Indoor Environment Map Building from RGB-D Matterport3D Data

Steven Brewer
Robotic Engineering
Worcester Polytechnic Institute
Worcester, MA, United States
scbrewer@wpi.edu

Peter Nikopoulos
Robotics Engineering
Worcester Polytechnic Institute
Worcester, MA, United States
phnikopoulos@wpi.edu

Chris Rutka
Robotics Engineering
Worcester Polytechnic Institute
Worcester, MA, United States
cjrutka@wpi.edu

***Abstract*—A digital camera views the three-dimensional world as a two-dimensional image. The depth of objects is lost and can only be calculated by taking multiple images from different known camera poses and the same objects in the image are analyzed to form a triangulated distance. Thus, reconstructed three-dimensional maps are generated. Images seen by robots can be used to re-create difficult to access locations, such as the interior of a condemned building in search of survivors.**

***Index Terms*—SIFT, RANSAC, Bundle Adjustment, indoor environment; geometric modeling; semantic modeling; topological modeling; scene reconstruction**

I. Introduction

The Environmental Protection Agency estimates that 75% of the population in the world reside in cities and towns. Of city dwelling population it is further estimated that the average person spends 90% of their day indoors [6]. Current spatial layouts of the indoor environments such as airports, hospitals and shopping malls are desirable but currently limited.

Indoor environment understanding is essential for computer vision, augmented reality, robotics, and scene modeling. Prior to 2010 gathering images with a depth channel was challenging. More recently large dataset have become easier to come by due to the decrease in sensor price [7] and increase in consumer products such as the Kinect. RGB-D /3D datasets have been small areas with limited scene coverage. Matterport3D greatly expanded research opportunities of indoor environments. The Matterport dataset is made up 90 builds with 194,400 RGB-D images. 10,800 of the images are 360 degree color and depth panoramas at human height of the indoor environment [1].

The Matterport3D dataset is comprehensive with accurate association between images which allows for keypoint matching, view overlap prediction, semantic segmentation, and region classification [1] techniques to be investigated. This paper looks to process one of the 90 buildings. The analysis pipeline developed is sufficient to scale to the remaining 89 buildings. Limiting the scope to one building reduced computer resources from over a terabyte of data to 5 gigabytes.

II. Literature Review

2D points can be detected using Scale invariant feature transform (SIFT). SIFT is algorithm that locates keypoints and their feature descriptors. The feature descriptor is a vector that describes the local area around the keypoint. The keypoint is found using the difference in Gaussians

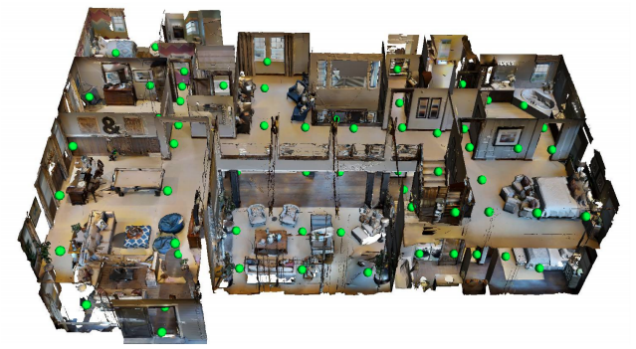

Fig. 1. Matterport3D building example. Green spheres represent image capture locations taken approximately 2.25m apart. [1]

where a Gaussian blur of varying blur rate are subtracted from one another. Extreme points in the x,y image space and blurred image stack become the keypoints. This also done on the image at different scales. The descriptor vector looks at the local neighborhood of the keypoint and computes the gradient vector. The gradient vector is robust to viewpoint changes [2].

Matching 2D points across two images can be done using RANdom SAmpling Consensus (RANSAC). RANSAC is a trial-and-error approach that partitions the key points into outliers and inliers (based on a statistical average) and then uses only the inliers. This is done by first sampling the data. The algorithm assumes the sample points are inliers and computes the statistical model parameters. The remaining datapoints are used to score the model by finding the number of points that fit to the line. This process is repeated for n number of iterations seen in equation 1. The model with the best score is used to find the data association between inliers [4].

$$n = log(1-p)/log(1-(1-e)^s) \tag{1}$$

where e is outlier ratio (outliers/datapoints), s is the number of sampled points, p is the probability [4].

Bundle Adjustment is a state estimation technique for estimating the 3D location of 2D points from sets of images. Bundle adjustment estimates both the 3D position of points seen in a camera frame as well as the 6 DOF position of this camera in space. Practically, bundle adjustment is solved using a large least squares approach that minimizes the preprojective error of all cameras poses and

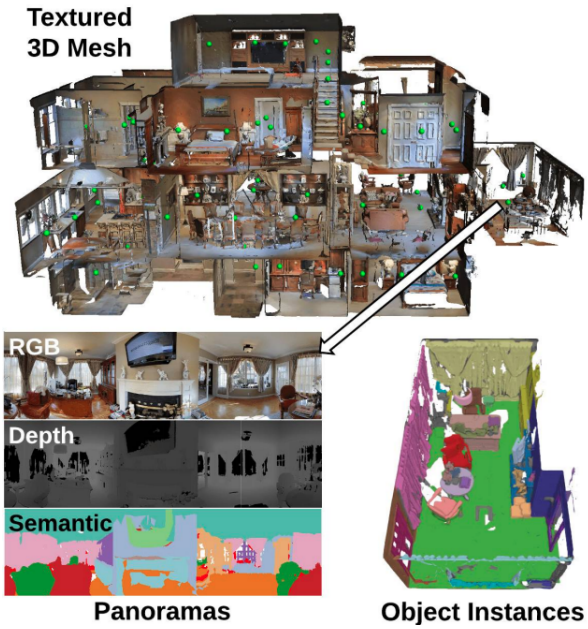

Fig. 2. Example of The Matterport3D dataset with RGB images, depth images, categories, and object semantic segmentations. [1]

keypoints in the dataset at the same time. The algorithm is performed by calculating initial guesses for the positions of all 3D points and camera poses in the scene and then projecting them into 2D points in each camera frame. The 2D positions of projected 3D points and the keypoints within each camera frame are compared, resulting in some error. This error is used to feed the least squares algorithm which attempts to minimize the overall preprojective error of the scene. This process is recursive and happens over several iterations to arrive at a good enough approximation of the scene [5].

Matterport3D is a diverse, large-scale RGB-D dataset consisting of 10,800 panoramic views from 194,400 RGB-D images of 90 building scale scenes. The dataset provides annotations of surface reconstructions, camera poses, and semantic segmentations. Matterport 3D is an ideal dataset for this application because it features a wide variety of complete indoor scenes. The dataset features the precise global alignment of diverse panoramic set of views enables a variety of computer vision algorithms to be applied. Tasks such as keypoint matching, view overlap prediction, and more can be applied to this dataset to validate computer vision systems [1].

3D reconstruction using structure from motion initializes from image pairs. It utilizes keypoint matching, fundamental matrix and essential matrix estimation, camera pose estimation, and bundle adjustment [6]. Classical bundle adjustment can be improved by incorporating depth estimates as seen in Melbouci et al. [8]. An unmanned aerial vehicle (UAV) approached depth cameras data by exploiting the range data and was used for feature tracking [10]. Combining geometric registration with global optimization based online processes can add robustness to geometric alignment errors [9].

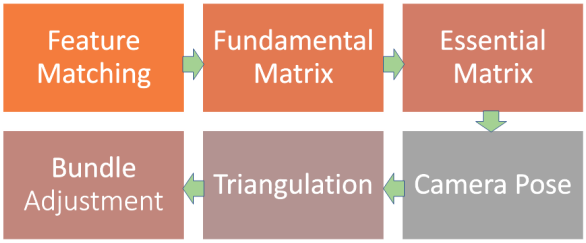

Fig. 3. Flowchart of structure from motion steps

III. Problem Description

The Matterport3D dataset contains interior information of building scenes. Based on the datasets, a visual reconstruction was generated that allows for human understanding of what the robot sees as it travels safely through the building. This has applications for disaster recovery or simply needing to get a robot from one place to the next. Mapping and 3D reconstruction is key for a robot to learn human level understanding of an indoor environment. From this understanding, the robot can learn how to move to avoid an obstacle or identify objects/people which are key to the mission success.

Matterport3D data contains RGB images, depth data, categories, and object semantic segmentations data. The full dataset is over a terabyte. One medium size building is approximately five gigabytes of data and was used to develop the techniques to convert the data into a 3D reconstruction. The RBG images coupled with the depth data must be used to identify overlap between neighboring images. Room by room a 3-Dimensional point cloud must be generated.

IV. Methodology

The implementation of structure from motion was used and followed using Melcouci et al. [8] which accounted for the depth data during bundle adjustment. The general workflow was as follows: first key points and point descriptors were identified using SIFT on image pairs. Outlier rejection utilized RANSAC. The Fundamental matrix was then estimated with the aid of RANSAC. Utilizing the camera intrinsics and the fundamental matrix, the essential matrix was estimated. From the essential matrix camera poses were estimated. This resulted in four possible camera poses between image pairs, triangulation solved this issue. The bundle adjustment set was performed using depth data to increase accuracy. The flow chart of the structure from motion process can be seen in figure 3.

Bundle adjustment minimizes 2D reprojection error of the 3D points.The error is the difference between the estimate projection Q_i through the homogeneous transform P and the observation $q_{i,j}$.The implementation used here match that in [8] where the SLAM error is defined in equation 2.

$$\epsilon_{slam} = \sum_{i=1}^{N_p} \sum_{j \in A_i} \rho_s(q_{i,j} - \pi(K P_j Q_i), a_s) \qquad (2)$$

where K is the intrinsic camera parameters seen in equation 3 where u and v are the camera center x and y coordinates and f are the respective the focal lengths.

$$K = \begin{bmatrix} f_u & 0 & u_0 \\ 0 & f_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

P is the homogeneous transformation matrix for world coordinates to camera frame.

$$P = \begin{bmatrix} R & t \\ 0_{1x3} & 1 \end{bmatrix} \tag{4}$$

$Q = (X, Y, Z, 1)^T$ is the homogeneous 3D points in the world frame. $q = (x, y, z)^T$is the 2D homogeneous point in the camera frame. $\pi(q) = (x/z, y/z)^T$ is the perspective projection and a_s is the reprojection threshold given by the median absolute difference [8].

Depth data was integrated into bundle adjustment to further constrain the problem, improve accuracy and robustness using equation 5. d_i represents the depth value.

$$\epsilon_{depth} = \sum_{i=1}^{N_p} \sum_{j \in A_i} \sum_{K \in A_i}^{k \neq j} \rho_d(q_{i,j} - \pi(K P_j P_k^{-1} \pi^{-1}(q_i, k, d_i, k)), a_d) \tag{5}$$

Optimization of the depth and SLAM error was done equation 6. The minimization used Levenberg-Marqurdt method.

$$\epsilon = \epsilon_{depth} + \epsilon_{slam} \tag{6}$$

V. Results and Discussion

A. Multithreading

The process of structure from motion is computationally intensive, to speed up the system multithreading was used. The computer splits up the program into data packages and sends them to multiple cores in the computers CPU. Doing this allows the program to run concurrently saving time and speeding up the process.

B. Results

The Matterport3D images were set up on a rig with three diagonal cameras. The rig was rotated 60 degrees in each subsequent image. There was little overlap between the images that proved challenging. Had the geometry of the camera rig been known a smart setup for recreation could be made. In lieu of that a naive approach was take that relied on image matching. Figure 4 shows the dining room scene. Some of the views were too zoomed in to properly match while others had little scene overlap. Bundle adjustment out performs EKF-SLAM where scene overlap is minimal. even so the bundle adjustment method struggled with many of the views.

Figure 5 was a subset of the dining room that was provided enough matches reconstruct the scene.

It can be seen in Figure 6 the mismatch in light fixture however the doorway to the adjacent room can be seen clearly. The added depth data allowed for better bundle adjustment.

C. Discussion

Matterport3D's camera and rotation set up is such that there is a small amount of overlap between images. Levenberg-Marquardt optimization does not work when the number of residuals is less than the number of variables. Figure 7 illustrates the images the system typically struggled with.

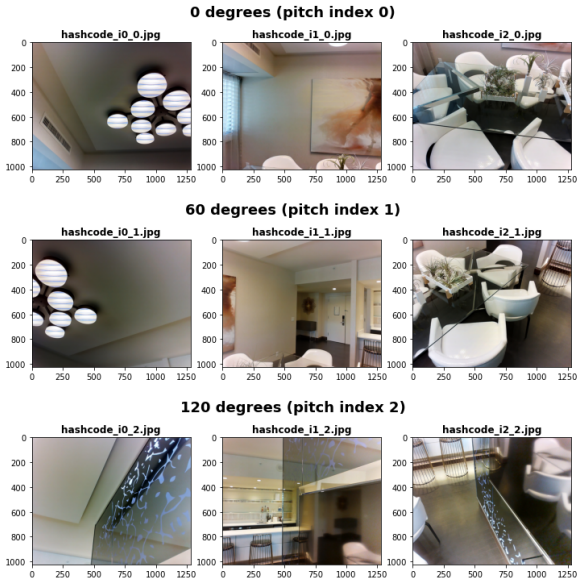

Fig. 4. Dining room scene

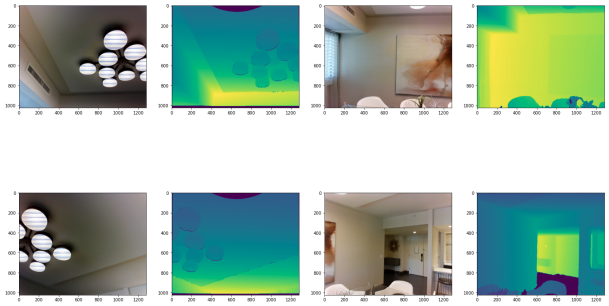

Fig. 5. Matched dining room 2D Scene with depth map

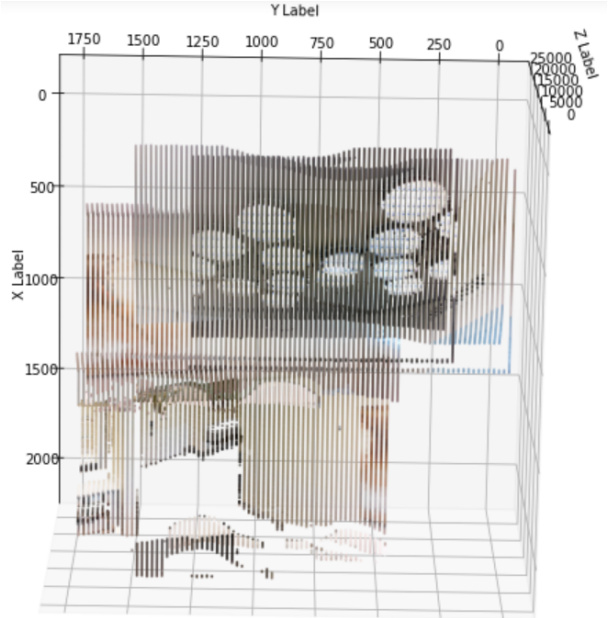

Fig. 6. 3D reconstruction of the Dining room

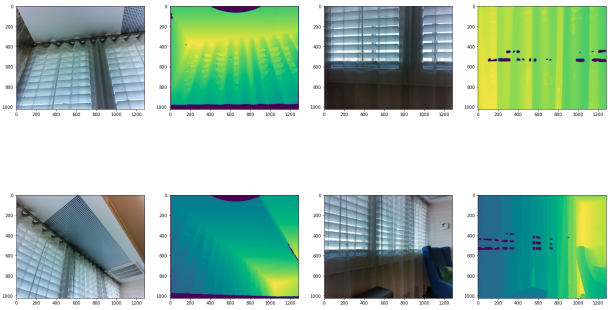

Fig. 7. 2D image and depth data of an example where the system struggled to recreate the environment

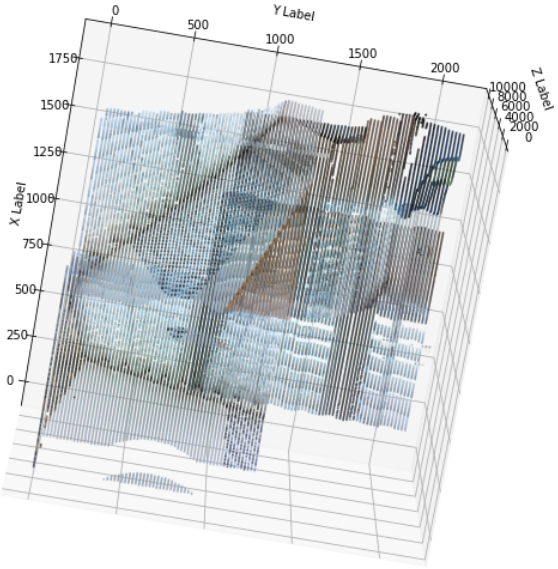

Fig. 8. 3D reconstruction of challenging scene

The 3D reconstruction mismatch is a function of minimal overlap and non unique features. The window blinds are very similar in the views and challenging for the system to match features.

Without the depth data the scale invariance of the pose estimation coupled with the minimal keypoints had little success at 3D reconstruction as seen in Figure 9.

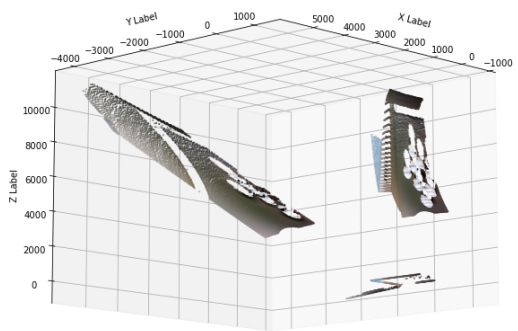

Fig. 9. 3D reconstruction without depth data

VI. Conclusion

3D reconstruction can be completed using stereo image cameras. The sift algorithm successfully detects key points in images. Feature matching using RANSAC allows the key-points in images to be matched and from the matched key points the fundamental matrix was estimated. Using the camera intrinsics and the fundamental matrix the essential matrix was estimated. Pose estimation was done from the essential matrix, triangulation improved the estimate. Bundle adjustment was performed first using solely Levenberg-Marquardt optimization of the slam reprojection error. The depth data further constrained the problem and allowed for more accurate 3D reconstruction. The Matterport3D data set had minimal overlap between images hampering the system's success however when tested on RGB-D data with over lap between images the results were successful.

Expanding on these results would entail further testing on the larger dataset as a whole. This would also include smarter image matching such using the images above, below, left, and right of the current image frame would be paired during feature matching and bundle adjustment. A focus on edge case errors and better outlier dampening should be explored. The proposed method could also be expanded beyond RGB-D images and tested on LiDAR data with corresponding images.

References

[1] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, Y. Zhang Matterport3D: Learning from RGB-D Data in Indoor Environments International Conference on 3D Vision (3DV 2017)

[2] Lowe, D. G. (2004), 'Distinctive Image Features from Scale-Invariant Keypoints', Int. J. Comput. Vision 60 (2), 91–110.

[3] H. R. Kher and V. K. Thakar, "Scale Invariant Feature Transform Based Image Matching and Registration," 2014 Fifth International Conference on Signal and Image Processing, 2014, pp. 50-55, doi: 10.1109/ICSIP.2014.12.

[4] Martin A. Fischler and Robert C. Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24, 6 (June 1981), 381–395. DOI:https://doi.org/10.1145/358669.358692

[5] D. Martinec and T. Pajdla, "Robust Rotation and Translation Estimation in Multiview Reconstruction," 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1-8, doi: 10.1109/CVPR.2007.383115.

[6] Kang, Zhizhong, Juntao Yang, Zhou Yang, and Sai Cheng. 2020. "A Review of Techniques for 3D Reconstruction of Indoor Environments" ISPRS International Journal of Geo-Information 9, no. 5: 330. https://doi.org/10.3390/ijgi9050330

[7] Armeni, I., Sax, S., Zamir, A., Savarese, S. (2017). Joint 2D-3D-Semantic Data for Indoor Scene Understanding. ArXiv, abs/1702.01105.

[8] K. Melbouci, S. N. Collette, V. Gay-Bellile, O. Ait-Aider, M. Carrier and M. Dhome, "Bundle adjustment revisited for SLAM with RGBD sensors," 2015 14th IAPR International Conference on Machine Vision Applications (MVA), 2015, pp. 166-169, doi: 10.1109/MVA.2015.7153159.

[9] Choi, Sungjoon, Zhou, Qian-Yi Koltun, Vladlen. (2015). Robust Reconstruction of Indoor Scenes.10.1109/CVPR.2015.7299195.

[10] Belmonte, L., Castillo, José, Fernández-Caballero, Antonio, Almansa-Valverde, Sergio, Morales, R. (2015). Flying Depth Camera for Indoor Mapping and Localization. 10.1007/978-3-319-19695-4-25.

[11] Belmonte, L., Morales, R. Fernández-Caballero, Antonio. (2019). Computer Vision in Autonomous Unmanned Aerial Vehicles—A Systematic Mapping Study. Applied Sciences. 9. 3196. 10.3390/app9153196.