Multiview 3D Reconstruction

Abhishek Aditya BS*
Department of Computer Science
PES University
Bangalore

PES1UG19CS019
abhishek.aditya10@gmail.com

T Vijay Prashant*
Department of Computer Science
PES University
Bangalore

PES1UG19CS536
tvijayprashant@gmail.com

Vishal R*
Department of Computer Science
PES University
Bangalore

PES1UG19CS571
vishalramesh01@gmail.com

Yashas KS*
Department of Computer Science
PES University
Bangalore

PES1UG19CS589
yashas120@gmail.com

***Abstract*—Obtaining 3D representation of scenes has been a challenging task in computer vision and remains to be an active area of research. Numerous techniques have been proposed over the last few years ranging from traditional image processing techniques to advanced methods like deep neural networks, transformer models. In this work, we use traditional image processing techniques to construct 3D point cloud of objects. We use Incremental Structure from Motion (SfM), a popular SfM algorithm for 3D reconstruction for reconstruction. The method is then evaluated using certain 3D reconstruction datasets.**

Keywords—Structure From Motion, Multi View 3D Reconstruction, 3D Reconstruction

I. Introduction

One of the main difficulties in image-based modeling and computer vision is creating a 3D model from 2D images that is as realistic as possible. The development of three-dimensional models from a group of photos is known as 3D reconstruction from multiple photographs. Recent and rapid advances in the domains of autonomous driving and augmented reality, which rely significantly on precise 3D reconstructions of the surrounding world, are approximated by combining depth readings from sensors such as LIDAR, structured light, and other specific sensors. The disadvantage of these sensors is that they require special hardware, which makes them more effective but also more complicated to acquire and use than systems that rely solely on RGB camera systems. Additionally, due to the low albedo and glossy, reflective, or obstructive surfaces, the sensors suffer from noise and missing measurements.

The current advancements in digital cameras, as well as an improvement in picture resolution and clarity, have opened up new approaches to rebuild 3D images utilizing various techniques that employ merely these cameras rather than pricey special sensors, making the reconstruction process relatively affordable.

The goal of the reconstruction is to derive the geometrical structure of a scene from a set of photos, assuming that the camera position and internal parameters are known or can be guessed from the set of images. This is accomplished by employing numerous photos in which the 3D information may be (partially) retrieved by applying the Structure from Motion approach to solve the pixel-wise correspondence problem.

The process of reconstructing a three-dimensional structure using projections acquired from a succession of photographs from diverse viewpoints is known as Structure

*- *Equal Contribution*

from Motion (SfM). This technique produces advanced state-of-the-art findings, but the technique's primary concerns are resilience, precision, completeness, and scalability, which are handled using an incremental approach for the structure from motion. LIDAR-based 3D reconstruction of a scene is costly and prone to artifacts from GPS and IMU. As a result, we will employ the Structure from Motion method, which uses just low-cost camera images to rebuild a 3D scene while also obtaining the camera poses of the monocular camera in relation to the provided scene.

II. Literature Survey

The 3D reconstruction of images from large scale structure from motion using unordered datasets have seen a huge development in recent years. Feature matching between images and bundle adjustment are the two major tasks performed for an unordered SfM. Hence in order to increase the efficiency, the above two methods must be either optimized by parallelizing them or must be restricted to a clever subset of images. There are usually two ways in which the number of the candidate pairs used for the above processes are reduced which are by either indexing the images via a shared visual word or by selecting a subset of images using clustering using a global image descriptor such as GIST. But even these methods tend to miss out a few key image pairs leading to unnecessary fragmentation. Hence the paper [1] proposes an efficient process to match all the pairs of images in a database without any need for an exhaustive test between each image pair.

The method in [1] is built on top of a huge, publicly available visual vocabulary which can accommodate a quantization of 16 million visual words. This vocabulary has 2 layers which use the SIFT descriptors after the Hessian-Affine interest points are extracted for the computation. The images from the database are resampled to a common resolution and the same features are extracted and are quantized. These processed images are used to generate an inverted file where each visual word has a list of all the images where the word appears. The image is only added to this list when a visual word describing the image appears <= 1% of all images in the database. This ensures that the quality and the efficiency of the method is improved. This inverted image file is then used to cluster the images using a threshold which is equal to the minimum number of correspondences required for the epipolar geometry estimation in the SfM pipeline. Using the tracks from the inverted file, the pairwise matches of all the possible two-view combinations are found. These image pairs are then passed to the Bundler which uses the default settings to return a sparse 3D reconstruction of the scene. The drawbacks of this method is that the vocabulary based matching is a little stricter than the conventional method of pairwise matching, hence the method was not able to find the complete track for an object point which led to the Bundler taking longer time to reconstruct the 3D model. Also the point clouds generated by the vocabulary are noisier which causes the correspondences detected with the proposed method to be not very accurate due to the high uncertainty of the triangulation and that they contain more epipolar-consistent miss matches.

The usual incremental SfM flow starts by feature matching between two images, followed by the 3D reconstruction of the scene from the two-view reconstructions which is achieved by repeated addition of the matched images and triangulation of the features matched. This process follows the bundle adjustment of the structure and motion finally yielding the 3D reconstructed scene. This requires a time complexity of $O(n^4)$ for n images which is very high for large scene reconstruction. Hence the paper [2] aims at improving the efficiency and the time complexity of the method by introducing a new preemptive feature matching technique which decreases the pairs of image matching by upto 95% while still recovering adequately good matches for the reconstruction of the scene. A re-triangulation step is also introduced into the pipeline which deals with the problem of accumulated drifts without an explicit loop closing without sacrificing the time-complexity of the method.

In paper [2] a novel preemptive feature matching technique is introduced to identify the good pairs of images efficiently hence decreasing the set of images by 75% - 98%. The technique basically involves matching the first h features of the two images in the pair and if the number of matches is smaller than a threshold the image pair is used. This reduces the time complexity of the process from $O(n^2)$ to O(n). The Bundle adjustment is another major bottleneck in the performance of the SfM, hence to make this process more efficient an improved Preconditioned Conjugate Gradient BA which uses the Hessian matrix instead of the Schur complement is used to bring the time complexity of the process from $O(n^3)$ to O(n). Thus the improvement in the time complexity of both the feature matching and the bundle adjustment gives an opportunity to make the incremental SfM pipeline close to O(n). The algorithm adds a single image at each iteration and runs either a full bundle adjustment or a partial bundle adjustment. After this step the algorithm either continues to the next iteration or a

re-triangulation takes place. The major drawback of this method is that the method might fail for extremely large reconstruction due to the high accumulated errors. Also the number of image pairs matched to test the incremental SfM is $O(n^2)$ which is higher than the vocabulary strategy which chooses the pairs to match in O(n). Hence this becomes a bottleneck to the whole pipeline.

III. Methodology

The basic idea in 3D image reconstruction is that given a set of images $\{I_1, \ldots, I_N\}$ where each image is taken from a different viewpoint, our goal is to use these images to reconstruct a three dimensional representation of the object. More specifically, we will find the motions of the cameras with respect to a world coordinate frame F_W. This motion of cameras is also known as camera projection matrices $\{P_!, \ldots P_N\}$. Using this set of camera projections we will then use different algorithms to recover the 3D structures of the scene.

To perform this, we will construct a pipeline that will consist of two main parts, data association, structure-from-motion (SfM). Data association is used to check whether a pair of images are similar to each other. Two images can be checked for similarity by using image correspondences and robust two-view geometry. Structure-from-motion is responsible for the initial reconstruction using pose estimation and triangulation techniques and refining this using the bundle adjustment algorithm. MVS is then applied on this to get a dense 3D representation.

Data association is the first part of the 3D reconstruction pipeline. Given a set of unstructured images, we first find the connected components in these images. This helps us to find overlapping views in images. To establish the connected components, we use the SIFT [3] algorithm that helps us to extract keypoints from the images. We then perform image or keypoints correspondences using a two-view geometry algorithm. This will map a feature in one image to a similar feature in another image. One of the problems in data associations is that in set of input images N is large, then searching through image pairs becomes intractable and the complexity of querying one image is of the order $O(N.K\text{^}2)$ where K is the number of keypoints in each image. An efficient tree based method is used for image retrieval. This makes the complexity of querying an image reduce to $O(K.b.L)$ where K is the features in the query image, b is the branches in the tree and l is the levels in the tree.

There are three different methods in SfM namely, incremental SfM, global SfM and hierarchical SfM. For our purpose, we used the incremental SfM for generating the sparse 3D reconstruction. The basic idea of incremental SfM is as follows. First we choose two non-panoramic views from the scene graph generated by the data association step. A 8-point algorithm is used to compute the fundamental or the essential matrix. The fundamental matrix can also be thought of as a camera projection which can be decomposed into two matrices P and P'. P' represents the intrinsic camera calibration. We then apply the linear triangulation algorithm to compute the correspondences and obtain the 3D points Bundle adjustment algorithm is then applied to refine the 3D points obtained from the previous step. We then find 2D - 3D correspondences and add more views into the system. 2D - 2D correspondence is first established between the newly added image and the previous image and 2D - 3D correspondence is then established to get the 3D points.

Once we establish the 2D - 3D correspondences for all images, we will use the Perspective-n-Point (PnP) algorithm to compute the pose of the images with respect to the world frame coordinates. Additionally, more 2D - 2D correspondences can be selected in the neighboring images and apply linear triangulation to find their 3D points. This helps us obtain a more dense point cloud for 3D reconstruction. After obtaining the 3D points, the next step is to refine these points using the bundle adjustment algorithm. The bundle adjustment is applied to both the 3D points as well as the camera pose estimates obtained from the PnP algorithm.

Figure 2: Shows sample input images of Gustav II Adolf statue and the corresponding 3D reconstruction on the right (Bundle Adjustment was not applied to this reconstruction).

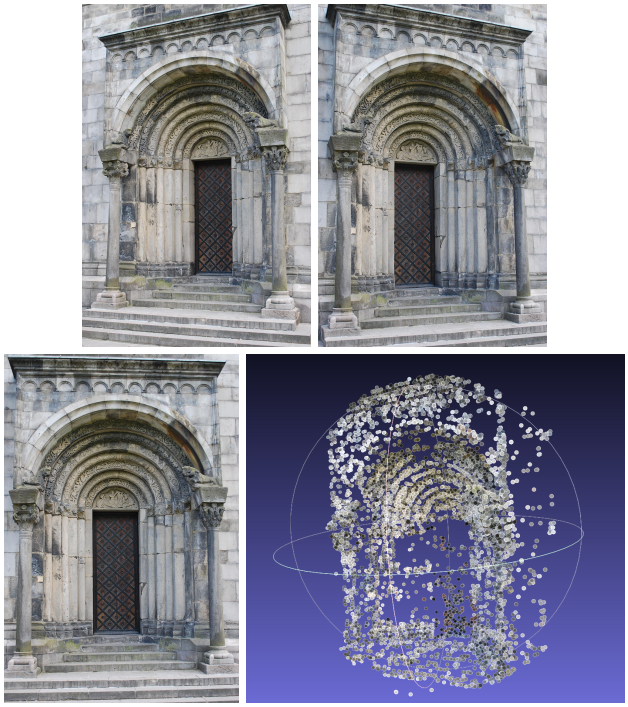

Figure 3: Shows images of a door and the corresponding 3D reconstruction of the door on the right.

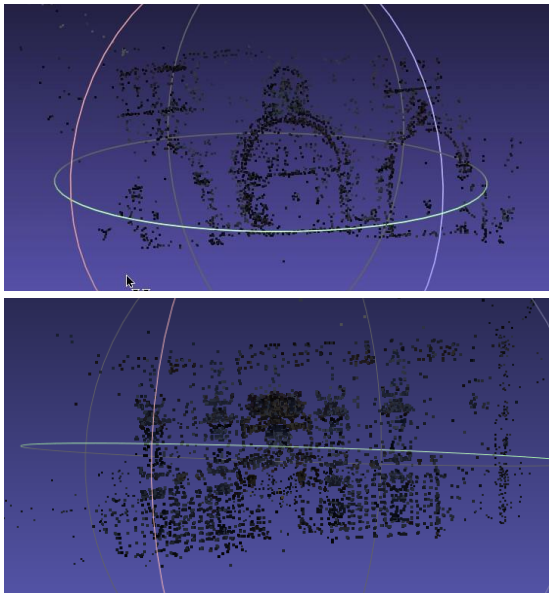

Figure 4: Shows image reconstructed using Bundle adjustment on the left, without Bundle Adjustment on the right.

IV. Experimental Results

For testing our models we used several images of different objects. The dataset for an object will consist of several images of the object taken from different positions and angles. Along with the images, the camera calibration matrix K must be provided. This encodes the information required for obtaining the projection matrix for each image.

Figure 2 shows sample images of Gustav II Adolf statue. Each image in the dataset is different in terms of its position and angles. These images will be passed to our SFM pipeline that we have created to obtain the final 3D reconstruction. Figure 2 (right) shows the output of our SFM pipeline. Figure 3 is a picture of a door. When we feed these images along with its camera parameters, the output 3D representation is obtained as shown in Figure 3 (right).

In our experiments, the Bundle Adjustment algorithm takes a very long time on large images. Hence we were able to only use it for smaller image datasets. We can obtain relatively decent 3D representation without applying bundle adjustment. However, applying bundle adjustment leads to better reconstruction quality.

V. Conclusion

This work presents an implementation of the incremental SFM algorithm using linear triangulation, Perspective-n-Points algorithms to find the pose and reconstruct the image as a 3D sparse model. Additionally bundle adjustment was implemented to enhance the quality of the model. The quality of models generated with bundle adjustment is excellent. Sparse representation of structure from motion lacks any definitive quantitative metrics to compare different models. This model has some disadvantages like the images should be passed to the model in some order meaning the overlap of two images must be present to some extent. Any changes in order will break the model. Although sparse representation is a very important step in generating 3D models from 2D images, dense image reconstruction is required for all practical usage. This dense image can be generated using the Multi-View Stereo (MVS) algorithm.

References

1. Havlena, Michal, and Konrad Schindler. "Vocmatch: Efficient multiview correspondence for structure from motion." In *European Conference on Computer Vision*, pp. 46-60. Springer, Cham, 2014.

2. Wu, Changchang. "Towards linear-time incremental structure from motion." In *2013 International Conference on 3D Vision-3DV 2013*, pp. 127-134. IEEE, 2013.

3. Lowe, David G. "Distinctive image features from scale-invariant keypoints." *International journal of computer vision*60, no. 2 (2004): 91-110.