Mapping the MIT Stata Center: Large-scale Integrated Visual and RGB-D SLAM

Maurice F. Fallon, Hordur Johannsson, Michael Kaess, David M. Rosen, Elias Muggler and John J. Leonard

***Abstract*—This paper describes progress towards creating an integrated large-scale visual and RGB-D mapping and localization system to operate in the MIT Stata Center. The output of a real-time, temporally scalable 6-DOF visual SLAM system is used to generate low fidelity maps that are used by the Kinect Monte Carlo Localization (KMCL) algorithm. This localization algorithm can track the camera pose during aggressive motion and can aid in recovery from visual odometry failures. The localization algorithm uses dense depth information to track its location in the map, which is less sensitive to large viewpoint changes than feature-based approaches, e.g. traversing in opposite direction up and down a hallway. The low fidelity map also makes the system more resilient to clutter and small changes in the environment. The integration of the localization algorithm with the mapping algorithm enables the system to operate in novel environments and allows for robust navigation through the mapped area—even under aggressive motion. A major part of this project has been the collection of a large dataset of the ten-floor MIT Stata Center with a PR2 robot, which currently consists of approximately 40 kilometers of distance traveled. This paper describes ongoing efforts to obtain centimenter-level ground-truth for the robot motion, using prior building models.**

I. Introduction

There are many challenges in building autonomous, large-scale visual mapping systems that can operate over extended periods of time. In this paper we report on our progress towards developing such a system. We have collected an extensive vision dataset using the PR2 with the purpose of mapping the Stata Center (Fig. 1). Many areas have been visited repeatedly to capture the longer term temporal shape of the building. We are working towards recovering ground truth for this entire dataset by automated alignment to floor plans.

II. Vision SLAM

We have developed a vision based mapping system that can construct large scale maps using an RGB-D camera [1] fusing data collected over many months. The mapping system integrates motion estimates from multiple sources, including an IMU and wheel odometry in addition to visual odometry (VO). An example of a map produced by the system is shown in Fig. 2.

The system consists of several modules: (i) visual odometry, (ii) appearance based loop proposals, (iii) view registration and (iv) map estimation. The visual odometry is implemented using the Fast Odometry from VISion (FOVIS) library [2].

This work was partially supported by ONR grants N00014-10-1-0936, N00014-11-1-0688, and N00014-12-10020. The authors are with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA mfallon,hordurj,kaess, rosen,muggler,jleonard@mit.edu

Fig. 1. Ray and Maria Stata Center designed by Frank O. Gehry and completed in 2004. The 10 floor, 67,000 m^2 building is the home to the Computer Science and Artificial Intelligence Laboratory.

The map is represented as a pose graph, where the nodes are camera poses and the edges are constraints between the poses. A configuration of the poses that best satisfies all the constraints is computed using incremental smoothing and mapping (iSAM) [3]. These constraints can come from the visual odometry module and other sensors that are on the robot, e.g. an IMU gives absolute constraints on roll and pitch and relative constraints on heading.

An appearance based index is used to recognize when a place is revisited [4]. It generates proposals which are verified by computing the rigid body transformation between two camera frames. This is done by minimizing the reprojection error of the corresponding features. This transformation can then be used to add additional constraints into the graph and improve the overall accuracy of the map.

To avoid continous growth of the graph, new poses are only added when the camera senses parts of the environment that are not sensed by other poses. This limits the number of nodes in the graph—yet information acquired when going from one place to another is still used to continually improve the map.

III. Creating a 3-D Building Model

Using the poses estimates from the vision SLAM system, evaluating a 3D model of the environment represented by point clouds is a trival matter of reprojection. An example of such a model is as illustrated in Figure 2 (top). This representation provides a visually appealing reconstruction of the building and could be further processed to form a volumetric mesh reconstruction. Additionally a voxel–based occupancy tree could be used to aid path planning.

For the purposes of robust localization, we can also generate a simplified model of large planar structure which is (1) unlikely to change (2) anchored on the pose graph and (3) of very small size (<10MB for the entire building). This approach extracts large planes from a single point cloud using RANSAC

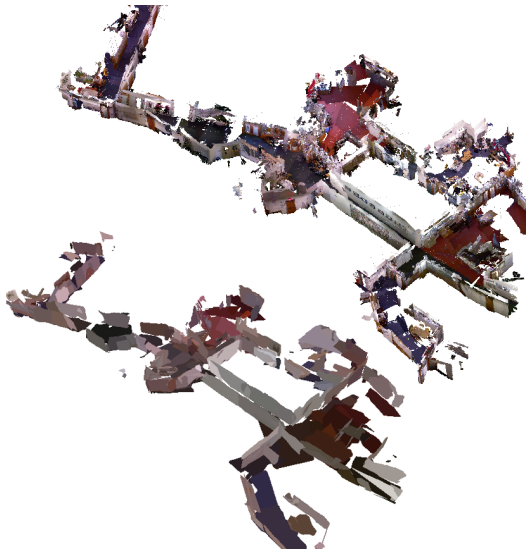

Fig. 2. Output of the Vision SLAM system: A dense 3-D point cloud of the 2nd floor of the Stata Center (top). Input to the KMCL Localization system: A simplified model made up of only large planar objects (bottom).

and then progressively grows the planes using nearby scans. The resultant plane–based model is also presented in Figure 2 (bottom).

IV. RGB-D Monte Carlo Localization

In our previous publication we presented Kinect Monte Carlo Localization (KMCL, [5]) which localized using as input a prior map produced in this manner. A particle filter algorithm propogated an estimate of the robot's pose using visual odometry and as many as 1000 particles. Using an optimized GPU rendering algorithm simulated depth images could be efficiently generated and compared to the measured data to evaluate the relative likelihood of each particle. Because the algorithm could support such a large number of particles, robust localization in motions as fast as 5 m/s were possible while only using the RGB-D data as input.

V. KMCL and vision SLAM integration

The two systems described above, KMCL and vision SLAM, have their strengths and weakness. The localization works well for tracking during aggressive motion, recovery from visual odometry failures, and tracking multiple hypothesis. Its main limitation is that it requires a prior map – thus limiting its use in new environments.

While the mapping system is not as robust as the localization system, it can construct a map of an unknown environment. These limitations of the two system are orthogonal to each other. So by integrating these two system it is possible construct a new system that improves over using the systems individually. We are not the first to separate the tracking and mapping, that was used with a monocular camera by Parallel Tracking and Mapping (PTAM, [6]) with good results.

Our approach is to integrate the two systems in the following way. Both the systems receive their input from the visual odometry. The mapping system starts constructing a map as before. As the map is built, planes are extracted and attached to poses in the map. This allows for easy readjustment of the map when loop-closure corrections occur. The plane map is then forwarded to the KMCL tracker.

When the KMCL tracker starts receiving the map it can track the robot position in parallel to the mapping process. Two things that change in this scenario: (i) each particle tracks the location relative to a nearby pose, thus if the map is corrected the particles will move with that correction; (ii) the robot might be exploring (in that case the map will not be available and the tracker will be made aware of that situation). It is important to note that the tracker is always tracking position relative to the map, it does not care about the global accuracy of the map.

In addition to providing the map to the tracker, the mapper can also propose feature-based relocalization using the bad of words. This allows the tracker to introduce particles in the proposed place and track that hypothesis to see if it receives enough support. The mapper can use this information to validate potential loop closures.

The tracker sends tracking information to the mapper and can use this information to search for new loop closures that can be used to refine the map. It can also be used to recover from motion failures, and because the tracker uses depth information it can better track when moving in the opposite direction to that originally explored, e.g. up and down a hallway—which is a challenge for feature based methods.

In this way these two algorithms complement each other and by combining them into an integrated system, better performance in both the mapping and localization processes can be achieved.

VI. Groundtruth Dataset

As mentioned above during this project we set the lofty goal of mapping the entire MIT Stata Center—not simply a single time but continously over an extended period of time. This has resulted in an interesting and realistic 10 floor dataset. It is particularly relevent given recent interest in long term SLAM and autonomy as evidenced by a series of workshops including the 2011 RSS SLAM Evaluation Workshop.

Some figures of merit regarding the dataset are as follows:

1) Period covered: January 2011 to May 2012.
2) Total distance: 40km over 10 floors.
3) Total filesize: about 2 Terabytes.
4) Total time: about 26 hours, 68 sessions
5) Sensors collected: Standard PR2 sensors plus a head–mounted Microsoft Kinect. Various configurations.

To the best of our knowledge this dataset is the largest dataset of continous operation in a single location. We hope that it will provide the basis for future research tackling the scaling limitations of SLAM as well as more generally contributing to long-term understanding and autonomy for robotic systems.

Recently we having been working towards determining ground-truthing for the data logs and making the dataset available to the research community. The first part of the

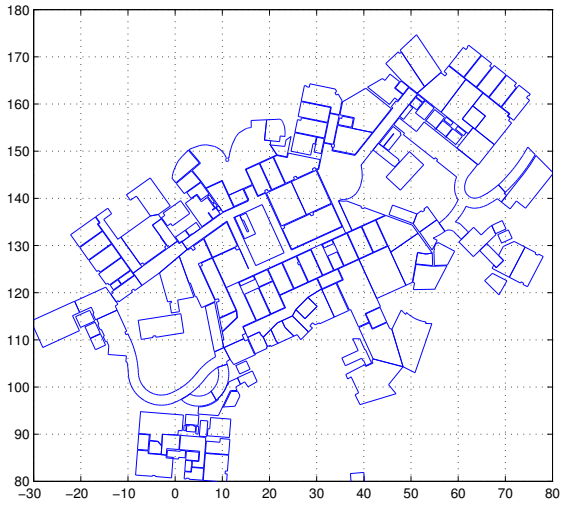

Fig. 3. Floorplan of the 3rd floor of the Stata Center, with units in meters.

dataset is now available from this location:
http://projects.csail.mit.edu/stata/

A. Requirements of a Ground Truthed Dataset

The progress of visual mapping, SLAM and 3D reconstruction have been motivated by and benchmarked using datasets such as the Middlebury Evaluation System [7] which can be used to make a clear comparison between different algorithms and implementations.

Additionally, simulated and real-world datasets such as the Victoria Park, Manhattan and Intel datasets have been widely used to compare the graphical SLAM backend, for example [3]. This dataset is intended to extend this approach to Visual SLAM by the provision of an extensive, rich multi-sensor dataset. Previously Sturm *et al.* [8] have developed an automated system for the comparison of 3D RGB-D SLAM systems using a motion capture system to provide ground truth—the Freiburg dataset.

However, due to the constraints of the motion capture system the Freiburg dataset is limited 36 minutes and 400 meters across 16 experiments. Our dataset is intended to be symbiotic with the Freiburg dataset but at a much larger scale.

B. Extraction of Ground Truth

We recognise that the utility of any mapping database is vastly increased by access to ground-truth robot and sensor pose measurements. Thankfully our building is relatively new and our institute's services department maintains accurate and reliable 2D building plans of each floor of the building as illustrated in Figure 3.

The ground truthing process involves a combination of (1) incremental alignment of scans from the PR2's Hokuyo URG-04LX-UG01 base laser, (2) global alignment of scans with the model where clearly correct alignment can be observed and (3) smoothing of these 2–3 second subproblems. The scan-matching algorithm is that presented in [9]. An example alignment is illustrated in Figure 4. We have taken particular care that the poses estimated are in no way correlated across the map—providing instantanous measurements of error for each pose to an accuracy of 2–3 cm.

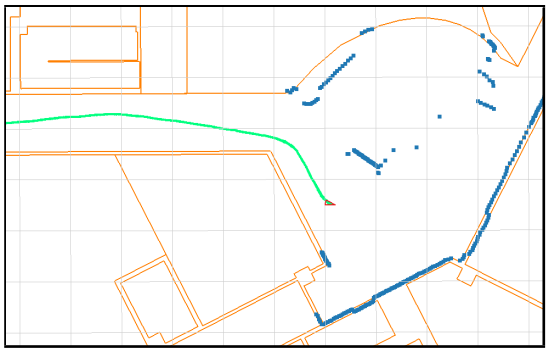

Fig. 4. Example of a portion of the ground truthed PR2 trajectory (green) and the alignment of a single scan to the floor plan. Note that door recesses and clutter cannot be matched to the model.

The PR2's ROS coordinate frame manager, **tf**, maintains the internal relative calibration of each sensor. This allows us to provide pose estimates of its tilting Hokuyo, Microsoft Kinect and stereo camera.

VII. Acknowledgements

We would like to acknowledge Alper Aydemir who made available the floor plans described in Section VI. Models of MIT and KTH are available from:
http://www.csc.kth.se/~aydemir/floorplans.html

References

[1] H. Johannsson, M. Kaess, M. F. Fallon, and J. J. Leonard, "Temporally scalable visual slam using a reduced pose graph," Tech. Rep. MIT-CSAIL-TR-2012-013, Computer Science and Artificial Intelligence Laboratory, MIT, May 2012.

[2] A. Huang, A. Bachrach, P. Henry, M. Krainin, D. Maturana, D. Fox, and N. Roy, "Visual odometry and mapping for autonomous flight using an RGB-D camera," in *Proc. of the Intl. Symp. of Robotics Research (ISRR)*, (Flagstaff, USA), Aug. 2011.

[3] M. Kaess, A. Ranganathan, and F. Dellaert, "iSAM: Incremental smoothing and mapping," *IEEE Trans. Robotics*, vol. 24, pp. 1365–1378, Dec. 2008.

[4] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Intl. Conf. on Computer Vision (ICCV)*, vol. 2, (Los Alamitos, CA, USA), p. 1470, IEEE Computer Society, 2003.

[5] M. Fallon, H. Johannsson, and J. Leonard, "Efficient scene simulation for robust Monte Carlo localization using an RGB-D camera," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, (St. Paul, MN), pp. 1663–1670, May 2012.

[6] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *IEEE and ACM Intl. Sym. on Mixed and Augmented Reality (ISMAR)*, (Nara, Japan), pp. 225–234, Nov. 2007.

[7] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, pp. 7–42, April–June 2002.

[8] J. Sturm, S. Magnenat, N. Engelhard, F. Pomerleau, F. Colas, W. Burgard, D. Cremers, and R. Siegwart, "Towards a benchmark for RGB-D SLAM evaluation," in *Proc. of the RGB-D Workshop on Advanced Reasoning with Depth Cameras at Robotics: Science and Systems Conf. (RSS)*, (Los Angeles, USA), June 2011.

[9] A. Bachrach, S. Prentice, R. He, and N. Roy, "Range - robust autonomous navigation in gps-denied environments," *J. of Field Robotics*, vol. 28, pp. 644–666, September 2011.