Robust Tumbling Target Reconstruction through Fusion of Vision and LIDAR for Autonomous Rendezvous and Docking On-Orbit

Jose Padial
Aerospace Robotics Laboratory, Stanford University
450 Lomita Mall, Stanford, CA 94305
jpadial@stanford.edu

Abstract

This milestones report details progress toward 3D target reconstruction through fusion of vision and LIDAR data. The utility of using both vision and LIDAR for on-orbit target reconstruction is first presented. The technical approach, including a new camera-LIDAR Structure from Motion (SfM) framework, is next presented, though this technical framework is provisional at present. Validation methodology and expected results through simulation and hardware experimentation is outlined.. Finally, specific milestones met thusfar are outlined.

1. Introduction

Target reconstruction is a necessary capability for safe and reliable autonomous rendezvous and docking capability on orbit. Vision is a natural sensor for object reconstruction as it is capable of providing frame-to-frame point correspondence and texture information. The field of Structure from Motion is a well-developed one, providing the capability to map a target (structure) and recover the camera motion, up to a similarity transformation (unknown overall scale), assuming calibrated cameras. This scale ambiguity is a problem for real operation on orbit, and cannot be resolved without more information.

Range data (LIDAR) provide 3D structure data directly. When using 3D LIDAR technology (*e.g*. Flash LIDAR), it is possible to solve for scan-to-scan correspondence through alignment of point clouds (typically, with a form of Iterative Closest Point algorithm [1]). Conversely, line-scanning (2D) LIDAR can only solve the correspondence problem in loop closure situations, other than for the degenerate case where the axis of target rotation is perpendicular to the line-scan plane. In terrestrial applications, the use of 3D LIDAR is most likely the correct choice. However, for applications on small satellite chaser vehicles, limitations on power, size, and weight may/will dictate use of the line-scanning LIDAR. The estimation framework proposed here is designed for the camera and line-scanning LIDAR sensor configuration. However, the results will be extensible to use with more complex LIDAR technology.

2. Approach

The approach presented here toward 3D reconstruction through fusion of visual imagery and LIDAR data is evolving. The following subsections outline building blocks that are being developed toward the algorithmic solution. Note that this outline is not a complete account of all the methods being investigated. For instance, there is no mention of projective factorization methods. However, projective factorization may be utilized in the final solution formulation.

2.1. Frame-to-Frame Relative Pose Estimation

Frame-to-frame relative pose estimation is accomplished via two-frame epipolar constraints. Assuming a calibrated camera, the Essential matrix E is estimated using the well-known 8-point algorithm [2]. Robust feature correspondences are input to the 8-point algorithm. These feature correspondences can be generated by any suitable method, *e.g*. SIFT, SURF, Harris-Laplace. For this implementation, SIFT features are used for correspondence. Accuracy of the Essential matrix estimation is vital to success of the algorithm. As such, performance may be substantially improved by inclusion of a nonlinear optimization step wherein the estimate of the Essential matrix is refined further. As yet, this has not been included. However, nonlinear refinement of the Essential matrix is well-known.

From the Essential matrix E, we extract the rotation and translation (${}^{i+1}R^{i}$, ${}^{i+1}t^{i/i+1}$) between frames C_i, C_{i+1}. A well-known method using the SVD is used to extract rotation and translation from E, yielding four possible solutions, from which one solution is selected based on chirality (triangulated feature depths should be positive in the camera frame).

Outlier rejection can be performed on the frame-to-frame

pose estimates by enforcing some measure of camera trajectory smoothness. Especially for high frame rate data, we know that the camera linear and angular velocities should vary smoothly, so we can reject camera motions that imply discontinuous velocity profiles. This has not yet been incorporated into the solution strategy.

2.2. Vision-Range Correspondence

Correspondence between image pixels and range returns is necessary for effective fusion of these two data sources. We assume that the camera and LIDAR are co-located such that the relative translation between the two sensors is small, and as such there is no occlusion from ranged point to the image plane. Under this assumption, and wiith known extrinsic calibration of the LIDAR to camera (rotation, translation), and known intrinsic calibration of the camera, we can unambiguously project range scans onto the image plane:

$$^{C}\tilde{x}_j = K[^{C}R^{L} \quad ^{C}t^{L/C}]^{L}\tilde{X}_j \tag{1}$$

Where K is the camera intrinsic matrix, $^{C}\tilde{x}_j$ is the homogeneous 2D image of ranged point j in camera frame C, and $^{L}\tilde{X}_j$ is the homogeneous 3D ranged point in the LIDAR frame L. Let x_i be an image interest point pixel location, let x_j be a range projection pixel location, let α be a distance threshold, and let M be the set of vision-range matches. A vision-range match is identified by the simple Euclidean distance measure:

$$if \ \|x_i - x_j\|_2 \leq \alpha \rightarrow (x_i, x_j) \in M \tag{2}$$

There is an inherent problem of sparseness in this vision-range correspondence. If we simply look for matches between our range points and robust interest points, *e.g.* SIFT features, then we will have very little matches. This will diminish the ability to successfully fuse these vision and LIDAR data. Instead, if we search for image interest points along the projected scan line, and then search for interest point matches in the successive image frame using the knowledge of the pairwise epipolar geometry estimated from robust correspondence, then we have the potential to inject more of the range information into the SfM solution. These image interest points may be simple gradients or Harris corners, as opposed to more complex robust interest points. Also, the matching may be done by a more simple method such as normalized correlation. This is a key investigation that is planned for the upcoming month.

2.3. Projective Depths and Absolute Translation Scale

Unlike the canonical SfM vision-only problem, with the addition of LIDAR sensing we can directly measure projective depth to 3D features. Given frame-to-frame relative pose estimates, we can formulate a global optimization problem for the projective scale to each 3D feature, and the proper scale of the frame-to-frame translation. These initial depth and scale estimates can then be used to further refine global pose estimates. The details of the global optimization problem will only be roughly outlined here. Full description of the optimization will appear in the final report.

The optimization formulation is adapted from the formulation presented in [3]. The formulation makes use of a clever cross-product trick in order to re-shape the problem into a manageable form. This paper adapts the canonical SVD solution from this framework into a new convex optimization problem that utilizes the direct measurements of range from range-vision correspondence in order to yield a scale unambiguous estimate of 3D structure and relative camera poses.

Let λ_i^j be the projective depth of feature j from camera frame $\{C_i\}$, measured with image pixel coordinates x_i^j. Let $^{i+1}R^i$, $^{i+1}t^{i/i+1}$ be the relative rotation and translation estimates from frame $\{C_i\}$ to $\{C_{i+1}\}$. Relative pose estimates are obtained initially from frame-to-frame Essential matrix estimation. Let $\gamma_{i,i+1}$ be the scale factor estimate between the relative translation estimate $^{i+1}t^{i/i+1}$ and the true relative translation. Let β_i^k be the projective depth of feature k from the camera center $\{C_i\}$, derived from the LIDAR measurement that was matched with vision feature k by the vision-range correspondence step. Finally, we define $[\tilde{x}_i^j]_x$ to be the cross-product matrix of homogeneous vector $\tilde{x}_i^j$.

$$([\tilde{x}_{i+1}^j]_x \, ^{i+1}R^i x_i^j)\lambda_i^j + ([\tilde{x}_{i+1}^j]_x \, ^{i+1}t^{i/i+1})\gamma_{i,i+1} = 0 \tag{3}$$

$$([\tilde{x}_{i+1}^j]_x \, ^{i+1}t^{i/i+1})\gamma_{i,i+1} = -\beta_i^k([\tilde{x}_{i+1}^k]_x \, ^{i+1}R^i x_i^k) \tag{4}$$

Relation (3) holds $\forall j = 1, ..., N_{i,i+1}^{\lambda}$, where $N_{i,i+1}^{\lambda}$ is the number of feature matches from frame $\{C_i\}$ to $\{C_{i+1}\}$ that have no vision-range correspondence. Relation (4) holds $\forall j = 1, ..., N_{i,i+1}^{\beta}$, where $N_{i,i+1}^{\beta}$ is the number of feature matches with vision-range correspondence in frame $\{C_i\}$. We take the relations (3), (4) for all frames i and features j, k, and form a large linear matrix equality $M\lambda = b$.

Further, we formulate constraints across frame-to-frame pairs by incorporating triangulation for features that are observed over 3 (or more) frames.

$$(^{i+1}R^i x_i^j)\lambda_{i+1}^j - x_{i+1}^j\lambda_{i+1}^j + ^{i+1}\, t^{i/i+1}\gamma_{i,i+1} = 0 \tag{5}$$

$$x_{i+1}^k\lambda_{i+1}^k - ^{i+1}\, t^{i/i+1})\gamma_{i,i+1} = \beta_i^k \, ^{i+1}R^i x_i^k \tag{6}$$

Once again, equations (5), (6) differ in that the latter is for a feature k obsserved in frames $i, i+1$ for which there is measured projective depth β_i^k in frames i from ranging. We take these constaints (5), (4), for all frames i, and features j, k for which the relations can be formed, and construct another linear matrix equality $A\lambda = c$.

Now we are able to form the following convex optimization problem to solve for our vector λ of unknown projective feature depths and translation scale factors.

$$\begin{aligned} \underset{\lambda,\epsilon}{\text{minimize}} \quad & \|M\lambda - b\|_2^2 + C\|\epsilon|_2^2 \\ \text{subject to} \quad & A\lambda + c + \epsilon = 0 \\ & D\lambda \succeq \zeta \end{aligned}$$

Slack variables ϵ are introduced to allow for minor deviations from the linear constraints. However, we penalize the size of the slack values by inclusion of the term $C\|\epsilon|_2^2$ in the objective, where C is some (large) positive scalar. Further, we include the elementwise constraint that each projective feature depth is greater than ζ. This requires some knowledge of the distance of the camera from the target, for which we can use our range returns to generate a conservative lower bound ζ. The inclusion of this constraint is a safety precaution against the solution $\lambda = 0$. The matrix D selects only the projective scale depths from λ, omitting the translation scale variables.

The sparseness of the matrices M, A make this a quickly solvable convex optimization problem.

2.4. Global Optimization and Refinement

Global refinement of the relative pose estimates and projective depths is necessary for accurate target reconstruction. In canonical SfM, this global refinement is termed bundle adjustment, and is typically a minimization of summed, squared reprojection error. With the addition of LIDAR sensing, we have an extra ability to also minimize error in the 3D ranged points.

Global refinement may be a single global optimization problem, or may be an iterative procedure involving estimation of relative poses given a prior estimate of projective scale factors, followed by re-estimation of projective scale factors given the new estimates of relative pose. This is a key area of work that will be investigated in the coming month.

3. Validation Methodology and Expected Results

Validation of the 3D reconstruction method will be conducted through simulated and experimental results. First, the method will be tested on data from our tumbling target simulation environment. In the simulation environment, a target model is flown with a specified state trajectory, *e.g.* torque free motion. An observer is populated in the environment, and simulated range returns and images (with simulated image features from 3D points) generated, as shown in figures 1, 2. From these simulated range returns and images, it is possible to test algorithms in a noise-free environment where perfect truth data is avaialble. Furthermore, this environment allows for noise to be injected at various stages of the measurement pipeline in a known way.

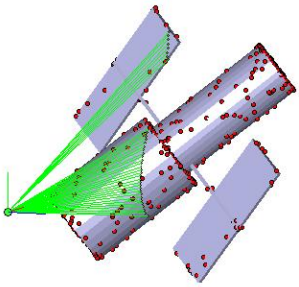

Figure 1. Simulation environment with simulated vision features (red) and simulated range scan (green).

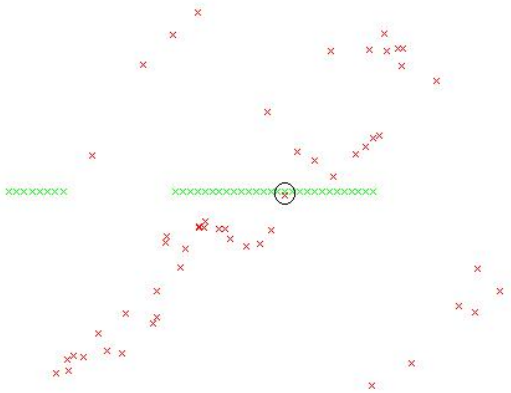

Figure 2. Simulated image of vision features (red) and 3D range scans (green) projected onto image plane. A candidate vision-range correspondence match is circled in black.

Hardware results are expected, and will be the 3D reconstruction of a real-world target measured by a co-located camera and URG Hokuyo line-scanning LIDAR. We have these sensors in the Aerospace Robotics Laboratory, and so collecting this data should not be problematic in the next month. While there will not be hard truth data with which to compare the real world reconstruction, the effectiveness of the reconstruction should be apparent by the quality of the final product.

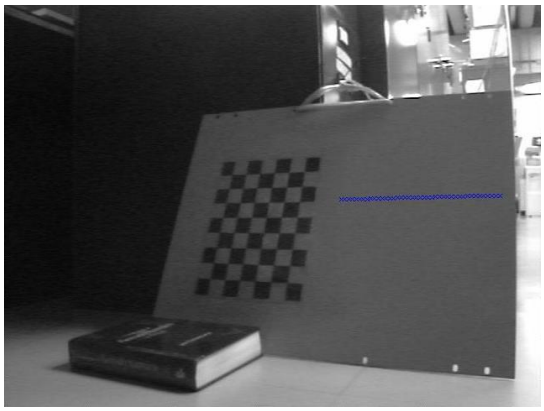

Figure 3. Projection of range scan onto image using camera-LIDAR extrinsics (rotation, translation) as estimated by camera-LIDAR calibration.

4. Intermediate Results

The main results to date include:

1. Simulated Frame-to-Frame Relative Pose Estimation: The frame-to-frame relative pose estimation method outlined in Section 2.1 has been successfully implemented in simulation. With zero noise and sufficient simulated image correspondences, the relative pose is estimated well by the code. The accuracy given varying noise will be investigated in the next month.

2. Simulated Depth Scale and Translation Scale Estimation: As outlined in Section 2.3, this optimization method is nearly working for simulated noise-free data. The recovery of true scale (given vision-range correspondences), is not quite working as yet. I anticipate this will be working shortly.

3. Simulated and Experimental Camera-LIDAR Calibration: This code, using the method of Zhang and Pless [4], is working. The calibration is perfect for noise-free simulation data. The accuracy of the calibration on the hardware has not yet been quantified. Figure 3 shows the projection of a line scan on a calibration image using the result of camera-LIDAR calibration.

References

[1] P. J. Besl and N. D. McKay. A method for registration of 3-d shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14:239–256, February 1992.

[2] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2 edition, 2003.

[3] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry. *An Invitation to 3D Vision: From Images to Geometric Models*. Springer Verlag, 2003.

[4] Q. Zhang and R. Pless. Extrinsic calibration of a camera and laser range finder (improves camera calibration). *2004 IEEERSJ International Conference on Intelligent Robots and Systems IROS IEEE Cat No04CH37566*, 3(314):2301–2306, 2004.