching, bundle adjustment and feature recognition to create a logical three-dimensional model that aids in precise spatial comprehension. With possible applications in robotics, virtual reality, and archaeological documentation, the results show how successful SfM is at creating detailed 3D reconstructions. The value of image-based approaches in capturing intricate settings and objects is highlighted by this work, which advances automated 3D reconstruction techniques.

1. Introduction

The progress made in computer vision algorithms has greatly improved our capacity to extract and decipher visual information from images. Because it can reconstruct three-dimensional structures from two-dimensional image sequences, Structure from Motion (SfM) is one such technology that has gained popularity. SfM uses overlapping photos taken from various angles to estimate the 3D coordinates of points in a scene. It is based on the concepts of photogrammetry and geometric computer vision.

In order to identify objects and recreate them in three dimensions, this study applies SfM on a dataset of fourteen photos. Using the spatial information included in the image set to create a precise and comprehensive 3D representation of the scene is the main goal. SfM algorithms are able to triangulate the relative positions of essential features and rebuild the 3D geometry of a scene by recognising and matching these features across many photos.

There are a few crucial steps in the procedure. The first step in the process is feature matching, which finds and matches important traits throughout the collection of images to create similarities. Next comes the application of RANSAC for match outlier rejection and the estimation of the basic matrix through the use of epipolar geometry. The necessary matrix is estimated from the basic matrix in order to comprehend the motion of the camera.

Next, based on the essential matrix, the camera pose is estimated, which gives the location and orientation of the camera in space. The cheirality criterion is verified by non-linear triangulation in order to guarantee legitimate 3D points. After then, the Perspective-n-Points (PnP) technique is used, which includes non-linear PnP for fine-tuning camera poses, PnP RANSAC, and linear camera pose estimation. The 3D structure and camera parameters are then thoroughly refined through bundle adjustment, which is carried out using the visibility matrix. This study is important because it has the potential to be applied in many other fields, such as robotics, augmented and virtual reality, cultural heritage preservation, and environmental monitoring. SfM provides a practical and economical substitute for conventional techniques by automating the 3D reconstruction process, opening up the technology to a broad range of applications.

This study provides insights into the difficulties and solutions faced during the reconstruction process by outlining the SfM project's methodology, implementation, and outcomes. The findings' ramifications and prospective advancements to raise the precision and effectiveness of 3D reconstruction methods are also covered.

2. Methodology

We followed the pipeline which was on the CMSC 733 website. The process is broken down into several critical steps, each essential for the accurate reconstruction of 3D structures from a set of 14 images.

2.1. Data Loading and Feature Matching

The first step in the SfM pipeline is feature matching to identify distinctive and repeatable points in each image that can be reliably matched across multiple views. Keypoints are detected in each image using the Scale-Invariant Feature Transform (SIFT) algorithm. SIFT is chosen for its robustness to changes in scale, rotation, and illumination. After detecting keypoints, feature descriptors are computed for each keypoint. To establish correspondences between keypoints detected in different images, which are essential for 3D reconstruction these descriptors are then matched across images using a nearest-neighbor search. To improve the reliability of matches, a ratio test is applied to filter out ambiguous matches. Ratio test, proposed by David Lowe as part of the SIFT algorithm, the ratio test improves the reliability of matches by filtering out ambiguous matches. For a given keypoint in one image, the nearest neighbor and the seWith the matched features, the next step involves establishing the epipolar geometry between image pairs. Epipolar geometry describes the geometric relationship between two views of the same scene, constraining the possible locations of matching points. cond nearest neighbor in another image are identified, and the ratio of the distance of the nearest neighbor to the distance of the second nearest neighbor is computed. If this ratio is below a certain threshold (typically 0.7-0.8), the match is considered reliable; otherwise, it is discarded. This helps in eliminating matches that are not distinctive enough, which could lead to incorrect correspondences and errors in 3D reconstruction. The process of keypoint detection, descriptor computation, and feature matching is fundamental to the SfM pipeline. Detecting reliable keypoints that are invariant to various transformations ensures that features can be robustly matched across different images. The use of techniques like the ratio test further enhances the reliability of these matches, leading to more accurate 3D reconstruction. This foundational step sets the stage for subsequent processes like triangulation, pose estimation, and bundle adjustment, ultimately enabling the creation of detailed and accurate 3D models from 2D images.

2.2. Estimating Fundamental Matrix

With the matched features, the next step involves establishing the epipolar geometry between image pairs. Epipolar geometry describes the geometric relationship between two views of the same scene, constraining the possible locations of matching points.

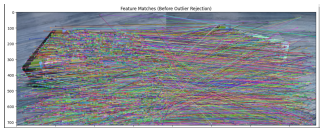

Figure 1. Epipolar lines for original images

The fundamental matrix (F) encapsulates this epipolar geometry. It is estimated using the eight-point algorithm, which requires at least eight pairs of matched points.

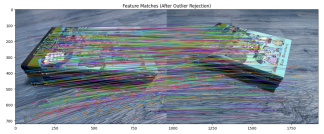

Figure 2. Epipolar lines after outlier rejection

Since the point correspondences are computed using SIFT or other feature descriptors, the data is typically noisy and often contains several outliers. To remove these outliers, we use the RANSAC algorithm to obtain a more accurate estimate of the fundamental matrix. Out of all possible estimates, the fundamental matrix F with the maximum number of inliers is chosen.

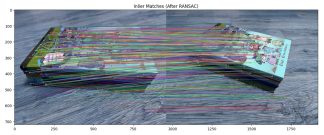

Figure 3. Inliers after using RANSAC

2.3. Estimating Essential Matrix

We can now identify the relative camera positions between the two pictures since we have computed the fundamental matrix **F** utilising epipolar constraints. The Essential Matrix **E** is used to do this. Unlike the Fundamental matrix **F** , the essential matrix is another 3x3 matrix that, under the assumption that the cameras follow the pinhole camera model, links the equivalent spots in two pictures.

The fundamental matrix **F** and the camera intrinsic matrix *K* may be used to produce the essential matrix *E*. The connection may be expressed specifically as follows:
$\mathbf{E} = \mathbf{K}^T\mathbf{F}\mathbf{K}$, where **K** is the camera calibration matrix that contains the intrinsic characteristics of the camera. With the help of this relationship, we can include the inherent characteristics of the camera in the computation and provide a more realistic depiction of the relative motion between the cameras.

The calculated essential matrix **E**, however, could not have the ideal singular values of (1, 1, 0) because to noise

and errors in the intrinsic matrix $\mathbf{K}$. Maintaining the pinhole camera model's geometric limitations requires that the singular values of a perfect essential matrix be (1, 1, 0). Using singular value decomposition (SVD) to rebuild $\mathbf{E}$, we may rectify this by enforcing these unique values.

By ensuring that $\mathbf{E}$ has the correct singular values, we can accurately estimate the relative rotation and translation (camera poses) between the two images. This step is crucial for reconstructing the 3D structure of the scene, as it allows us to establish the relative positions and orientations of the cameras, thereby enabling the triangulation of 3D points from the 2D correspondences.

2.4. Triangulation and Cheirality Check

Triangulation is a fundamental technique in Structure from Motion (SfM) used to reconstruct the 3D coordinates of points from their 2D projections in multiple images. This section describes the mathematical formulation and implementation of triangulation, as well as the cheirality check to ensure the correctness of the reconstructed 3D points.

2.4.1 Mathematical Formulation

Given two images with corresponding points $\mathbf{p}_1 = [x_1, y_1, 1]^T$ and $\mathbf{p}_2 = [x_2, y_2, 1]^T$, and their respective camera projection matrices $\mathbf{P}_1$ and $\mathbf{P}_2$, we aim to solve for the 3D point $\mathbf{X} = [X, Y, Z, 1]^T$.

The relationship between the 3D point $\mathbf{X}$ and its 2D projections is given by:

$$\mathbf{p}_1 = \mathbf{P}_1\mathbf{X} \tag{1}$$

$$\mathbf{p}_2 = \mathbf{P}_2\mathbf{X} \tag{2}$$

Expanding these equations, we get:

$$\begin{aligned} x_1 &= \frac{\mathbf{P}_1^{(1)}\mathbf{X}}{\mathbf{P}_1^{(3)}\mathbf{X}}, \quad y_1 = \frac{\mathbf{P}_1^{(2)}\mathbf{X}}{\mathbf{P}_1^{(3)}\mathbf{X}} \\ x_2 &= \frac{\mathbf{P}_2^{(1)}\mathbf{X}}{\mathbf{P}_2^{(3)}\mathbf{X}}, \quad y_2 = \frac{\mathbf{P}_2^{(2)}\mathbf{X}}{\mathbf{P}_2^{(3)}\mathbf{X}} \end{aligned} \tag{3}$$

where $\mathbf{P}_i^{(j)}$ represents the j-th row of the i-th camera projection matrix.

To solve for $\mathbf{X}$, we rearrange the above equations into a linear system represented in matrix form:

$$\mathbf{A}\mathbf{X} = \mathbf{0} \tag{4}$$

where $\mathbf{A}$ is a 4×4 matrix formed as:

$$\mathbf{A} = \begin{bmatrix} x_1\mathbf{P}_1^{(3)} - \mathbf{P}_1^{(1)} \\ y_1\mathbf{P}_1^{(3)} - \mathbf{P}_1^{(2)} \\ x_2\mathbf{P}_2^{(3)} - \mathbf{P}_2^{(1)} \\ y_2\mathbf{P}_2^{(3)} - \mathbf{P}_2^{(2)} \end{bmatrix} \tag{5}$$

This system can be solved using Singular Value Decomposition (SVD) to find the solution $\mathbf{X}$ that minimizes the algebraic error. Perform SVD on $\mathbf{A}$:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{6}$$

The solution $\mathbf{X}$ corresponds to the last column of $\mathbf{V}$, scaled such that the fourth element is 1:

$$\mathbf{X} = \frac{\mathbf{V}_4}{\mathbf{V}_4^{(4)}} \tag{7}$$

2.4.2 Cheirality Check

The cheirality check ensures that the reconstructed 3D point $\mathbf{X}$ lies in front of both cameras. The depth of $\mathbf{X}$ with respect to a camera is given by the third element of the reprojected point $\mathbf{P}_i\mathbf{X}$:

$$\mathbf{X}_{depth} = (\mathbf{P}_i\mathbf{X})^{(3)} \tag{8}$$

For the point to be in front of the camera, its depth should be positive. Thus, the cheirality condition is:

$$(\mathbf{P}_1\mathbf{X})^{(3)} > 0 \quad \text{and} \quad (\mathbf{P}_2\mathbf{X})^{(3)} > 0 \tag{9}$$

If both conditions are satisfied, the point $\mathbf{X}$ is considered to be correctly triangulated.

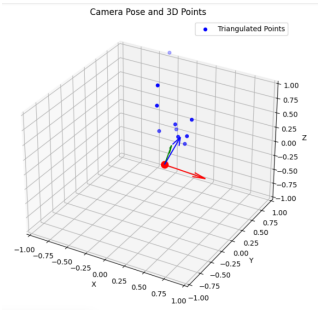

Figure 4. Points after cheirality check and triangulation

2.5. Perspective n Points

Now, since we have a set of n 3D points in the world, their 2D projections in the image, and the intrinsic parameters, the 6 Degrees of Freedom (DOF) camera pose can be estimated using linear least squares. This fundamental problem is generally known as Perspective-n-Point (PnP). For a solution to exist, $n \geq 3$. There are multiple methods to solve the PnP problem, most of which assume that the camera is calibrated. However, methods such as Unified PnP do not adhere to this assumption as they estimate both intrinsic and extrinsic parameters. In this section, we will discuss a simpler version of PnP. Specifically, we will register a new image given 2D-3D correspondences ($X \leftrightarrow x$), followed by nonlinear optimization.

2.5.1 Linear Pose Estimation

Now that we have the rotation and translation from camera pose estimation and also the 3 D points from triangulation. We find the best camera pose for these points using algebraic methods. This pose has huge error to it because of the several factors like noise, outliers.

2.5.2 PnP RANSAC

The camera poses are checked with subsets of the 3D points and checks for which pose aligns with most number of 3D points. The pose with the highest number of alignments is selected.

2.5.3 Non-Linear PnP

The camera poses obtained from the previous step are robust but they are solved for non linear conditions. Now we check the camera poses even in non linear conditions. This minimises the 3D projection error and a more accurate and reliable camera pose estimation.

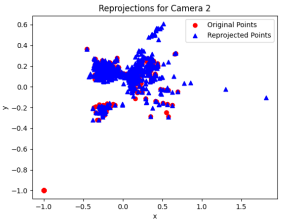

Figure 5. Reprojection error after PnP is performed

3. Bundle Adjustment

Bundle adjustment is a critical optimization technique in the Structure from Motion (SfM) pipeline. It refines the estimated 3D structure and camera parameters by minimizing the reprojection error, thus improving the accuracy of the reconstructed model. This report provides a detailed technical overview of bundle adjustment, including the mathematical formulation, parameter estimation, and optimization process.

3.1. Mathematical Formulation

The primary objective of bundle adjustment is to minimize the sum of squared reprojection errors across all images and 3D points. The optimization problem can be formulated as:

$$\min_{\mathbf{P},\mathbf{X}} \sum_{i,j} \|\mathbf{p}_{ij} - \pi(\mathbf{P}_i, \mathbf{X}_j)\|^2 \tag{10}$$

where:

- $\mathbf{P}_i$ represents the parameters of the i-th camera, including both intrinsic and extrinsic parameters.
- $\mathbf{X}_j$ represents the 3D coordinates of the j-th point.
- $\mathbf{p}_{ij}$ is the observed image point of the j-th 3D point in the i-th image.
- $\pi(\mathbf{P}_i, \mathbf{X}_j)$ is the projection function that maps the 3D point $\mathbf{X}_j$ onto the image plane of the i-th camera.

3.2. Camera Parameters

The camera parameters $\mathbf{P}_i$ consist of:

- **Intrinsic Parameters:** These include the focal length, principal point, and distortion coefficients, defining the internal characteristics of the camera.
- **Extrinsic Parameters:** These include the rotation matrix $\mathbf{R}_i$ and translation vector $\mathbf{t}_i$, which describe the camera's position and orientation in the 3D space.

3.3. Point Parameters

The 3D point parameters $\mathbf{X}_j$ represent the coordinates of the reconstructed points in the 3D space. These points are initially estimated through triangulation using matched feature points from multiple images.

3.4. Optimization Process

Bundle adjustment employs nonlinear optimization techniques to iteratively refine the camera and point parameters. The process can be broken down into the following steps:

3.4.1 Jacobian Calculation

The Jacobian matrix $\mathbf{J}$ represents the partial derivatives of the reprojection error with respect to the camera and point parameters. It indicates how changes in the parameters affect the reprojection error. The Jacobian is calculated as follows:

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{p}_{11}}{\partial \mathbf{P}_1} & \cdots & \frac{\partial \mathbf{p}_{1m}}{\partial \mathbf{P}_1} & \cdots & \frac{\partial \mathbf{p}_{11}}{\partial \mathbf{X}_1} & \cdots & \frac{\partial \mathbf{p}_{1m}}{\partial \mathbf{X}_n} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{p}_{k1}}{\partial \mathbf{P}_k} & \cdots & \frac{\partial \mathbf{p}_{km}}{\partial \mathbf{P}_k} & \cdots & \frac{\partial \mathbf{p}_{k1}}{\partial \mathbf{X}_1} & \cdots & \frac{\partial \mathbf{p}_{km}}{\partial \mathbf{X}_n} \end{bmatrix} \tag{11}$$

3.4.2 Update Step

The update step involves computing the parameter updates using the Jacobian and the current error. Depending on the chosen optimization algorithm (e.g., Levenberg-Marquardt or Gauss-Newton), the update step can be expressed as:

$$\Delta \mathbf{p} = -(\mathbf{J}^T\mathbf{J} + \lambda \mathbf{I})^{-1}\mathbf{J}^T\mathbf{e} \tag{12}$$

where:

- $\Delta \mathbf{p}$ represents the updates to the parameters.
- $\mathbf{J}^T$ is the transpose of the Jacobian matrix.
- $\mathbf{e}$ is the vector of reprojection errors.
- λ is a damping factor in the Levenberg-Marquardt algorithm.
- $\mathbf{I}$ is the identity matrix.

3.4.3 Parameter Adjustment

The computed updates $\Delta \mathbf{p}$ are applied to the camera and point parameters:

$$\mathbf{P}_i \leftarrow \mathbf{P}_i + \Delta \mathbf{P}_i \tag{13}$$

$$\mathbf{X}_j \leftarrow \mathbf{X}_j + \Delta \mathbf{X}_j \tag{14}$$

3.4.4 Convergence Check

The optimization process evaluates the change in the reprojection error. If the change is below a predefined threshold, or if a maximum number of iterations is reached, the optimization process terminates. The convergence criterion can be expressed as:

$$\|\Delta \mathbf{p}\| < \epsilon \tag{15}$$

where ϵ is the convergence threshold.

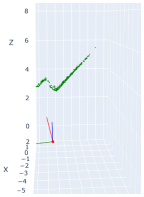

Figure 6. Side view of the book after bundle adjustment

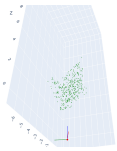

Figure 7. Isometric view of the book after bundle adjustment

3.5. Reconstruction

We saved all the 3D point we got into a polygon file and used meshlab to show a 3D representation of all camera poses and 3D points of the books.

4. Results

By analysing the pictures that were shown in the earlier sections, we are able to make a number of significant deductions. Let's look first at outlier rejection and feature matching. Prior to outlier rejection, a large number of features shared by the two images are initially identified. Unfortunately, noise and false matches make it impossible to interpret all of these features accurately. We may drastically decrease the amount of features to just the most dependable ones (inliers) by using the RANSAC algorithm for outlier rejection. Because it guarantees that only the most reliable and precise matches are utilised for reconstruction, this step is essential to increasing the accuracy of the processes that follow.

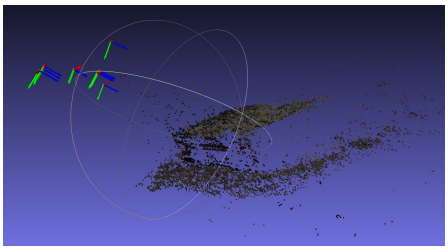

Figure 8. Reconstruction result with Camera Poses and 3D points

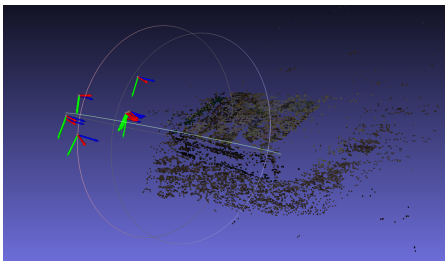

Figure 9. Reconstruction result with Camera Poses and 3D points

Even though RANSAC works well, we found that several traits that are recognised are not related to the main item of interest, which is the book. This problem arose because the book's various orientations across several photos prevented us from applying a mask on it. Thus, several irrelevant characteristics are nevertheless identified, underscoring the difficulty of reliable object identification from different angles.

The points acquired during triangulation and cheirality verification verify that the reconstructed points are in front of the camera, hence verifying the precision of our camera pose estimation. This is further corroborated by the reprojection plot, which demonstrates how closely the original and reprojected points align. The precise estimation of camera parameters using the Perspective-n-Points (PnP) approach is substantiated by the close alignment, which suggests that reprojection error is negligible.

Even if it captures the majority of the points, the 3D reconstruction of the image is not flawless. The feature matching is the main source of restriction. An imprecise reconstruction resulted from certain crucial sites being either ignored or wrongly matched due to improper feature matching. Enhancing feature matching methods might greatly improve the quality of the 3D reconstruction, maybe by util-

ising machine learning techniques or more resilient algorithms.

5. Conclusion

We successfully reconstructed the image in 3D space, capturing most of the essential features and translating them into a three-dimensional model. Although the outcome was not ideal, it did show the fundamental properties of the Structure from Motion (SfM) method. Further improvements to the reconstruction might be achieved by using more sophisticated feature matching methods. For example, the robustness and precision of the feature correspondences may be improved by employing techniques like ORB or SURF, which may provide superior results in particular cases.

Furthermore, improving the bundle adjustment procedure with more advanced optimisation techniques may greatly improve the 3D model's accuracy and consistency. More accurate and dependable 3D reconstruction could result from the use of advanced bundle adjustment techniques, which could more effectively minimise the reprojection error.

The amount and quality of the dataset is another important aspect that needs to be improved. More detailed information and greater scene coverage can be obtained by using higher-resolution photos and by expanding the number of photographs in the dataset. Finer characteristics can be extracted from higher-quality photos, and more detailed 3D modelling can be achieved by using a larger dataset, which captures more subtleties of the scene.

Additionally, investigating different strategies like Neural Radiance Fields (NeRF) may provide even better 3D reconstruction outcomes. NeRF represents and renders complicated 3D scenes with excellent resolution by using neural networks. It does, however, involve a lot of computation and takes a lot of time and processing power. In spite of this, NeRF has demonstrated amazing promise in producing incredibly accurate and comprehensive 3D reconstructions, indicating that it is a viable area for further research.

All things considered, these advancements—better matching strategies, optimised bundle modification, better datasets, and investigation of cutting-edge approaches like NeRF—may result in more accurate and comprehensive 3D reconstructions. With these improvements, the SfM method would be even more dependable and efficient for a range of applications, including the protection of cultural assets and virtual reality.

6. GitHub Repository

The following link takes you to the github repository of our final project: gitrepo

References

CMSC website which has the pipeline: link