Tumbling Target Reconstruction and Pose Estimation through Fusion of Monocular Vision and Sparse-Pattern Range Data

Jose Padial[*], Marcus Hammond[*], Sean Augenstein[**], and Stephen M. Rock[*,†]

[*]Department of Aeronautics & Astronautics, Stanford University, Stanford, CA, USA
[**]Skybox Imaging, Mountain View, CA, USA
[†]Monterey Bay Aquarium Research Institute, Moss Landing, CA, USA

***Abstract*— A framework for 3D target reconstruction and relative pose estimation through fusion of vision and sparse-pattern range data (*e.g.* line-scanning LIDAR) is presented. The algorithm augments previous work in monocular vision-only SLAM/SfM to incorporate range data into the overall solution. The aim of this work is to enable a more dense reconstruction with accurate relative pose estimation that is unambiguous in scale. In order to incorporate range data, a linear estimator is presented to estimate the overall scale factor using vision-range correspondence. A motivating mission is the use of resource-constrained micro- and nano-satellites to perform autonomous rendezvous and docking operations with uncommunicative, tumbling targets, about which little or no prior information is available. The rationale for the approach is explained, and an algorithm is presented. The implementation using a modified Rao-Blackwellised particle filter is described and tested. Results from numerical simulations are presented that demonstrate the performance and viability of the approach.**

I. Introduction

Accurate estimates of target geometry and relative pose are necessary for safe and reliable autonomous rendezvous and docking. When dealing with an uncommunicative target for which no prior information is known, the target shape and relative pose must be estimated using only exteroceptive sensors. Stereo vision enables 3D feature tracking and reconstruction through triangulation, but depth estimation accuracy is limited by the distance between the cameras. For the problem being considered, the observer/chaser craft is a micro- or nano-satellite. Based on the size of these craft, the ratio of feature distance to available stereo baseline is very small, making accurate depth estimation with stereo techniques difficult. For this reason, this approach will focus on using monocular vision and augmenting it with range data.

Mapping and pose estimation using monocular vision data is a well-researched area, broadly covered by the fields of Structure from Motion (SfM) and vision-only SLAM. However, there are limitations when using only monocular vision. One limitation is that the map and pose estimates are only recoverable up to an overall scale factor. It is impossible to determine whether the target is large, far away, and moving quickly, or small, close and moving slowly relative to the chase satellite. This scale ambiguity is a problem for real operation on orbit, and cannot be resolved without more information. A second limitation is that map density is often quite low when building a map of tracked vision features.

Range data (LIDAR) provide 3D structure 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 [3], or using 3D feature tracking). However, these 3D ranging devices are high energy consumers, and as such may not be suitable for use on micro- or nano-satellites. Therefore, the focus of this paper is limited to the use of sparse-pattern LIDAR sensing, *e.g.* a line-scanning device. The key limitation assumed in this work is that the range sensor pattern or resolution is not capable of providing reliable frame-to-frame correspondence.

The estimation formulation presented fuses monocular vision and sparse-pattern range data for accurate pose estimation and target reconstruction. This work extends a vision-only SLAM/SfM hybrid estimation algorithm [2], based on a modified Rao-Blackwellised particle filter (RBPF), to include the estimation of an overall scale factor through vision-range correspondence. This scale factor is essential for projection of the range data into the body-fixed target frame and for recovery of true target pose relative to the chaser. At a fundamental level, visual feature tracking is used to estimate a scale-ambiguous state history, and vision-range correspondences are used as measurements to estimate an overall scale factor to recover the scale-unambiguous state history. With a scale-unambigous state history, the range measurements at each time step can be projected into a body frame target map that is considerably more dense than that recovered from the previous vision-only formulation.

II. Related Work

The fusion of range and vision data for 3D reconstruction is not new. Liu *et al.* proposed a method of automatic alignment of 2D image sequences with 3D range data [10] that assumes the presence of strong lines present in the scene structure. While this assumption is suitable for many urban/man-made scenes, it is unsuitable for observation of natural terrain/debris (*e.g.* asteroid) or targets for which one cannot assume there are strong edges. Mastin *et al.* proposed a 2D-3D registration technique based on the maximization of mutual information between 2D images and 3D LIDAR features projected onto the 2D image plane [11]. Their method is specifically aimed at registration of airborne LIDAR measurements with aerial imagery of urban scenes. They

explored different methods for evaluating mutual information between images and LIDAR projections, *e.g.* the mutual information between elevation in LIDAR and luminance in the optical image, where higher elevations of the point cloud are rendered with higher intensities.

Progress has been made in high-resolution 3D reconstruction through fusion of optical imagery and time-of-flight (ToF) ranging. In [5] the authors present an algorithm for improved 3D resolution ("super resolution") using Markov Random Fields. More recently, these authors and colleagues have demonstrated in [8] a multi-view system that fuses imagery and ToF sensing for impressive, dense reconstruction. The approach presented in this paper differs from that of [8] in that it is assumed that a 3D ToF ranging sensor is currently infeasible for the power and space constraints of a nano-satellite. As such, there is far less information available from the ranging sensor pattern considered in this work, and visual feature tracking must be relied upon for frame-to-frame correspondence.

A crucial component of achieving good vision-range registration is resolving the scale factor ambiguity that arises from monocular vision SfM. A number of approaches have been taken to estimate absolute scale factor, each making their own assumptions about the scene's structure and motion, or leveraging additional information.

In [15], Nutzi *et al.* present two different schemes involving inertial measurements to resolve scale ambiguity while performing monocular SfM. The implicit assumption in this approach is that the scene is stationary in an inertial reference frame. In the problem addressed by this paper, both target and observer are free-floating, so this assumption does not hold.

Another method used to estimate absolute scale is Depth from Defocus (DfD). In [9], images are taken with a single camera at several focal settings. The setting that produces the sharpest image is used to infer the absolute depth. While useful for providing a rough estimate of scene depth, this approach suffers from performance degradation with distance similar to stereo, and is not suitable for accurate reconstructions at more than a few meters depth.

Prior knowledge of relative geometry or motion can be leveraged to achieve scale estimates. Scaramuzza *et al.* presented a novel approach for estimating absolute scale by taking advantage of non-holonomic rolling constraints on the motion of car-mounted cameras [16]. By judicious choice of camera placement on a car, they were able to estimate the scale whenever the car turned. Since the approach in this paper is limited to having no prior target knowledge and assumes free-floating bodies, such an approach could not be taken.

III. Approach

The approach presented here extends work in vision-only tumbling target pose estimation and reconstruction, as presented in [1], [2], to incorporate range data for improved map density and recovery of a scale-unambiguous target pose and map estimate. Monocular vision feature tracking is used for frame-to-frame correspondence in order to estimate a scale-ambiguous target pose trajectory, and vision-range correspondences are used as measurements to estimate an overall scale factor to recover the scale-unambiguous map and pose trajectory.

Section III-A outlines the quantities estimated. A brief overview of Rao-Blackwellised particle filters is given in Section III-B, where the reader is directed toward [1] for additional detail on the implementation of the RBPF for tumbling target estimation. The hybrid SLAM/SfM motion prediction algorithm used in this work, which allows for smooth and physically realistic motion prediction of a tumbling target with less computational burden [2], is outlined in Section III-C. Section III-D presents the main contribution of this paper: a new method for overall scale factor estimation using vision-range correspondence between tracked visual features and range returns. Scale factor estimation is necessary for accurate projection of range returns into the target body frame. Section III-E provides a brief note on camera-LIDAR calibration. Finally, initialization of the particle set pose distribution by use of two-frame batch Structure from Motion (SfM) estimation is briefly summarized in Section III-F.

A. State Definition

The estimation variables in the tumbling target problem considered here are target poses $\bar{s}_t$, a visual feature map m of the target that is scale ambiguous, and overall scale α. The pose at time-step t is defined as the orientation, angular velocity, and position of the moving target with respect to the observing craft $\bar{x}_p$ as given in (1). This is in contrast to a traditional SLAM formulation, where the pose vector describes the moving observer pose, with the target's pose known due to it being a static, unmoving environment.

$$\bar{s} = \begin{bmatrix} \bar{\theta} & \bar{\omega} & \bar{x}_p \end{bmatrix}_t^T \tag{1}$$

A visual feature map of the target is constructed simultaneously. The map is expressed as the positions of N points, relative to the reference point p on the target, in the target's frame F. It should be noted that this vision feature map is scale ambiguous as it is tracked solely with bearings-only (monocular vision) measurements.

$$m = \left\{ \ \bar{m}_{1/p}, \ \bar{m}_{2/p}, \ \cdots, \bar{m}_{j/p} \ \cdots, \bar{m}_{N/p} \ \right\}^F \tag{2}$$

The pose relates 3D locations between the target's reference frame F and the camera's reference frame C:

$$\bar{x}_j^C = \mathbf{R}(\bar{\theta})^{C/F} \bar{x}_{j/p}^F + \bar{x}_p \tag{3}$$

where the rotation matrix $\mathbf{R}()^{C/F}$ is a function of the relative orientation $\bar{\theta}$.

Additionally, and newly presented in this paper, an overall scale factor α is estimated. When using bearings-only measurements for pose estimation and mapping, the final solution is only recoverable up to an overall scale factor. By incorporating range data into the solution, it is possible to

estimate this scale factor and recover a scale-unambiguous estimate of target pose and shape.

B. Rao-Blackwellised Particle Filter Formulation

Particle filters are a sequential Monte Carlo method of stochastic estimation that can track multi-modal belief distributions. One specific class of particle filters, known as Rao-Blackwellised particle filters, have the added advantage that they can maintain a much larger map of features than other Bayesian estimators [13], [17]. They have been previously applied in the vision-only SLAM/SFM problem [1], [6]. Accurate operation was achieved at frame-rate (30Hz) with 50 particles[6].

As an overview, the underlying distribution of target pose is estimated by particle sampling. Each particle contains an estimate of the visual feature map m along with its target pose estimate, $\bar{s}$. In addition, and newly presented in this work, each particle also contains an estimate of overall scale α as detailed in Section III-D.

A novel hybrid pose prediction algorithm, outlined in Section III-C and first presented in [2], is used to propagate the particle distribution forward in time. The pose prediction step is run for each particle i in the filter (containing M particles in total).

Following pose prediction, each particle's pose vector is perturbed with process noise to generate particle diversity. Following the approach of the FastSLAM 2.0 algorithm [12], this process noise takes into account the latest measurement z_t. FastSLAM 2.0 helps focus the proposal particle cloud towards the highest likelihood regions of the state space.

The Rao-Blackwellised particle filter proceeds to estimate the 3D locations of the visual features and weight the particles based on how well the predicted feature locations agree with measurements, and then to the resampling step, where likely particles are maintained and unlikely particles are discarded. Further detail on the specifics of the RBPF application to tumbling target tracking can be found in [1].

C. Hybrid Motion Prediction

A hybrid motion prediction algorithm is implemented in order to obtain smooth and physically realistic pose prediction of a tumbling target with low computational burden [2]. The algorithm combines concepts from two existing approaches to pose tracking: Bayesian estimation methods and measurement inversion techniques. Rotation is predicted in a Bayesian filter using the available process model. Translation is predicted via measurement inversion, to avoid using the part of the process model with large covariance noise. Using the rotational process model preserves a feasible smooth trajectory for relative orientation. Further, given relative orientation, a translation-only measurement inversion is not susceptible to jumps and non-smoothness, and is a fast calculation.

The rotational states are predicted using (4) (without noise), where $\mathbf{M}(\bar{\theta}_{t-1}^{[i]})$ is the gradient of the nonlinear attitude update equation, linearized about the current state.

$$\begin{aligned} \bar{\theta}_t^{[i]} &= \bar{\theta}_{t-1}^{[i]} + \Delta t \cdot \mathbf{M}(\bar{\theta}_{t-1}^{[i]})\bar{\omega}_{t-1}^{[i]} \\ \bar{\omega}_t^{[i]} &= \bar{\omega}_{t-1}^{[i]} \end{aligned} \tag{4}$$

Using the latest measurements available, $\bar{z}_t$, each particle uses its predicted orientation $\bar{\theta}_t^{[i]}$ and estimate of feature locations $m^{[i]}$ to perform a linear least squares measurement inversion for the translational states, yielding $\bar{x}_{p_t}^{[i]}$:

$$\bar{x}_{p_t} = \bar{x}_{p_{t-1}} + \Delta\bar{x}_{LS}^{C} \tag{5}$$

Space limitations prohibit greater discussion of this approach, but more details on the hybrid motion update estimation scheme and its motivations can be found in [2].

D. Overall Scale Factor Estimation

In order to incorporate range data into the existing monocular vision-only filter, absolute scale α must be estimated. The range data considered here is incapable of providing frame-to-frame correspondence, and so vision-range correspondence must be relied upon in order to estimate scale. Scale could be added to the pose vector, $\bar{s}$, that is estimated through re-sampling, but this would have negative consequences. The most severe negative consequence would be an increase in the filter computational complexity, which is exponential in the length of the pose vector. This would lead to an increased number of particles required for adequate exploration of the state space.

Instead of adding scale to the pose vector, the strategy adopted here is to track scale within each particle. This holds the advantage of letting the tracking of monocular vision features drive re-sampling, with each successful particle holding its best estimate of overall scale. In order to follow this estimation strategy, the full SLAM posterior is factorized in (6).

$$p(y^t|z^t, c^t) = p(m|\alpha, s^t, z^t, c^t)p(\alpha|s^t, z^t, c^t)p(s^t|z^t, c^t) \tag{6}$$

where

$y_t \equiv [\bar{s}_t, m, \alpha]$

$c_t \equiv$ correspondence from measurement to visual map feature

$\text{var}^t \equiv \text{var}_{0:t}$ for any variable *var*

An important observation is that the probability of the visual feature map is not dependent on scale. This is a valid statement as long as bearings-only measurements are used to update visual map feature estimates, which is obeyed in this filter formulation. It should be explicitly noted that in so doing, the definition of the visual feature map m is scale ambiguous. Thus the dependence on scale can be dropped in the conditional probability for the visual map, and each map feature may be tracked independently as shown in [12].

$$p(m|\alpha, s^t, z^t, c^t) = p(m|s^t, z^t, c^t) = \prod_j p(\bar{m}_j|s^t, z^t, c^t) \tag{7}$$

Bayes rule is exploited to break the scale estimation factor in (6) into a familiar motion/measurement model factorization as follows:

$$
\begin{aligned}
p(\alpha_t|s^t, z^t, c^t) &= \frac{p(\bar{z}_t|\alpha_t, s^t, z^{t-1}, c^t)p(\alpha_t|s^t, z^{t-1}, c^t)}{p(\bar{z}_t|s^t, z^{t-1}, c^t)} \\
&= \underbrace{\eta p(\bar{z}_t|\alpha_t, s^t, z^{t-1}, c^t)}_{\text{measurement update}} \underbrace{p(\alpha_t|s^t, z^{t-1}, c^t)}_{\text{motion update}}
\end{aligned} \tag{8}
$$

The first term in (8) is the scale measurement update. The second term is the motion update. As scale is a static parameter, the motion update is trivial. In order to evaluate the scale measurement update, vision-range correspondence must be determined and all possible values of the corresponding vision map feature location integrated over. For a given particle, each vision map feature is tracked in a separate EKF, and thus its probability distribution is estimated by a Gaussian with a mean and covariance.

$$
\begin{aligned}
&p(\bar{z}_t|\alpha_t, s^t, z^{t-1}, c^t) = \\
&= \int p(\bar{z}_t|\bar{m}_{c_t}, \alpha_t, s^t, z^{t-1}) \underbrace{p(\bar{m}_{c_t}|\alpha_t, s^t, z^{t-1})}_{\sim\mathcal{N}(\bar{m}_{c_t}; \bar{\mu}_{c_t}, \Sigma_{c_t})} \mathrm{d}\bar{m}_{c_t}
\end{aligned} \tag{9}
$$

Vision-range correspondence c_t is estimated by finding close alignment between image feature measurements and projections of range measurements onto the image plane. A vision-range correspondence is chosen as the closest such alignment given that the distance between the image feature measurement and the range projection onto the image is below a threshold β:

$$
\begin{aligned}
c_t = \arg\min_i \quad & ||P_I(\bar{m}_i) - P_I(\bar{z}_t)||_2 \\
\text{subject to } & ||P_I(\bar{m}_i) - P_I(\bar{z}_t)||_2 \leq \beta
\end{aligned} \tag{10}
$$

In (10) the function $P_I()$ maps a three dimensional point onto image plane I. The mapping $P_I(\bar{m}_{c_t})$ is the measurement of the visual feature c_t in image I. In order to project the range measurement $\bar{z}_t$ onto the image plane, the camera intrinsic matrix K and the extrinsic rotation and translation from the range sensor frame L to the camera frame C must be known. A calibrated camera-LIDAR system, in which these quantities are known, is assumed in this work, with further detail provided in Section III-E.

All that remains to evaluate (9) is to evaluate the first term in the integral. To do so, the measurement equation for the scale estimation system is defined as follows:

$$
\begin{aligned}
\bar{z}_t &= (\mathbf{R}(\bar{\theta}_t)^{F/C}\bar{x}_{p_t} + \bar{\mu}_{c_t})\alpha_t + \bar{\delta}_z \\
\bar{\delta}_z &\sim \mathcal{N}(0, \Gamma_{z_t})
\end{aligned} \tag{11}
$$

Figure 1 provides a schematic of the scale estimation system. The range sensor measurement $\bar{z}_t$ is the 3D ranged point from the camera origin C_o, expressed in target reference coordinates F. The pose vector tracks translation of the body reference point p in camera coordinates rather than body coordinates, so in order to evaluate the expectation of the measurement the translation vector to point p must be rotated into the body frame basis F representation by $\mathbf{R}^{F/C}(\bar{\theta}_t)$.

Returning to (9), the scale measurement update can now be expressed as the convolution of two Gaussians, which is itself a Gaussian distribution as shown in (12). The covariance

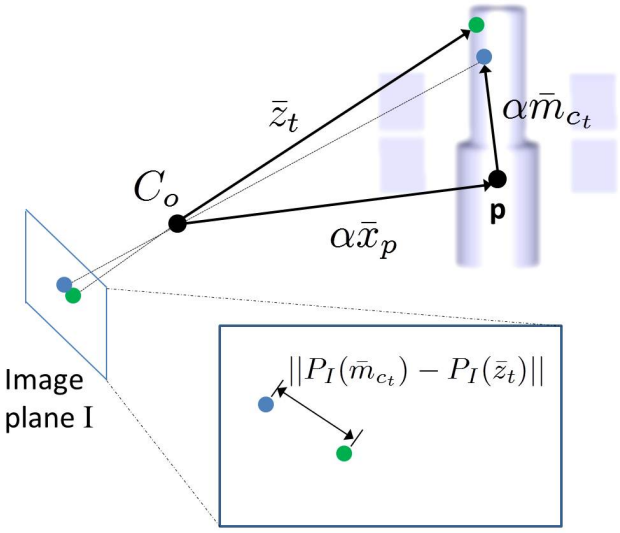


Fig. 1: Scale measurement system

result makes intuitive sense, as it is the sum of the range measurement covariance Γ_{z_t} and the scaled visual feature covariance $\alpha_t^2\Sigma_{c_t}$.

$$
\begin{aligned}
&p(\bar{z}_t|\alpha_t, s^t, z^{t-1}, c^t) = \\
&= \int \underbrace{p(\bar{z}_t|\bar{m}_{c_t}, \alpha_t, s^t, z^{t-1})}_{\sim\mathcal{N}(\bar{z}_t; (\mathbf{R}(\bar{\theta}_t)^{F/C}\bar{x}_{p_t} + \bar{m}_{c_t})\alpha_t, \Gamma_{z_t})} \underbrace{p(\bar{m}_{c_t}|\alpha_t, s^t, z^{t-1})}_{\sim\mathcal{N}(\bar{m}_{c_t}; \mu_{c_t}, \Sigma_{c_t})} \mathrm{d}\bar{m}_{c_t} \\
&\sim \mathcal{N}(\bar{z}_t; (\mathbf{R}(\bar{\theta}_t)^{F/C}\bar{x}_{p_t} + \bar{\mu}_{c_t})\alpha_t, \Gamma_{z_t} + \alpha_t^2\Sigma_{c_t})
\end{aligned} \tag{12}
$$

The linear scale estimation system can be tracked using a standard linear Kalman filter, where each particle maintains its own Kalman filter tracking scale per the Rao-Blackwell factorization:

$$
\begin{aligned}
\overline{\alpha_t} &= \alpha_{t-1} \\
\overline{\Sigma_t} &= \sigma_\gamma^2 + \sigma_{\alpha,t-1}^2 \\
K_t &= \overline{\Sigma_t}C_t^T(C_t\overline{\Sigma_t}C_t^T + \Gamma_{c_t} + \overline{\alpha_t}^2\Sigma_{c_t})^{-1} \\
\alpha_t &= \overline{\alpha_t} + K_t(z_t - C_t\overline{\alpha_t}) \\
\sigma_{\alpha,t}^2 &= (1 - K_tC_t)\overline{\Sigma_t} \\
&\text{for} \\
C_t &= (\mathbf{R}(\bar{\theta}_t)^{F/C}\bar{x}_{p_t} + \bar{\mu}_{c_t})
\end{aligned} \tag{13}
$$

It should be noted that the scale motion noise variance σ_γ^2 is not stricly necessary, as true scale is a static parameter. The motion noise variance is included in the filter to allow for heuristic choice of scale motion noise such that the filter does not become overconfident in the scale estimate.

If an accurate scale estimate is not required until a later time, then it may be preferable to accumulate measurements and perform a batch update on the scale estimate closer to that time. The Rao-Blackwell factorization implies that each particle's scale estimate is conditioned on that particle's full state trajectory, and so from a probabilistic standpoint it is entirely valid to accumulate measurements and perform the update at a later time. For the case when there is no prior

estimate of scale, this batch update reduces to a weighted least squares problem:

$$\alpha_t = (C^T W C)^{-1} C^T W \bar{z}$$
$$\sigma_t^2 = (C^T W C)^{-1}$$

for

$$\underbrace{\begin{bmatrix} \bar{z}_1 \\ \bar{z}_2 \\ \vdots \\ \bar{z}_N \end{bmatrix}}_{\bar{z}} = \underbrace{\begin{bmatrix} \mathbf{R}(\bar{\theta}_{t,1})^{F/C} \bar{x}_{p_{t,1}} + \bar{\mu}_{c_1} \\ \mathbf{R}(\bar{\theta}_{t,2})^{F/C} \bar{x}_{p_{t,2}} + \bar{\mu}_{c_2} \\ \vdots \\ \mathbf{R}(\bar{\theta}_{t,N})^{F/C} \bar{x}_{p_{t,N}} + \bar{\mu}_{c_N} \end{bmatrix}}_{C} \alpha_t + \underbrace{\begin{bmatrix} \bar{\delta}_{z,1} \\ \bar{\delta}_{z,2} \\ \vdots \\ \bar{\delta}_{z,N} \end{bmatrix}}_{\bar{\delta}} \tag{14}$$

$$W \equiv \text{Cov}(\bar{\delta})^{-1} = \begin{bmatrix} \Lambda_1 & & & \\ & \Lambda_2 & & \\ & & \ddots & \\ & & & \Lambda_N \end{bmatrix}$$

$$\Lambda_i \equiv (\Gamma_{z_i} + \Sigma_{c_i})^{-1}$$

E. Camera-LIDAR Calibration

The reconstruction algorithm outlined in this paper assumes a calibrated camera-LIDAR system. Extrinsic calibration of the camera to LIDAR (i.e. rotation and translation from LIDAR frame L to camera frame C) is accomplished using the method of Zhang and Pless [18]. This method utilizes standard camera calibration [4]. Hence, it is assumed that the camera intrinsic matrix K, radial distortion parameters k_c, LIDAR-to-camera frame rotation matrix $\mathbf{R}^{L/C}$ and camera-to-LIDAR frame translation $\bar{t}^C_{L/C}$ are known.

F. Particle Initialization using the Essential Matrix

In order to guide the initial particle distribution, the state space is populated using two-frame batch estimates of observer-target relative motion using epipolar constraints. Assuming a calibrated camera, the Essential matrix E is estimated using the well-known 8-point algorithm [7], provided that there are 8 or more point correspondences. In the event that less than 8 point correspondences are available (but more than 4), E is estimated using the 5-point algorithm presented by Nister [14].

From the Essential matrix E_i, the rotation and translation ($\mathbf{R}^{i+1/i}$, $\bar{t}^{i+1}_{i/i+1}$) between frames C_i, C_{i+1} are extracted. A well-known method using the SVD is used to extract rotation and translation from E, yielding four possible solutions, from which bad solutions are pruned according to chirality (triangulated feature depths should be positive in the camera frame).

IV. Results

Simulated data were generated to verify the performance and viability of the approach. A CAD model of the Hubble telescope was rotated along a torque free motion profile for 100 time-steps. An observer craft with a monocular camera and line-pattern range sensor was simulated at a distance of 12 meters from the Hubble target. Sparse vision feature points were generated randomly on the surface of the Hubble CAD model. At each time-step, feature points were projected onto the image plane of the simulated observer camera, generating virtual feature pixel measurements. Range measurements were simulated by ray-tracing from the observer sensor origin to the CAD model surface.

Noise was added to the virtual pixel measurements and range measurements at each time-step. Vision noise was characterized as pixel noise variances in camera x- and y-directions. Pixel measurement noise was sampled from a zero-mean Gaussian with camera x- and y-pixel variances σ_u^2, σ_v^2. As is common in computer vision literature, σ_u^2, σ_v^2 were set to be 1 pixel each. Noise was added to each virtual range measurement as a function of range. Range noise was sampled from a zero-mean Gaussian with standard deviation of ζr, where r is the true range and ζ is some multiplier. ζ was set to 0.01 to represent a 1% standard deviation as a function of true range. This value matches the specifications of the Hokuyo URG04-LX line-scanning LIDAR.

A total of 50 simulated datasets were generated. For each dataset, the estimation algorithm presented in this paper was run using virtual pixel measurements and range measurements as inputs. Scale error for each run was calculated using the best estimate of scale and truth data. Using this best estimate of scale, a scale unambiguous translation trajectory was estimated and a translation error ϵ_t was calculated as given by (15). Angular velocity error ϵ_ω was calculated for each run as given by (15).

$$\epsilon_t = \frac{1}{N_f} \sum_{t=1}^{N_f} \left(\frac{||\bar{x}_{p_t} - \hat{x}_{p_t}||}{||\bar{x}_{p_t}||} \right)$$
$$\epsilon_\omega = \frac{1}{N_f} \sum_{t=1}^{N_f} \left(\frac{||\bar{\omega}_t - \hat{\omega}_t||}{||\bar{\omega}_t||} \right) \tag{15}$$

In the above equation, N_f is the number of simulation timesteps, t indexes the timestep in a simulation run, $\bar{()}$ is truth, $\hat{()}$ is an estimate, x_p is translation from camera to body frame, and ω is angular velocity.

Table I provides statistics on the scale, translation and angular velocity errors. Mean scale error of 2.14% was achieved, with all scale errors 4.36% or less. Mean translation error of 2.18% was achieved, with all translation errors 4.60% or less. These results suggest that the main source of translation error in our simulations stem from scale estimation error, underlining the importance of accurate scale estimation. Table I presents results for angular rate tracking as well, which do not depend on the scale estimate, but provide statistics for evaluation of the overall effectiveness of the algorithm.

Estimate Error	Mean	Std. Dev.	Max
Scale	2.14%	0.86%	4.36%
Translation	2.18%	0.83%	4.60%
Angular Velocity	3.62%	0.71%	5.77%

TABLE I: Error statistics for 50 simulation runs

Reconstruction results for two of the runs are shown in Figure 2. The first run had the best scale estimation, while the second run had the worst scale estimation. Where the

LIDAR ranged the target, there is considerable increase in map density. Furthermore, the quality of the reconstruction is qualitatively high even in the presence of the worst scale estimation, as seen in the third frame of Figure 2. On the surfaces that were ranged by the LIDAR, the visual features in blue are closely aligned with the green range projection surfaces. On the surfaces not ranged by the LIDAR, the sparsity of the visual features is evident. Figure 4 shows angular rate tracking for the same two runs presented in Figure 2.

In order to demonstrate the utility of the proposed scale estimation approach, it is compared to a naïve scale estimation approach in Figure 3. Scale is estimated in the naïve approach as the mean of all range returns divided by the mean of all visual feature ranges from the camera origin. Figure 3 shows that the proposed method significantly outperforms the naïve method.

V. Conclusion

A method for 3D tumbling target reconstruction and relative pose estimation through fusion of vision and sparse-pattern range data was presented. The algorithm extends a monocular vision-only estimation filter presented in previous work to incorporate range data for true-scale pose estimation and significantly increased map density. A linear estimator was presented to recover the overall scale factor. By proposing correspondences between image features and range returns, the algorithm was able to recover scale factor reliably on simulated data with sensor noise on par with what is achievable in lab. Knowledge of this scale factor made it possible to project the range data into the frame of the monocular vision-only map features, fusing the information from the two sensors into a single, dense reconstruction of the target.

VI. Acknowledgments

This work is supported by NASA National Space Technology Research Fellowship #NNXAM92H and FAA COE-CST Grant #10-C-CST-SU-14. The authors thank Stephen Russell, Andrew Smith, and Nicolas Lee for their assistance and expertise.

References

[1] S. Augenstein and S. Rock. Simultaneous Estimation of Target Pose and 3-D Shape using the FastSLAM Algorithm. In *Proc. AIAA Guidance, Navigation, and Control Conference (GNC)*, Chicago, IL, USA, 2009.

[2] S. Augenstein and S. Rock. Improved Frame-to-Frame Pose Tracking during Vision-Only SLAM/SFM with a Tumbling Target. In *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, 2011.

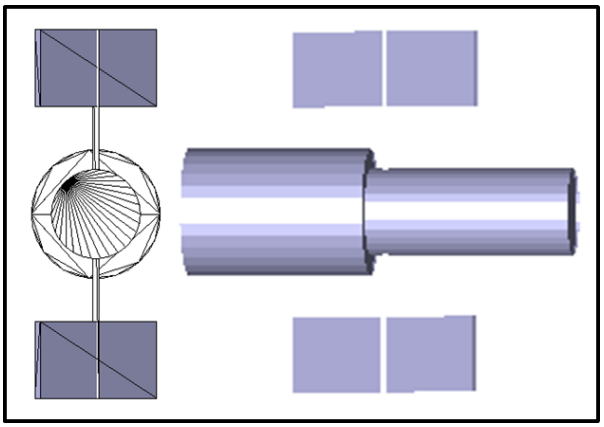
[3] P. J. Besl and N. D. McKay. A method for registration of 3-d shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14:239–256, February 1992.

[4] J. Y. Bouguet. Camera calibration toolbox for Matlab, 2008.

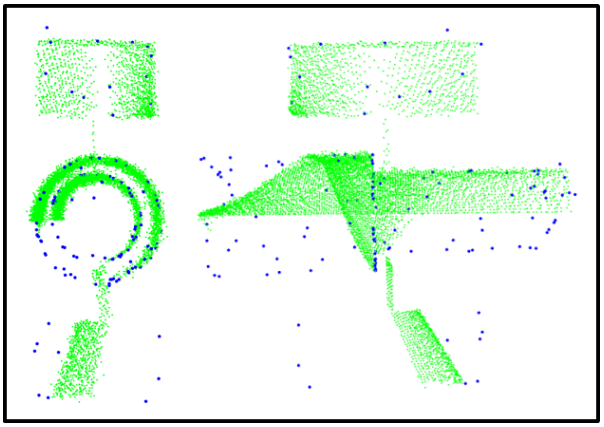
[5] J. Diebel and S. Thrun. An application of markov random fields to range sensing. In *In NIPS*, pages 291–298. MIT Press, 2005.

[6] E. Eade and T. Drummond. Scalable Monocular SLAM. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, New York, NY, USA, 2006.

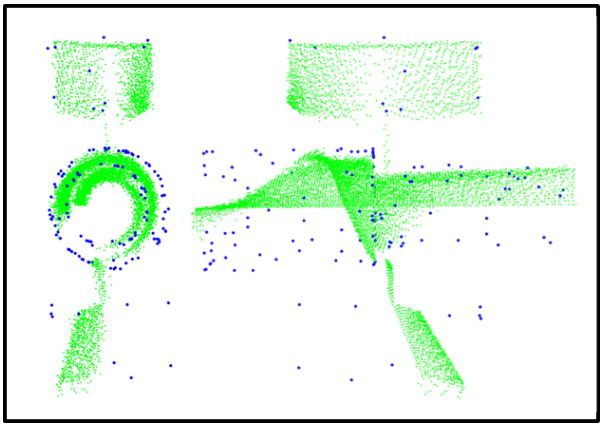
[7] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2 edition, 2003.



(a) Hubble model



(b) Run A: 0.42% scale error, 0.6% translation error, 3.42% angular velocity error.



(c) Run B: 4.36% scale error, 4.60% translation error, 3.68% angular velocity error. Note: this was the largest scale error of simulated runs.

Fig. 2: Hubble model and reconstructed targets corresponding to max-weighted particle in two selected runs. Range projections are shown in green, vision features in blue. Note that because the satellite was only partially scanned with the LIDAR, a number of the image features do not closely correspond with range returns, even though they lie on the satellite surface.

[8] Y. M. Kim, C. Theobalt, J. Diebel, and J. K. B. Miscusik. Multi-view image and tof sensor fusion for dense 3d reconstruction.

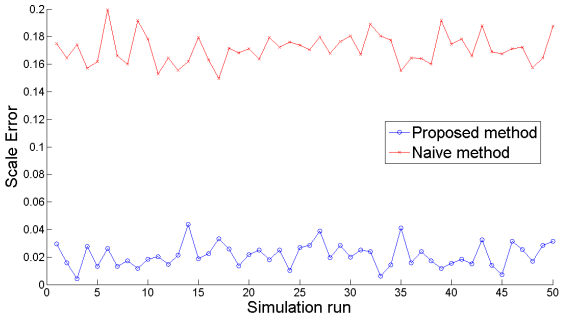
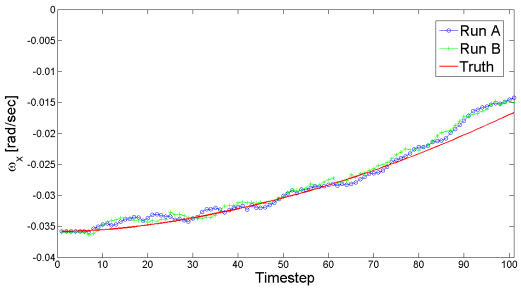
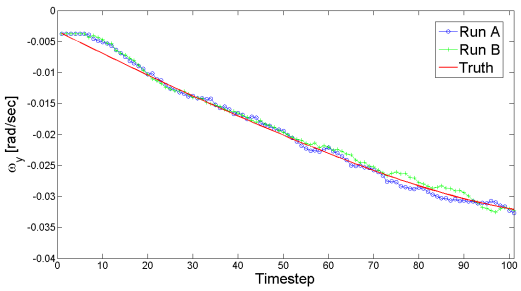


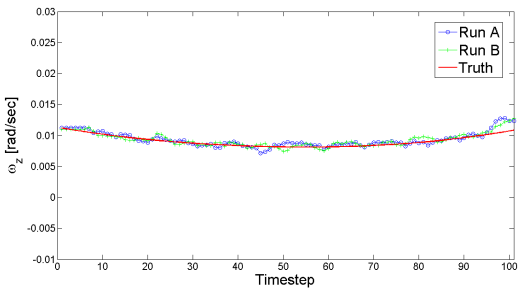
Fig. 3: Comparison of proposed scale estimation approach to naïve method.



(a) ω_x



(b) ω_y



(c) ω_z

Fig. 4: Angular rate estimation for the two runs in Figure 2

[9] A. Kuhl, C. Wöhler, L. Krüger, and H. michael Groß. Monocular 3d scene reconstruction at absolute scales by combination of geometric and real-aperture methods. In *Pattern Recognition. Proc. 28th DAGM Symposium*, pages 607–616. Springer Verlag, 2006.

[10] L. Liu and I. Stamos. Multiview geometry for texture mapping 2d images onto 3d range data. In *In CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2293–2300. IEEE Computer Society, 2006.

[11] A. Mastin, J. Kepner, and J. Fisher. Automatic registration of lidar and optical images of urban scenes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2639 – 2646, June 2009.

[12] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit. FastSLAM 2.0: An Improved Particle Filtering Algorithm for Simultaneous Localization and Mapping that Provably Converges. In *International Joint Conference on Artificial Intelligence*, volume 18, pages 1151–1156, 2003.

[13] K. Murphy and S. Russell. Rao-Blackwellised Particle Filtering for Dynamic Bayesian Networks. In A. Doucet, N. de Freitas, and N. Gordon, editors, *Sequential Monte Carlo Methods in Practice*, pages 499–515. Springer-Verlag, New York, NY, 2001.

[14] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26:756–777, June 2004.

[15] G. Nützi, S. Weiss, D. Scaramuzza, and R. Siegwart. Fusion of imu and vision for absolute scale estimation in monocular slam. *Journal of Intelligent and Robotic Systems*, 61:287–299, 2011. 10.1007/s10846-010-9490-z.

[16] D. Scaramuzza, F. Fraundorfer, M. Pollefeys, and R. Siegwart. Absolute scale in structure from motion from a single vehicle mounted camera by exploiting nonholonomic constraints. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1413 –1419, 29 2009-oct. 2 2009.

[17] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. MIT Press, Cambridge, MA, USA, 2005.

[18] 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.