(Preprint) AAS 23-405

LEVERAGING CAMERA ATTITUDE PRIORS FOR STRUCTURE FROM MOTION OF SMALL, NONCOOPERATIVE TARGETS

Kaitlin Dennison[*] and Simone D'Amico[†]

This paper derives and evaluates a formulation of structure from motion (SfM) that incorporates rotation priors, called rotation-informed SfM (RISfM). RISfM uses spacecraft inertial attitude measurements in addition to pixel measurements of the target's features to estimate their 3D points and the translation of the camera. Traditional SfM, which only uses pixel measurements to estimate 3D structure and camera pose, is often considered for low size, weight, power, and cost depth perception and state estimation. However, given a prescribed relative geometry between cameras and specific camera intrinsic parameters, there is a limit to the size of the target, below which SfM breaks down. Even with a wide, inter-image baseline, it is difficult to maintain a safe distance from a small, noncooperative target. In contrast, RISfM has 59.4% less 3D reconstruction error than traditional SfM in orbit configurations where safe separation and keypoint correlation can be achieved. When evaluated on simulated imagery of a 1.48 m width target spacecraft using keypoint descriptors as 2D point measurements, RISfM results in a 3D reconstruction error of 13.25 cm while 15.8 m from the target.

INTRODUCTION

Autonomous spacecraft rendezvous with noncooperative targets is essential for on-orbit servicing and debris removal. Great progress has been made in the field of cooperative target pose estimation where target characteristics and shape are known a priori.[1,2] However, if a model of the target is not available, there are three common pathways to recover its 3D structure: LiDAR, binocular, and monocular systems.[3] Monocular systems are becoming more popular for their low size, weight, power, and cost (SWaP-C) properties and measurement spread.[4,5] State of the art algorithms for monocular-based 3D shape recovery are severely limited by the noise of the input data (pixel measurements or camera pose knowledge) and the scene relative geometry. Here, relative geometry refers to the distance to the target and angular separation between cameras. To overcome these limits, this paper develops a formulation of structure from motion (SfM) that incorporates rotation measurements, which is called rotation-informed SfM (RISfM) throughout this paper.

A previous publication by the authors, Dennison and D'Amico (2023),[6] showed an analysis of traditional SfM and synthetic stereovision (SV) under varying noise and relative geometry cases. It is possible to infer from the analysis that, given a prescribed relative geometry between cameras and specific camera intrinsic parameters, there is a limit to the size of the target, below which SfM performance becomes unacceptable. Even with a wide baseline (translation between camera poses),

[*] Doctoral Candidate, Stanford University Aeronautics & Astronautics Dept., Durand Building, 496 Lomita Mall Stanford, CA 94305.

[†] Associate Professor, Stanford University Aeronautics & Astronautics Dept., Durand Building, 496 Lomita Mall Stanford, CA 94305.

it is difficult to maintain a safe separation from a small, noncooperative target. This presents a challenge because many cases of space rendezvous approach SmallSats. The vast majority of state of the art SfM architectures such as VisualSfM,[7] ORB-SLAM,[8] and COLMAP[9] rely on traditional SfM methods like the one evaluated in Dennison and D'Amico (2023) to initialize their procedure. How do we maximize the distance where SfM is possible given a certain geometry of the problem and number of cameras?

Incorporating a partial camera pose prior (position or rotation) may be the answer. For Earth-orbiting spacecraft, position measurements sometimes available through GNSS or pseudorange and Doppler and are on the order of 1 to 10m, depending on the source and orbit.[10] Attitude measurements are typically acquired onboard spacecraft with the use of star-trackers and have low uncertainties on the order of 60 arcsec.[11] The authors' previous study showed that noisy position measurements propagated significantly more error than noisy rotation measurements did during SV applications because of the scale of the typical position and rotation measurement errors for spacecraft rendezvous.[6]

A position-informed SfM formulation exists where position measurements are incorporated into SfM in Carceroni et al. 2006.[12] Another previous publication by the authors, Dennison, Stacey, and D'Amico (2023),[4] used position and attitude priors for multi-agent SV to initialize an unscented Kalman filter for asteroid rendezvous.

As for SfM that incorporates rotational knowledge, Cui et al. (2017)[13] developed a factor-graph SfM architecture[14] called Hybrid SfM (HSfM) that uses a SfM formulation that assumes rotation is provided. However, Cui focuses on a method for estimating the rotation matrices from correlated images and the overall post-initialization architecture. The SfM formulation is simply a means to an end and no consideration is given for cases where external rotation measurements are provided in addition to the images.

This paper derives RISfM using the same SfM formulation as Cui et al. (2017)[13] but in a context usable for any SfM architecture where rotation is either estimated or measured externally from SfM. The performance of RISfM is compared to traditional SfM formulations with respect to camera relative geometry using simulated imagery of a target spacecraft. It is also evaluated with respect to image and attitude measurement noise as well as three difference keypoint descriptors. RISfM is able to achieve higher precision 3D reconstruction and camera translation estimates from two images at larger distances from the target than traditional SfM techniques.

The remainder of this paper is set up as follows. First, the relevant mathematics of traditional SfM techniques are explained. Then RISfM is derived in the context of inertial rotation measurements. Next, RISfM is evaluated using synthetic images and keypoint descriptors. Finally, the paper concludes and discusses ways forward.

MATHEMATICAL PRELIMINARIES

This section describes the camera projection model as well as traditional SfM techniques that estimate both 3D structure and camera pose. The purpose is to establish the notation, equations, and context for the derivation of RISfM in the following section.

Image projection

The finite projective camera model uses the camera projection matrix $\mathbf{M}$ to transform a 3D point $\boldsymbol{P}$ into a 2D point $\boldsymbol{p}$. It is defined as

$$\boldsymbol{p} = \mathbf{M}\boldsymbol{P}. \tag{1}$$

This uses the homogeneous form of the points such that $\boldsymbol{p} = [u\ v\ 1]^T$ and $\boldsymbol{P} = [xw\ \ yw\ \ zw\ \ w]^T$, where w is a scale factor. $\boldsymbol{P}$ is expressed in the world frame W, which is any 3D orthonormal frame where $\boldsymbol{P}$ is constant,[15] such target surface points oriented in a target body-fixed right-handed triad. The camera projection matrix corresponding to the jth image is defined as

$$\mathbf{M}_j = \mathbf{K}_j \left[\mathbf{R}_W^{C_j} \;\middle|\; {}^{C_j}\boldsymbol{t}_W \right], \tag{2}$$

where $\mathbf{K}_j$ is the camera's calibration matrix. The rotation matrix from W to the jth camera-fixed frame C_j is denoted as $\mathbf{R}_W^{C_j}$. In this work, C is defined such that the z-axis points along the boresight of the camera, the x-axis is parallel to the image frame horizontal axis, and the y-axis completes the right-handed triad (pointing down in the image frame). The vector ${}^{C_j}\boldsymbol{t}_W$ is the 3D vector from camera j's optical center to the origin of W, expressed in the C_j frame.

When working with multiple cameras or views from the same camera, it is useful to define the world frame W to be aligned with a single camera frame, typically the frame of the first image $(W = C_1)$. In other words, the body-fixed axes are assumed to be aligned with the C_1 camera frame without loss of generality. The first and jth projection matrices (see Eq. (2)) become

$$\mathbf{M}_1 = \mathbf{K}_1 \left[\mathbf{I}_3 \mid \mathbf{0}_{3\times1} \right] \quad \text{and} \quad \mathbf{M}_j = \mathbf{K}_j \left[\mathbf{R}_{C_1}^{C_j} \;\middle|\; {}^{C_j}\boldsymbol{t}_{C_1} \right], \tag{3}$$

where $\mathbf{I}_3$ is a 3×3 identity matrix and $\mathbf{0}_{3\times1}$ is a 3×1 vector of zeros. Eq. (3) is called the *canonical form* of a pair of camera matrices.[16] From this point forward, the canonical form will be used where $W = C_1$ and notation will be simplified in the context of two images such that $\mathbf{R} = \mathbf{R}_{C_1}^{C_2}$ and $\boldsymbol{t} = {}^{C_2}\boldsymbol{t}_{C_1}$. Any values of $\boldsymbol{P}$ are in C_1.

It is important to note that because $\boldsymbol{P}$ must be constant in W, if the images are taken at different times, C_1 represents the scene at the moment that image 1 is taken. This means that $\mathbf{R}_{C_1}^{C_j}$ and ${}^{C_j}\boldsymbol{t}_{C_1}$ encodes the target's motion between the time step of image 1 and the time step of image j. The implications of this constraint will be further explored in a later section in the context of RISfM and inertial attitude measurements.

Traditional Structure from Motion

Structure from motion (SfM) is the recovery of the motion of the camera ($\mathbf{R}$ and $\boldsymbol{t}$) between two images and the structure of the scene (the 3D points $\boldsymbol{P}$) purely from 2D measurements $\boldsymbol{p}$ of the points matched between the images. This can be computed when the camera is and is not calibrated. The uncalibrated case will be discussed first as it was the first method developed for SfM.

Uncalibrated Cameras A 3×3, rank 2 matrix, the fundamental matrix $\mathbf{F}$, relates the two image frames. This matrix has seven degrees of freedom and is only defined up to a scale.[16] It is related to the camera matrices via,

$$\mathbf{F} = \mathbf{K}_2^{-T} [\boldsymbol{t}]_\times \mathbf{R} \mathbf{K}_1^{-1} \tag{4}$$

where $[\boldsymbol{t}]_\times$ is the skew-symmetric form of $\boldsymbol{t}$.

Uncalibrated SfM starts with the epipolar constraint, which relates $\mathbf{F}$ to a pair of points ($\boldsymbol{p}_1$ from image 1 and $\boldsymbol{p}_2$ from image 2) matched between two images. It is defined mathematically by

$$\boldsymbol{p}_2^T \mathbf{F} \boldsymbol{p}_1 = 0. \tag{5}$$

Eq. (5) can be rearranged in typical least-squares fashion such that,

$$[u_1 u_2 \quad v_2 u_1 \quad u_1 \quad u_1 v_1 \quad v_1 v_2 \quad v_1 \quad u_2 \quad v_2 \quad 1]\boldsymbol{f} = 0 \tag{6}$$

where $\boldsymbol{f}$ is the vector form of $\mathbf{F}$. Eight or more constraints are required to determine $\boldsymbol{f}$ from noisy point measurements. These can be obtained by appending at least seven more point pairs (for a total of eight point pairs) by vertically stacking paired 2D measurements into a single measurement matrix represented as

$$\mathbf{C}_F = \begin{bmatrix} u_{1,1}u_{2,1} & v_{2,1}u_{1,1} & u_{1,1} & u_{1,1}v_{1,1} & v_{1,1}v_{2,1} & v_{1,1} & u_{2,1} & v_{2,1} & 1 \\ & & & & \vdots & & & & \\ u_{1,N}u_{2,N} & v_{2,N}u_{1,N} & u_{1,N} & u_{1,N}v_{1,N} & v_{1,N}v_{2,N} & v_{1,N} & u_{2,N} & v_{2,N} & 1 \end{bmatrix}. \tag{7}$$

This results in the equation, $\mathbf{C}_F \boldsymbol{f} = \mathbf{0}_{N\times 1}$, which can be solved for $\boldsymbol{f}$ using singular value decomposition (SVD) where $\mathbf{USV}^T = \mathrm{SVD}(\mathbf{C}_F)$ and $\boldsymbol{f}$ is the last column of $\mathbf{V}$. This computation of $\mathbf{F}$ is called the 8-point algorithm. Hartley and Zisserman discuss a more reliable version of this algorithm in chapter 11 of *Multiple View Geometry*[16] called the normalized 8-point algorithm. Specifically, the points should be normalized and the rank constraints of $\mathbf{F}$ must be enforced. Hartley and Zisserman also describe multiple versions of the normalized 8-point algorithm that improve results and robustness such as the polynomial method and the gold standard algorithm.[16]

Calibrated Cameras A similar matrix exists for the calibrated camera case: the essential matrix $\mathbf{E}$. This matrix is also 3×3, rank 2, and of ambiguous scale, but has only five degrees of freedom. It is represented mathematically by

$$\mathbf{E} = [\boldsymbol{t}]_\times \mathbf{R} = \mathbf{K}_2^T \mathbf{F} \mathbf{K}_1. \tag{8}$$

The matrix $\mathbf{E}$ has two constraints. The first is $2\mathbf{E}\mathbf{E}^T\mathbf{E} - tr(\mathbf{E}\mathbf{E}^T)\mathbf{E} = 0$ and the second is $\det(\mathbf{E}) = 0$ because $\mathbf{E}$ is rank 2. Consequently, two of the singular values of $\mathbf{E}$ are one and the third is zero.

$\mathbf{E}$ provides a similar relationship between points as $\mathbf{F}$, but $\mathbf{E}$ operates under the assumption of camera-normalized coordinates where $\hat{\boldsymbol{p}} = \mathbf{K}^{-1}\boldsymbol{p}$. The epipolar constraint (Eq. (5)) becomes

$$\hat{\boldsymbol{p}}_2^T \mathbf{E} \hat{\boldsymbol{p}}_1 = 0 \tag{9}$$

when the essential matrix is used. Eq. (9) can be solved for $\mathbf{E}$ using the same general concept as the 8-point method. Solving for $\mathbf{E}$ is typically called the 5-point method because it only requires five points due to $\mathbf{E}$ having only five degrees of freedom. Nister (2004);[17] Hondong and Hartley (2006);[18] and Kukelova et al. (2009)[19] are examples of solutions to this problem with additional consideration given to constraints and robustness.

Recovering the Camera Matrices

Once $\mathbf{F}$ or $\mathbf{E}$ is estimated, it is possible to determine $\mathbf{R}$ and $\boldsymbol{t}$ using it and the pairs of $\boldsymbol{p}$. While it is possible to retrieve $\mathbf{R}$ and $\boldsymbol{t}$ directly from $\mathbf{F}$, solutions retain a projective ambiguity because $\mathbf{F}$ is not projective invariant.[16] However, $\mathbf{E}$ only has an affine ambiguity and there are four possible solutions based on variants of $\mathbf{R}$ and $\boldsymbol{t}$ and the true solution can be determined from the 2D measurements. It is common to use a calibrated camera, meaning the $\mathbf{K}_j$ matrices are already known. Therefore, even if $\mathbf{F}$ is computed, it is assumed in this work that $\mathbf{E}$ is then computed from $\mathbf{F}$ using Eq. (8). The method for computing $\mathbf{R}$ and $\boldsymbol{t}$ from $\mathbf{E}$ and a set of corresponding 2D points is derived in Ch. 9 of *Multiple View Geometry*[16] and in Nister (2004)[17] but the final, relevant steps are summarized here for later comparison and manipulation.

Four possible camera projection matrices ($\mathbf{M}_2 = \mathbf{K}_2[\mathbf{R} \mid \boldsymbol{t}]$) are comprised of the combinations of two possible $\mathbf{R}$ matrices and two possible $\boldsymbol{t}$ vectors based on SVD$(\mathbf{E}) = \mathbf{U}\text{diag}(1,1,0)\mathbf{V}^T$. The possible values of $\mathbf{R}$ are

$$\mathbf{R} = \mathbf{U}\mathbf{W}\mathbf{V}^T \quad \text{or} \quad \mathbf{U}\mathbf{W}^T\mathbf{V}^T \tag{10}$$

where

$$\mathbf{W} = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{11}$$

The two possible values of $\boldsymbol{t}$ are $\pm\boldsymbol{u}_3$ where $\boldsymbol{u}_3$ is the last column of $\mathbf{U}$. Note that scale is not recovered and $||\boldsymbol{t}|| = 1$. The true $\mathbf{M}_2$ is determined by the chirality constraint: if a point is visible, it must lie in front of the camera.[20] Thus, $\mathbf{M}_1$ from Eq. (3) can be used with each possible $\mathbf{M}_2$ to perform triangulation (see Hartley and Zisserman (2003)[16] or Henry and Christian (2022)[21]) on the set of all paired $\boldsymbol{p}$ and obtain estimates for the respective $\boldsymbol{P}$ values. The true $\mathbf{M}_2$ will have the highest number of $\boldsymbol{P}$ estimates in front of the camera ($\boldsymbol{P}$ will have a positive z-value).

Robust Structure from Motion

The SfM methods discussed so far in this section are not typically used on their own. They are often incorporated as the model of a maximum likelihood estimator sample consensus (MLESAC) algorithm and they are typically part of a larger shape, pose, or state estimation architecture.

Feature detection and tracking is never perfect, meaning there are always outliers and, even when matched correctly, the 2D points may not be exact.[22–24] The SVD method for recovering $\mathbf{F}$ or $\mathbf{E}$ can be erroneous if there are deviations from the ground truth. Therefore, the 8-point and 5-point algorithms are almost always employed as the model for MLESAC with the symmetric epipolar distance or the Sampson distance used as the error metric. While MLESAC is a general outlier rejection algorithm, the authors of the original paper, Torr and Zisserman (2000),[25] explicitly describe its use for estimating image geometry. The symmetric epipolar distance and the Sampson distance are variations of the reprojection error; Fathy et al. (2011)[26] evaluates various error metrics for estimating $\mathbf{F}$.

SfM can be included in a shape, pose, or state estimation architecture that falls under the umbrella of visual simultaneous localization and mapping (visual SLAM). There are two common visual SfM frameworks: the filter-based approach and the factor-graph (or key-frame) approach.[14] In the filter-based approach, SfM is used to initialize part of an extended or unscented Kalman filter's state vector with the camera pose and 3D points, then the 2D points are used as measurement inputs with a variation of Eqs. (1) and (2) as the measurement model.[4,5,27] This can achieve real-time performance at the cost of accuracy when compared to the factor-graph approach.[14] The factor-graph approach can be performed incrementally or in large batches (globally) where the scene is first initialized using SfM on two images. Additional images are incorporated using perspective-n-point solvers on the reconstructed scene and 2D points; then the overall scene is optimized using bundle adjustment.[7–9, 13] This is the more popular approach when computation time or power is less of a concern.[14]

ROTATION-INFORMED STRUCTURE FROM MOTION

The mathematical preliminaries behind basic SfM have been covered and the derivation of RISfM will be performed in this section. RISfM begins similar to the 5-point algorithm but with a modified epipolar constraint. First, Eqs. (8) and (9) are combined as

$$\hat{\boldsymbol{p}}_2^T [\boldsymbol{t}]_\times \mathbf{R} \hat{\boldsymbol{p}}_1 = 0. \tag{12}$$

Next, expand Eq. (12) and collect the coefficients of the components of $\boldsymbol{t}$ such that $[C_x \; C_y \; C_z]\boldsymbol{t} = 0$. These coefficients are

$$C_x = \mathbf{R}_{21}\hat{u}_2 + \mathbf{R}_{22}\hat{v}_2 + \mathbf{R}_{23} - \mathbf{R}_{31}\hat{v}_1\hat{u}_2 - \mathbf{R}_{32}\hat{v}_1\hat{v}_2 - \mathbf{R}_{33}\hat{v}_1 \tag{13}$$

$$C_y = -\mathbf{R}_{11}\hat{u}_2 - \mathbf{R}_{12}\hat{v}_2 - \mathbf{R}_{13} + \mathbf{R}_{31}\hat{u}_1\hat{u}_2 + \mathbf{R}_{32}\hat{u}_1\hat{v}_2 + \mathbf{R}_{33}\hat{u}_1 \tag{14}$$

$$C_z = \mathbf{R}_{11}\hat{v}_1\hat{u}_2 + \mathbf{R}_{12}\hat{v}_1\hat{v}_2 + \mathbf{R}_{13}\hat{v}_1 - \mathbf{R}_{21}\hat{u}_1\hat{u}_2 - \mathbf{R}_{22}\hat{u}_1\hat{v}_2 - \mathbf{R}_{23}\hat{u}_1 \tag{15}$$

where the subscript indices ab of $\mathbf{R}_{ab}$ denote the corresponding entry of $\mathbf{R}$ and the camera-normalized point from image j is represented as $\hat{\boldsymbol{p}}_j = [\hat{u}_j \; \hat{v}_j \; 1]^T$.

Finally, vertically stack the coefficients ($[C_{x,i} \; C_{y,i} \; C_{z,i}] \; \forall i \in [1, N]$) formed by each pair of correlated 2D measurements into one measurement matrix

$$\mathbf{C}_R = \begin{bmatrix} C_{x,1} & C_{y,1} & C_{z,1} \\ & \vdots & \\ C_{x,N} & C_{y,N} & C_{z,N} \end{bmatrix}. \tag{16}$$

The linear equation $\mathbf{C}_R\boldsymbol{t} = \mathbf{0}_{N\times 1}$ can be solved for $\boldsymbol{t}$ using SVD, like in the 8-point and 5-point algorithms. This also enforces the constraint that $||\boldsymbol{t}|| = 1$, meaning $\boldsymbol{t}$ only has two degrees of freedom and neither scale nor sign is recovered. As a result of $\boldsymbol{t}$ having 2 degrees of freedom, only two pairs of image points are necessary. Once $\boldsymbol{t}$ is computed, the essential matrix is easily computed using Eq. (8).

Now the camera projection matrix can be recovered, but RISfM allows the computation cost of the $\mathbf{R}$ and $\boldsymbol{t}$ recovery process described previously to be significantly reduced by two aspects. First, because $\mathbf{E}$ is already decomposed into $\boldsymbol{t}$ and $\mathbf{R}$, SVD does not need to be used to obtain them from $\mathbf{E}$. This reduces the linear algebra overhead per iteration. Second, knowing $\mathbf{R}$ means two of the four possible affine transformations can be eliminated. This cuts the number of triangulation computations in half. Therefore, $\mathbf{M}_2 = \mathbf{K}_2[\mathbf{R} \mid \boldsymbol{t}]$ or $\mathbf{K}_2[\mathbf{R} \mid -\boldsymbol{t}]$ and the true $\mathbf{M}_2$ is found by checking the chirality constraint of both as described in the previous section.

RISfM can replace the 8-point or 5-point algorithm in any factor-graph architecture and is, in fact, what HSfM does.[13] However, HSfM also estimates the rotation of the camera and has not been evaluated for when rotation measurements have been provided instead. Additionally, RISfM can be used to initialize a navigation filter for target shape and pose estimation purposes.[5, 14, 28] Both of these use cases are outside the scope of this paper but would be useful avenues for future work.

Inertial Attitude Measurements

It is important to discuss when RISfM is possible to use based on the rotation measurements provided. Spacecraft attitude measurements are typically taken in the inertial frame I so the rotation measurements available are $\mathbf{R}_I^{C_j}$. It is imperative that $\boldsymbol{P}$ is constant in W or triangulation (which is the final step of SfM and RISfM) will not work.[15] The points on the surface of the target $\boldsymbol{P}$ move in the inertial frame with the rotation of the target. Therefore, if $W = C_1$, the rotation of the points in the inertial frame between images must be accounted for when using inertial attitude measurements. To do so, the rotation used in Eq. (3) for SfM and RISfM is defined as

$$\mathbf{R}_{C_1}^{C_j} = \mathbf{R}_I^{C_j} \mathbf{R}_B^I(j) \mathbf{R}_I^B(1) \mathbf{R}_{C_1}^I, \tag{17}$$

where $\mathbf{R}_B^I(j)$ is the rotation from a target body-fixed frame B to the inertial frame I at the time when image j is taken. Baldini et al. (2018)[29] provides an in-depth derivation of Eq. (17).

Given $\mathbf{R}_I^{C_j}$ measurements, there are three cases where $\mathbf{R}_{C_1}^{C_j}$ can be calculated for use in RISfM. First, images are taken at arbitrary time steps and the target is not moving in the inertial frame. Second, there is a single camera taking sequential images of a moving target but $\mathbf{R}_B^I(j)$ is known or easily estimated (e.g. a cooperative target[1,30]). Third, there are two or more cameras taking images simultaneously such that $\mathbf{R}_B^I(j) = \mathbf{R}_B^I(1)$. The third case is the only viable option for spacecraft rendezvous with an unknown, noncooperative target, which necessitates the use of a distributed space system or binocular camera.

PERFORMANCE EVALUATION

RISfM is evaluated in three Monte Carlo simulations. First is a comparison between the normalized 8-point algorithm, the 5-point algorithm, and RISfM with respect to relative geometry. Second tests the robustness of RISfM with respect to pixel noise and attitude jitter. Third assesses the performance of RISfM with respect to relative geometry when three different keypoint descriptors are used on synthetic images.

Simulation Setup

All three simulations use synthetic images of a 1.48m width spacecraft, Tango from the PRISMA mission,[30] from two cameras. The cameras are both modeled as a Grasshopper3[31] with $f = 17.5$mm, which is the same camera used in Dennison and D'Amico (2023)[6] and Park et al. (2022).[32] The relative geometry used for all simulations is illustrated in Figure 1. Two cameras are pointed directly at Tango's center of mass, have an angular separation β between each other, and are the same distance d from Tango. The canonical camera pose components $\mathbf{R}$ and $\boldsymbol{t}$ are also illustrated. The inertial frame is represented so show that Tango and the two cameras are fixed in I at the moment the two images are taken. Without loss of generality, I can be positioned and oriented anywhere in the scene with respect to C_1 and C_2.

A noisy inertial attitude measurement $\tilde{\mathbf{R}}_I^{C_j}$ for each camera is assumed known. To represent these, jitter is applied to the true attitude of each camera using a 3-2-1 rotation,[11]

$$\tilde{\mathbf{R}}_I^{C_j} = \mathbf{R}_1\left(\mathcal{N}\left(0, (0.5\sigma_J)^2\right)\right) \mathbf{R}_2\left(\mathcal{N}\left(0, (0.5\sigma_J)^2\right)\right) \mathbf{R}_3\left(\mathcal{N}\left(0, \sigma_J^2\right)\right) \mathbf{R}_I^{C_j}, \tag{18}$$

where $\mathbf{R}_1$, $\mathbf{R}_2$, and $\mathbf{R}_3$, are x-, y-, and z-axis rotations, respectively. $\mathcal{N}\left(0, \sigma^2\right)$ represents a sample from a normal distribution with a mean of zero and standard deviation of σ. Unless otherwise

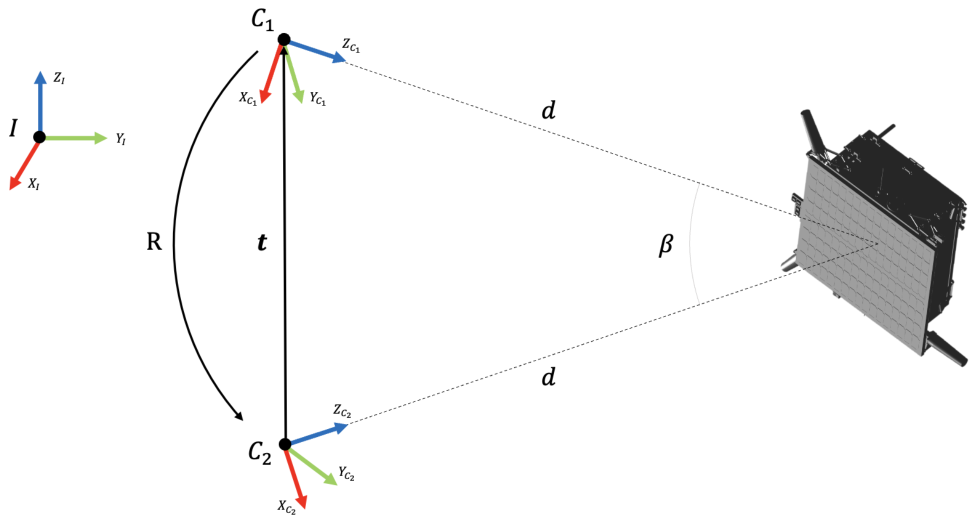

Figure 1: The relative geometry between two synchronized cameras and the target, Tango.

stated, $\sigma_J = 120$ arcsec. All simulations assume synchronized cameras so $\mathbf{R}_B^I(1) = \mathbf{R}_B^I(2)$ and $\mathbf{R} = \tilde{\mathbf{R}}_I^{C_2}\tilde{\mathbf{R}}_{C_1}^I$ from Eq. (17). This is the same relative geometry setup used in Dennison and D'Amico (2023)[6] and the reader is referred to that paper for further details.

In each simulation, two parameters are varied and the results are averaged over twenty samples for each parameter pair where Tango's attitude $(\mathbf{R}_B^I)$ and the lighting direction are randomized. The first and third simulation vary β and d while the second simulation varies the pixel noise (σ_p) and σ_J. Furthermore, the 3D points are randomly sampled from the target's 3D model in the first two simulations. Pixel measurements are modeled as 2D projections of the points with $\sigma_p = 2$px standard deviation Gaussian white noise added. In the third simulation, the pixel measurements are centers of the keypoints detected in synthetic images and the ground truth 3D points are the centers ray-traced through Tango's 3D model.[22] The simulation setup is summarized in Table 1.

Table 1: Variables and measurement sources for the three simulations.

Simulation	Variables	3D Points $\boldsymbol{P}$	2D Measurements $\boldsymbol{p}$
1	β and d	Model Samples	Projected Points
2	σ_J and σ_p	Model Samples	Projected Points
3	β and d	Ray-Traced Centers	Keypoint Centers

Each synthetic image is corrupted with Gaussian white noise with a standard deviation of 0.0022, Gaussian blur with standard deviation of 0.8, and salt and pepper noise with a standard deviation of 0.001. These parameters are chosen as representative noise values from various studies.[33,34] A median filter and a Wiener filter are applied to each image to de-noise and aid keypoint detection and matching. Keypoints are matched between images using their feature descriptors only. Example images before de-noising are shown in Figure 2; note that the white pixels in the background are salt and pepper noise, not stars.

Results and Discussion

Figure 2 shows a RISfM sample case where $\beta = 20°$, $d = 15.8$ m, and $\sigma_J = 120$ arcsec. The zoomed-in, synthetic images from both cameras are displayed in the top row with the detected and matched ORB points overlaid. The bottom left corner shows Tango's 3D model along with the ray-traced points (green circles) and RISfM-estimated 3D points (red crosses) multiplied by their true scale. Finally, the bottom right corner shows the entire scene to provide scale perspective in the X-Z plane of the C_1 frame. Camera 1 is indicated in blue and is, by definition of C_1, placed at the origin of C_1 with its boresight aligned with the Z-axis. Camera 2's true pose is indicated in black while the RISfM-estimated pose is indicated in red. In this example, the root mean square error (RMSE) in 3D reconstruction is $\delta P = 13.25$ cm and translation error is $\delta t = 15.59$ cm.

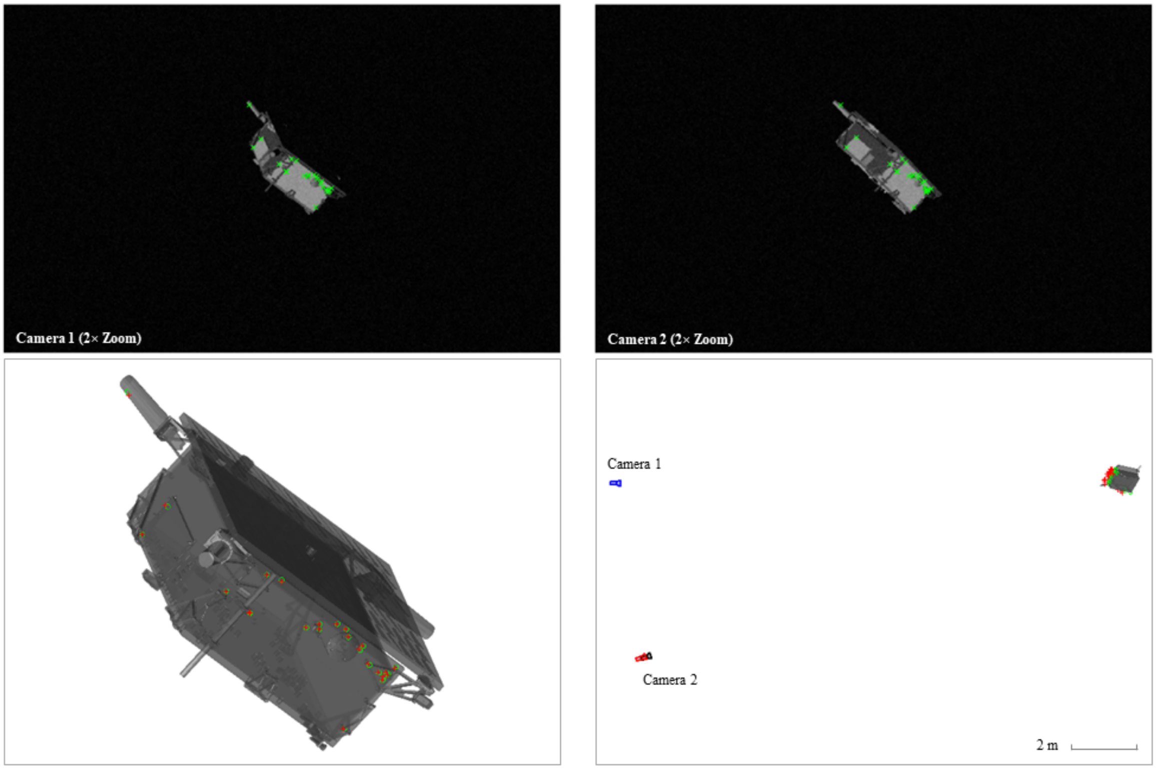

Figure 2: An example of RISfM estimation using ORB descriptors on synthetic images.

Simulation 1. Figure 3 shows the performance of the two traditional SfM methods compared to RISfM for varying relative geometry parameters: angular separation between cameras and distance from the target. All three methods are implemented in a MLESAC architecture with Sampson distance as the error metric. For each angular separation and target distance pair, the color represents δP between the true 3D points and the estimated 3D points (multiplied by their true scale) of all points estimated across 20 samples with Tango given a random orientation and lighting condition.

For the entire parameter space, the RISfM error is, on average, 58.3% and 7.06% less than the error of the normalized 8-point algorithm and 5-point algorithm, respectively. Furthermore, the respective median of δP is 1.50 m, 1.95 m, and 0.026 m for the normalized 8-point, 5-point, and RISfM algorithms. Thus, RISfM is able to improve 3D reconstruction performance in general for targets on the scale of 1.48 m in width that are observed with two synchronized, wide field of

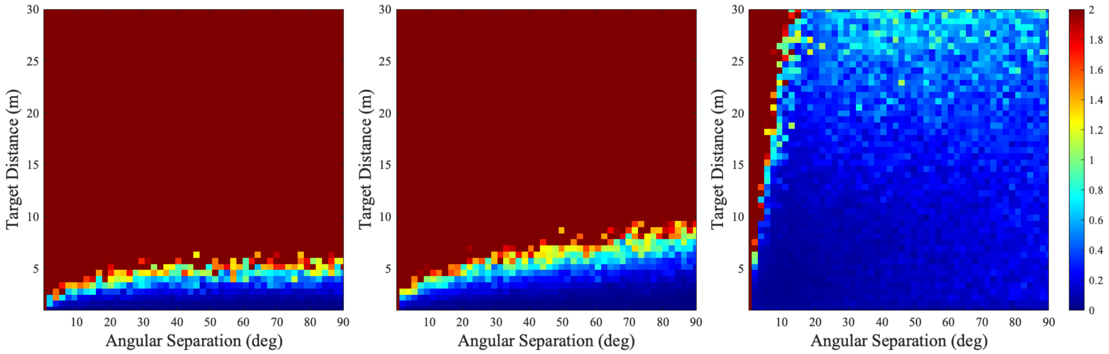

Figure 3: A comparison between the normalized 8-point algorithm (left), the 5-point algorithm (middle), and RISfM (right) with respect to relative geometry

view cameras that have associated attitude measurements. One interesting behavior is that RISfM performance decreases as angular separation increases unlike traditional SfM, meaning traditional SfM performs better than RISfM when the angular separation is large but the distance to the target is small.

The region where $\beta \leq 40°$ and $d \geq 8$ m is of particular interest because it represents the angular separation limit for keypoint tracking[22] and minimum distance for safe observation.[32] In this region, RISfM error is, on average, 77.4% and 59.4% less than the error of the normalized 8-point algorithm and 5-point algorithm, respectively. Thus, RISfM expands the limits on relative geometry beyond that of traditional SfM techniques.

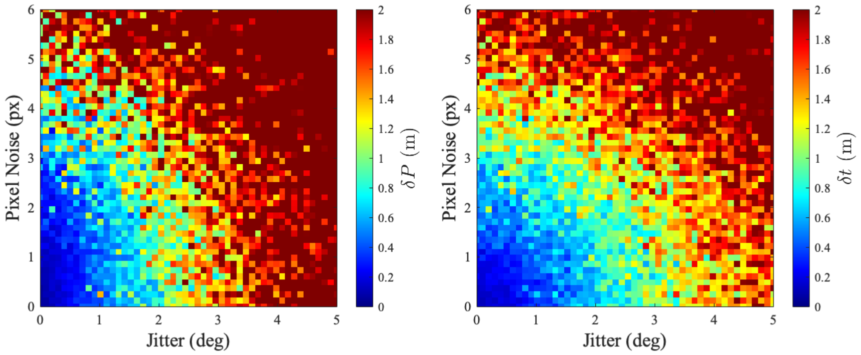

Figure 4: Performance of RISfM with respect to measurement noise and jitter. Shown is the RMSE of the 3D reconstruction (left) and the translation (right) estimates.

Simulation 2. Figure 4 shows the RMSE in 3D point and translation estimation for a Monte Carlo simulation varying attitude jitter and pixel noise when $\beta = 30°$ and $d = 15$ m. Nominal attitude jitter for a star tracker is around 60 arcsec[11] and pixel noise for key point descriptors is around two to five pixels.[4] These results show that RISfM is highly robust to both pixel noise and attitude jitter as long as the attitude priors used have an error less than approximately 1°.

Simulation 3. In Figure 5, RISfM is evaluated using three different keypoint descriptors (ORB,[35] SIFT,[36] and SURF[37]) instead of projections of points sampled from the model. For the entire parameter space, the median of δP is 19.47 cm, 13.21 cm, and 16.15 cm for ORB, SIFT, and

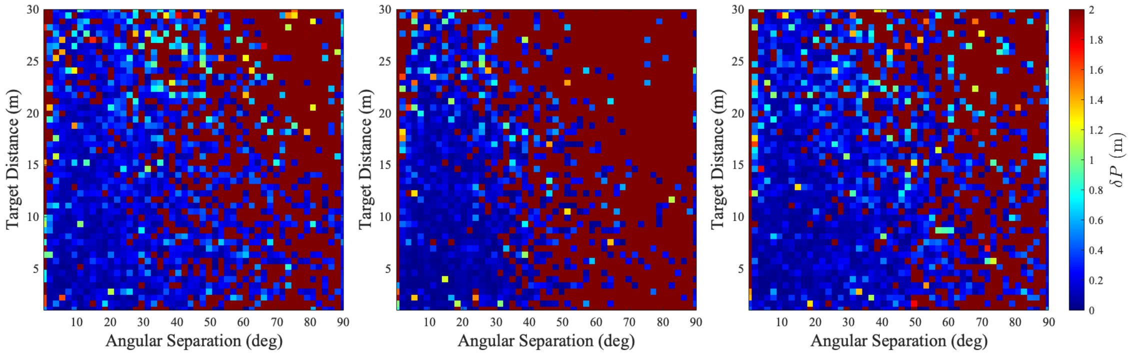

Figure 5: Performance of RISfM using keypoint descriptors on synthetic images of Tango with respect to relative geometry parameters. The individual plots use ORB (left), SIFT (middle), and SURF (right).

SURF, respectively. Usually space missions will include state estimation-based outlier rejection in MLESAC so the actual performance is expected to be better than the results shown.[4,25]

As the angular separation increases, matching performance decreases because the feature descriptors are not projective invariant. This concurs with previous keypoint descriptor analyses.[22,24] Interestingly, the keypoint descriptors do not appear to have as high of a minimum separation as the 2D projections do in Figure 3. This is likely because the 2D projections always had 2 px standard deviation of noise applied while the keypoints detected (and thus, the centroid error) change with the distance to the target. At 1° separation and 30 m to the target, 2 px of error can easily lead to the image projection rays crossing where they should not.

CONCLUSION

In an effort to overcome the limitations of traditional structure from motion (SfM) techniques, this paper introduced and evaluated a formulation of SfM that incorporates camera rotation knowledge, called rotation-informed structure from motion (RISfM). Traditional SfM techniques must be within a short distance from their target, proportional to the target's size and angular separation between images. This makes SfM difficult to perform for small, noncooperative targets while maintaining a safe operating distance and minimal angular separation for keypoint tracking.

The performance of RISfM was evaluated with respect to relative geometry parameters (distance to the target and angular separation between cameras) and noise parameters (pixel measurement and attitude jitter). It was compared to the performance of the two most common SfM techniques: the normalized 8-point algorithm and the 5-point algorithm. RISfM was shown to have more accurate 3D reconstruction performance than traditional SfM techniques when using two synchronized, wide field of view cameras that each have an associated inertial attitude measurement from a star tracker. When angular separation is less than 40° and the distance to a 1.48 m target is greater than 8 m, RISfM 3D reconstruction performance was 59.4% less on average than the best performing traditional SfM technique (the 5-point algorithm). When RISfM was evaluated using keypoint descriptors on synthetic images of the target spacecraft, the median 3D reconstruction error for ORB descriptors was 19.47 cm.

In future research, it would be useful to compare RISfM performance to multi-agent or sequential stereovision as well as a binocular camera system. It is possible that a binocular camera may have better performance under some conditions because the cameras are rectified and their poses are known to the tolerance of manufacturing. Although, the baseline between cameras is limited by the spacecraft that the binocular camera is mounted on, which limits the performance of stereovision in the presence of noise. Furthermore, additional research will be needed to investigate the robustness of RISfM for initializing a state estimation architecture, especially in terms of image synchronization and the accuracy of attitude priors. However, the simulations in this paper show that RISfM is robust with respect to star-tracker noise tolerances and keypoint descriptors, making it promising for initializing spacecraft navigation filters.

ACKNOWLEDGMENTS

This research has been supported by SpaceWERX Orbital Prime Small Business Technology Transfer (STTR) contract number FA8649-23-P-0560 as awarded to Ten One Aerospace in collaboration with the Stanford Space Rendezvous Laboratory.

REFERENCES

[1] T. H. Park, S. Sharma, and S. D'Amico, "Towards Robust Learning-Based Pose Estimation of Noncooperative Spacecraft," *2nd RPI Space Imaging Workshop*, 2019, pp. 1–2.

[2] S. Sharma and S. D'Amico, "Neural Network-Based Pose Estimation for Noncooperative Spacecraft Rendezvous," *IEEE Transactions on Aerospace and Electronic Systems*, 2020, pp. 1–1, 10.1109/TAES.2020.2999148.

[3] R. Opromolla, G. Fasano, G. Rufino, and M. Grassi, "A review of cooperative and uncooperative spacecraft pose determination techniques for close-proximity operations," *Progress in Aerospace Sciences*, Vol. 93, 2017, pp. 53–72. Publisher: Elsevier.

[4] K. Dennison, N. Stacey, and S. D'Amico, "Autonomous Asteroid Characterization Through Nanosatellite Swarming," *IEEE Transactions on Aerospace and Electronic Systems*, 2023, pp. 1–22, 10.1109/TAES.2023.3245997.

[5] V. Pesce, M. Lavagna, and R. Bevilacqua, "Stereovision-based pose and inertia estimation of unknown and uncooperative space objects," *Advances in Space Research*, Vol. 59, Jan. 2017, pp. 236–251, 10.1016/j.asr.2016.10.002.

[6] K. Dennison and S. D'Amico, "Vision-Based 3d Reconstruction for Navigation and Characterization of Unknown, Space-borne Targets," Austin, TX, Jan. 2023.

[7] C. Wu, "Towards Linear-Time Incremental Structure from Motion," *2013 International Conference on 3D Vision - 3DV 2013*, June 2013, pp. 127–134. ISSN: 1550-6185, 10.1109/3DV.2013.25.

[8] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM," *IEEE Transactions on Robotics*, Vol. 37, Dec. 2021, pp. 1874–1890. Conference Name: IEEE Transactions on Robotics, 10.1109/TRO.2021.3075644.

[9] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[10] M. D'Errico, *Distributed Space Missions for Earth System Monitoring*. Springer Science & Business Media, Sept. 2012.

[11] C. Pong, "On-Orbit Performance & Operation of the Attitude & Pointing Control Subsystems on ASTERIA," *32nd Annual Small Satellite Conference*, Utah, USA, Aug. 2018, p. 20.

[12] R. Carceroni, A. Kumar, and K. Daniilidis, "Structure from Motion with Known Camera Positions," *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1 (CVPR'06)*, Vol. 1, New York, NY, USA, IEEE, 2006, pp. 477–484, 10.1109/CVPR.2006.296.

[13] H. Cui, X. Gao, S. Shen, and Z. Hu, "HSfM: Hybrid Structure-from-Motion," 2017, pp. 1212–1221.

[14] D. Zou, P. Tan, and W. Yu, "Collaborative visual SLAM for multiple agents:A brief survey," *Virtual Reality & Intelligent Hardware*, Vol. 1, Oct. 2019, pp. 461–482, 10.1016/j.vrih.2019.09.002.

[15] M. R. U. Saputra, A. Markham, and N. Trigoni, "Visual SLAM and Structure from Motion in Dynamic Environments: A Survey," *ACM Computing Surveys*, Vol. 51, Mar. 2019, pp. 1–36, 10.1145/3177853.

[16] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge: Cambridge University Press, 2003.

[17] D. Nister, “An efficient solution to the five-point relative pose problem,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 26, June 2004, pp. 756–770, 10.1109/TPAMI.2004.17.

[18] Hongdong Li and R. Hartley, “Five-Point Motion Estimation Made Easy,” *18th International Conference on Pattern Recognition (ICPR’06)*, Hong Kong, China, IEEE, 2006, pp. 630–633, 10.1109/ICPR.2006.579.

[19] Z. Kukelova, M. Bujnak, and T. Pajdla, “Polynomial Eigenvalue Solutions to the 5-pt and 6-pt Relative Pose Problems,” *Procedings of the British Machine Vision Conference 2008*, Leeds, British Machine Vision Association, 2008, pp. 56.1–56.10, 10.5244/C.22.56.

[20] S. Agarwal, A. Pryhuber, R. Sinn, and R. R. Thomas, “The Chiral Domain of a Camera Arrangement,” *Journal of Mathematical Imaging and Vision*, Vol. 64, Nov. 2022, pp. 948–967, 10.1007/s10851-022-01101-2.

[21] S. Henry and J. A. Christian, “Absolute Triangulation Algorithms for Space Exploration,” *Journal of Guidance, Control, and Dynamics*, Vol. 46, No. 1, 2023, pp. 21–46. Publisher: American Institute of Aeronautics and Astronautics _eprint: https://doi.org/10.2514/1.G006989, 10.2514/1.G006989.

[22] K. Dennison and S. D’Amico, “Comparing Optical Tracking Techniques in Distributed Asteroid Orbiter Missions Using Ray-Tracing,” Charlotte, NC, Feb. 2021. Virtual Event.

[23] N. Takeishi, A. Tanimoto, T. Yairi, Y. Tsuda, F. Terui, N. Ogawa, and Y. Mimasu, “Evaluation of Interest-region Detectors and Descriptors for Automatic Landmark Tracking on Asteroids,” *Transactions of the Japan Society for Aeronautical and Space Sciences*, Vol. 58, No. 1, 2015, pp. 45–53, 10.2322/tjsass.58.45.

[24] L. P. Cassinis, R. Fonod, E. Gill, I. Ahrns, and J. G. Fernandez, “Comparative Assessment of Image Processing Algorithms for the Pose Estimation of Uncooperative Spacecraft,” Glasgow, UK, July 2019, p. 21.

[25] P. H. S. Torr and A. Zisserman, “MLESAC: A New Robust Estimator with Application to Estimating Image Geometry,” *Computer Vision and Image Understanding*, Vol. 78, Apr. 2000, pp. 138–156, 10.1006/cviu.1999.0832.

[26] M. E. Fathy, A. S. Hussein, and M. F. Tolba, “Fundamental Matrix Estimation: A Study of Error Criteria,” *Pattern Recognition Letters*, Vol. 32, Jan. 2011, pp. 383–391. arXiv:1706.07886 [cs], 10.1016/j.patrec.2010.09.019.

[27] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, “MonoSLAM: Real-Time Single Camera SLAM,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, June 2007, pp. 1052–1067, 10.1109/TPAMI.2007.1049.

[28] V. Capuano, “Monocular-based pose determination of uncooperative space objects,” *Acta Astronautica*, Vol. 166, Jan. 2020, pp. 493–506, 10.1016/j.actaastro.2019.09.027.

[29] F. Baldini, A. Harvard, S.-J. Chung, I. Nesnas, and S. Bhaskaran, “Autonomous Small Body Mapping and Spacecraft Navigation Via Real-Time SPC-SLAM,” Bremen, Germany, International Astronautical Federation, Oct. 2018, p. Art. No. 47373.

[30] S. D’Amico, J.-S. Ardaens, and R. Larsson, “Spaceborne autonomous formation-flying experiment on the PRISMA mission,” *Journal of Guidance, Control, and Dynamics*, Vol. 35, No. 3, 2012, pp. 834–850.

[31] FLIR, “Grasshopper3 2.3 MP Mono GigE Vision,”

[32] T. H. Park and S. D’Amico, “Adaptive Neural Network-based Unscented Kalman Filter for Robust Pose Tracking of Noncooperative Spacecraft,” Nov. 2022. arXiv:2206.03796 [cs, eess], 10.48550/arXiv.2206.03796.

[33] R. C. Gonzalez and R. E. Woods, *Digital image processing*. Pearson, 2018.

[34] P. Patidar, M. Gupta, S. Srivastava, and A. K. Nagawat, “Image De-noising by Various Filters for Different Noise,” *International Journal of Computer Applications*, Vol. 9, Nov. 2010, pp. 45–50, 10.5120/1370-1846.

[35] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “ORB: An efficient alternative to SIFT or SURF,” *2011 International Conference on Computer Vision*, Barcelona, Spain, Nov. 2011, pp. 2564–2571, 10.1109/ICCV.2011.6126544.

[36] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” *International Journal of Computer Vision*, Vol. 60, Nov. 2004, pp. 91–110, 10.1023/B:VISI.0000029664.99615.94.

[37] H. Bay, T. Tuytelaars, and L. Van Gool, “SURF: Speeded Up Robust Features,” *Computer Vision – ECCV 2006*, Vol. 3951, Berlin, Heidelberg, Springer Berlin Heidelberg, 2006, pp. 404–417, 10.1007/1174402332.