Rover Localization for Tube Pickup: Dataset, Methods and Validation for Mars Sample Return Planning

Tu-Hoa Pham, William Seto, Shreyansh Daftry, Alexander Brinkman, John Mayo, Yang Cheng, Curtis Padgett, Eric Kulczycki and Renaud Detry
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA
tu-hoa.pham@jpl.nasa.gov

***Abstract*—The Mars 2020 rover mission is intended to collect samples which will be stored in metal tubes and left on the surface of Mars, for possible retrieval and return to Earth by a future mission. In the proposed Mars Sample Return (MSR) campaign concept, a follow-up mission would collect the sample tubes and load them into a Mars Ascent Vehicle to be launched into orbit for subsequent transfer and return to Earth. In this work, we study the problem of autonomous tube localization and pickup by a "Fetch" rover during the MSR campaign. This is a challenging problem as, over time, the sample tubes may become partially covered by dust and sand, thereby making it difficult to recover their pose by direct visual observation. We propose an indirect approach, in which the Fetch rover localizes itself relative to a map built from Mars 2020 images. The map encodes the position of rocks that are sufficiently tall not to be affected by sand drifts. Because we are confident that tubes will remain immobile until Fetch arrives, their pose within the Mars 2020 map can be used to plan pickup maneuvers without directly observing the tubes in Fetch images. To support this approach, we present a dataset composed of 4160 images collected from two sets of stereo cameras placed at thirteen different view angles, two different heights from the ground, two distances from a tube, in five different lighting conditions, and ground-truthed with a motion capture setup. This dataset allows us to quantify the sensitivity of terrain-relative tube localization with respect to lighting conditions and camera pose.**

TABLE OF CONTENTS

1. INTRODUCTION

The Mars 2020 rover mission is intended to collect samples which will be stored in metal tubes and left on the surface of Mars, for possible retrieval and return to Earth by a future mission. In the proposed Mars Sample Return (MSR) campaign concept, a follow-up mission would collect the sample tubes and load them into a Mars Ascent Vehicle to be launched into orbit for subsequent transfer and return to Earth. In this work, we study the problem of autonomous tube localization and pickup by a "Fetch" rover during the MSR campaign. This is a challenging problem as, over time, the sample tubes may become partially covered by dust and sand, thereby making it difficult to recover their pose by direct visual observation. We thus propose an indirect approach, in which the rover localizes itself with respect to landmarks – tall rocks, mainly – that will not be affected by sand drifts, and that have previously been mapped by Mars 2020. Matching the map to the terrain allows us to pick up tubes without directly observing them. In this paper, we assume that the map is simply a collection of stereo images wherein the pose of the tube is manually encoded. The problem of localizing the Fetch rover with respect to the map simplifies into computing the relative pose of two stereo pairs: one belonging to the Fetch rover, the other being the nearest pair captured by Mars 2020 (see Section 2).

The main contribution of this paper is a novel dataset designed to benchmark the performance of terrain-relative rover localization on Mars. The dataset comprises over 4000 images of terrain that resembles the Martian surface, annotated with ground-truth camera poses obtained through a motion capture system (MoCap). The dataset also features a sample tube prototype set at the center of the setup and includes motion-capture ground truth pose for the tube, to support follow-on studies on direct tube localization (see Section 3). The dataset was collected to span multiple parameters representative of different capture configurations, i.e., different Fetch poses relative to the nearest Mars 2020 image-capture poses, namely, for two different rock densities, camera type, exposure, stereo baseline, viewpoint angle, distance, height, and lighting direction (see Section 4).

The second contribution of this paper is a benchmark of a sparse-feature localizer using this dataset (see Section 6). We estimate the camera motion between two stereo pairs captured under different capture configurations by four-way feature matching and reprojection error minimization between the resulting correspondences. We perform this calculation for all possible pairs of capture configurations. By comparing the transformation estimates from our localizer to ground-truth transformations from MoCap, we obtain 270,920 pose estimation errors expressed as translation and rotation errors for the tube across 6 capture dimensions. We use these estimates to identify and characterize parameter configurations enabling robust localization, which in turn can be used to make feasibility assessments and recommendations for the Mars 2020 and Fetch rover missions (see Section 7). We further examine the vulnerability of feature matching to variations in lighting conditions and discuss alternative localization methods using synthetic relighting and rendering (see Section 8).

978-1-7281-2734-7/20/$31.00 ©2020 IEEE

Surface features emerging from sand patterns may not be reliable in the current notional mission context, because sand may be affected by weather. To capture this risk, we restrict our benchmark to using features that emerge from rocks. We automate the pruning of features that emerge from sand by segmenting all images with a neural network that we specifically trained to discriminate between rocks and sand (see Section 5).

2. Mission Concept and Problem Statement

The notional MSR campaign [1], [2] is a four-phase campaign that includes three launches, two rovers, one lander, and a Mars rocket. The first phase is the collection of rock samples with the Mars 2020 rover, and the release of those tubes on the ground, at one or multiple sites referred to as *tube depots*. Within each depot, tubes would be released 2 to 5 meters apart, along a straight or curved line. Mars 2020's control over the layout is limited by the release mechanism, which consists of dropping the tubes from the rover's belly onto the ground.

The next phase would be the recovery of the tubes and launch to Martian orbit. A lander set to reach the Martian surface in 2027 or 2028 would carry a MER-class tube-fetching rover and a Mars Ascent Vehicle (MAV). The fetching rover (Fetch) would collect the tubes and bring them back to the lander, where all the tubes would be transferred to a single container, that the MAV would launch to orbit. The launch would coincide with the arrival of the third phase: a probe designed to catch the container and bring it back to Earth, with a ballistic landing at a site to be determined. The fourth phase would be the Earth-based sample containment and analysis.

Traditionally, Mars rovers that are intended to survive the Martian winter have used radioisotope heating, via radioisotope heating units (RHU, Mars Exploration Rovers) or a radioisotope thermoelectric generator (RTG, Mars Science Laboratory). To limit cost, the baseline Fetch rover would have neither, and therefore must complete its mission in a single season, including a $10\,\mathrm{km}$ drive to one or multiple depots, picking up tubes, and driving back. As a result, surface mission planning requires that the rover must notionally pick up 30 tubes in 20 sols. The minimum time for picking up a tube with ground in the loop being 3 sols [2], Fetch would have to recover tubes autonomously. Because of the time constraint, Fetch would also need to be able to recover tubes independently of time-of-sol (lighting conditions), from an arbitrary approach vector, and be robust to the effect Martian weather may have on tubes and the ground. The localization software must run on a RAD750 processor accompanied by an FPGA comparable to the Virtex 5.

Tubes will not be released in areas where sand is abundant, or on slopes steeper than $25°$. Based on those constraints and our understanding of Mars weather, geology, and qualitative observation of lander and rover images, we conservatively hypothesize that: (a) tubes will not move, (b) sand or dust may pile up next to tube or rocks, forming drifts, conceivably burying a tube entirely, (c) sand drifts will not exceed $5\,\mathrm{cm}$ in height, leaving the upper part of rocks taller than $5\,\mathrm{cm}$ unaffected, and (d) dust will deposit everywhere, creating a dust layer that will not exceed $0.25\,\mathrm{mm}$ in thickness.

To maximize the probably of successfully recovering all tubes, we plan to implement two localization solutions characterized by orthogonal constraints. The first is the terrain-relative localizer discussed in this paper. It is robust to the accumulation of sand near a tube, but relies on the presence of landmarks that are not be affected by drifts, and requires that Fetch approaches each tube along a vector that is close to the viewpoint of an image captured by Mars 2020. The second solution is to directly segment tube pixels in Fetch camera images. It will only work if tubes are unaffected by drifts, but it does not constrain the approach vector, and does not require the presence of landmarks near the tube. As mentioned above, this paper focuses on the terrain-relative solution.

3. Platform and Experimental Setup

In this section we describe the setup to collect images mimicking the Fetch rover's navigation and hazard cameras.

Camera Setup

To validate and assess the accuracy of our machine vision techniques, we collected a large dataset of tube depot images taken in a large variety of viewpoints and acquisition conditions. Mars rovers, including MER, MSL, Mars 2020, and the Fetch concept, usually features multiple sets of cameras. Those that are relevant for tube pickup include the front-facing hazard stereo camera, typically situated at about $70\,\mathrm{cm}$ off the ground [3], and the navigation stereo camera, attached to a mast placing them at 1.5 to $2\,\mathrm{m}$ off the ground. To conveniently collect our dataset, we constructed a dedicated camera setup that can be carried, positioned and oriented by hand. The camera setup consisted of four cameras:

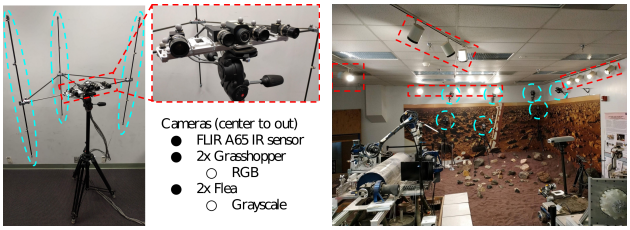
- Two PointGrey Grasshopper GRAS-50S5C-C cameras (2448×2048, color), stereo baseline $\approx 10\,\mathrm{cm}$, $27°$ field of view
- Two PointGrey Flea2 FL2G-50S5M cameras (2448×2048, grayscale), stereo baseline $\approx 30\,\mathrm{cm}$, $37°$ field of view

We rigidly mounted the four cameras on an aluminum plate, linked to a lightweight carbon-fiber frame holding MoCap markers for pose ground-truthing. The plate and frame were attached on a tripod of adjustable height and orientation, which allowed us to simulate the height and orientation of hazard and navigation cameras. The cameras were calibrated (intrinsics and stereo extrinsics) using a calibration checkerboard comprising 7×9 squares of individual size $5\,\mathrm{cm} \times 5\,\mathrm{cm}$. We depict the resulting camera setup in Fig. 1a.

The pixel angular resolution of those two setups are slightly better than Mars 2020's. We note that the objective of this paper is not to quantify our ability to complete Fetch's task with Mars 2020 cameras, but instead to assess conceptual feasibility and sensitivity to environmental conditions.

Testbed

We performed our experiments in the sandy area of the testbed, as depicted in Fig. 1b. The sand and rocks are representative of different terrain conditions that could be encountered on Mars. We operated within a $4\,\mathrm{m} \times 6\,\mathrm{m}$ area and consecutively assessed two types of rock distribution: dense or sparse. In terrains with dense distribution, approximately 4 to 10 rocks of diameter $10\,\mathrm{cm}$ or above were visible from each viewpoint, in contrast to only 1 to 4 rocks per viewpoint for terrains with sparse rock distribution. We placed a sample tube, equipped with MoCap markers on rods extending from both ends, centered in each depot. To capture the tube pose as well as the camera bar pose, we placed 10 MoCap cameras



(a) Camera acquisition setup. (b) Testbed.

Figure 1. (a): camera setup (red) with MoCap markers on rigid frame (blue). (b): data acquisition testbed with North/East/West/South lights (red) and MoCap cameras (7 depicted - blue).

(Vicon T-160) around the workspace for our test setup (6 cameras mounted high on the walls, 4 on tripods), which is a half-circle area of radius $2\,\mathrm{m}$ centered on the sample tube. Finally, we equipped the testbed with four groups of ceiling lights that were placed at the four cardinal directions and can be switched on and off separately, enabling us to assess the effect of different illumination conditions on tube localization algorithms.

4. DATASET

Reference Frames

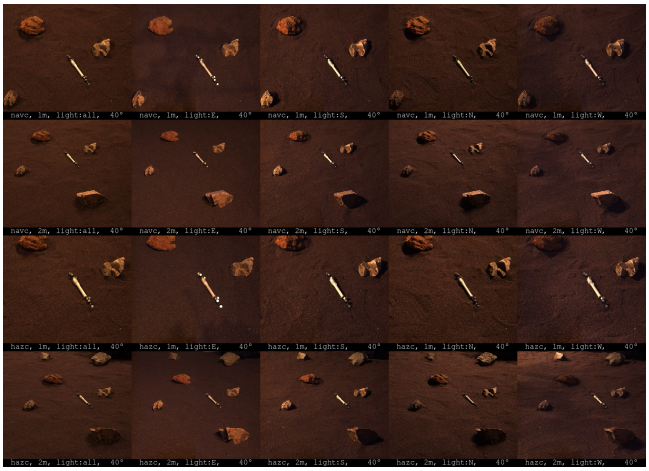
We use the notations from [4]. Given two coordinate frames $\{A\}$ and $\{B\}$, we denote by ${}^{A}\mathbf{T}_{B}$ the pose of $\{B\}$ relative to the frame defined by $\{A\}$, or the transformation from $\{A\}$ to $\{B\}$. A point $\mathbf{P}$ can thus be expressed in either coordinate frame as ${}^{A}\mathbf{p}$ and ${}^{B}\mathbf{p}$ following

$$ {}^{A}\mathbf{p} = {}^{A}\mathbf{T}_{B} \cdot {}^{B}\mathbf{p}. \tag{1} $$

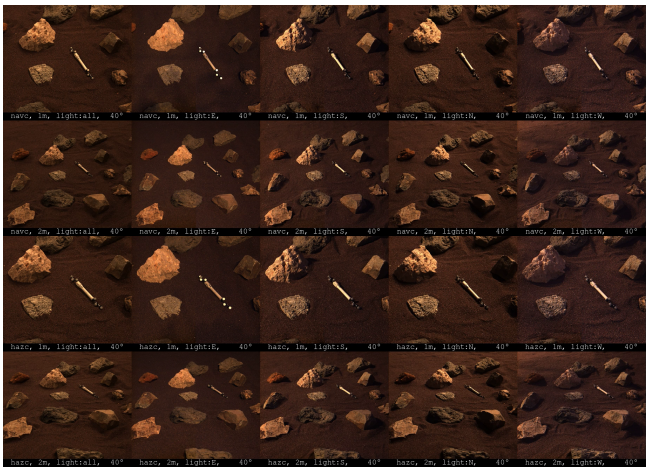
We denote by $\{O\}$ the world coordinate frame used by the MoCap motion capture system. The MoCap system outputs the 6D pose of groups (or *constellations*) of markers, expressed in frame $\{O\}$. Expressing the pose of a constellation requires the definition of a reference frame (origin and 3D orientation) for the constellation. We denote the frame of the camera bar's constellation by $\{C^V\}$. The MoCap system outputs the pose of this constellation relative to the MoCap's origin, ${}^{O}\mathbf{T}_{C^V}$. Denoting the cameras' optical frames by: $\{C^F\}$ (left Flea camera) and $\{C^G\}$ (left Grasshopper camera), we computed the rigid transformation ${}^{C^V}\mathbf{T}_{C^F}$ (resp. ${}^{C^V}\mathbf{T}_{C^G}$) between the MoCap marker constellation and the Flea (resp. Grasshopper) optical frames by simultaneously recording the 3D position of MoCap markers $\mathbf{p}_i$ tracked individually in the constellation frame ${}^{C^V}\mathbf{p}_i$ directly from the MoCap system and in the optical frame ${}^{C^F}\mathbf{p}_i$ (resp. ${}^{C^G}\mathbf{p}_i$) from stereo and aligning the correspondences $\left\{({}^{C^V}\mathbf{p}_i, {}^{C^F}\mathbf{p}_i)_i\right\}$ (resp. $\left\{({}^{C^V}\mathbf{p}_i, {}^{C^G}\mathbf{p}_i)_i\right\}$) using Eq. (1). Combining this with the constellation pose measured by MoCap ${}^{O}\mathbf{T}_{C^V}$ thus allowed us to obtain the optical frame $\{C^F\}$ and $\{C^G\}$ poses with respect to the world frame $\{O\}$:

$$ {}^{O}\mathbf{T}_{C^F} = {}^{O}\mathbf{T}_{C^V}\,{}^{C^V}\mathbf{T}_{C^F}, \tag{2} $$

$$ {}^{O}\mathbf{T}_{C^G} = {}^{O}\mathbf{T}_{C^V}\,{}^{C^V}\mathbf{T}_{C^G}. \tag{3} $$



(a) Sparse rock distribution.



(b) Dense rock distribution.

Figure 2. Images captured by the left Grasshopper camera on (a) sparse and (b) dense rock distributions. For (a) and (b), from top to bottom: viewpoints from ($0.8\,\mathrm{m}$ high, $1\,\mathrm{m}$ away), ($0.8\,\mathrm{m}$ high, $2\,\mathrm{m}$ away), ($1.6\,\mathrm{m}$ high, $1\,\mathrm{m}$ away), ($1.6\,\mathrm{m}$ high, $2\,\mathrm{m}$ away); from left to right: same viewpoint with all lights on, or only East, South, North, West.

For rest of the paper, we no longer consider camera marker constellation poses, only those of the optical frames. For the tube, a similar procedure could be taken for the tube to align marker positions from the tube MoCap constellation frame $\{T^V\}$ to an arbitrary tube frame $\{T\}$ (e.g., defined by CAD) through a transformation ${}^{T^V}\mathbf{T}_{T}$. In the following, we simply set ${}^{T^V}\mathbf{T}_{T} = \mathbf{I_4}$ the 4×4 identity matrix such that ${}^{O}\mathbf{T}_{T} = {}^{O}\mathbf{T}_{T^V}$.

Data Collection

We collected a large dataset of rock-sand-tube images by exhaustive exploration of the following capture configuration space $\mathcal{C}$:

- 2 rock distributions: sparse or dense
- 52 viewpoints for each rock distribution as follows:
  - 2 distances from the tube to the camera tripod: $1\,\mathrm{m}$ or $2\,\mathrm{m}$
  - For each such distance d, place the tripod at 13 angles along a half-circle of radius d centered on the tube: $0°$, $20°$, $40°$, $60°$, $70°$, $80°$, $90°$, $100°$, $110°$, $120°$, $140°$, $160°$, $180°$

– 2 camera heights: $0.8\,\mathrm{m}$ or $1.6\,\mathrm{m}$

• 5 illumination settings for each viewpoint: all lights on or only North, East, South, West

• 2 stereo pairs: Grasshopper (baseline $\approx 10\,\mathrm{cm}$) or Flea (baseline $\approx 30\,\mathrm{cm}$)

• 2 camera exposure settings: tuned to maximize the pixel intensity range while avoiding saturation for either the tube or the rocks. We used the latter for terrain-relative localization and reserve the former for a separate study on direct tube localization (out of scope of this paper).

In the rest of this paper, we refer to *capture configuration* as a combination of rock distribution, viewpoint, illumination, stereo pair and exposure. We thus collected a dataset of left and right stereo images over $|\mathcal{C}| = 2080$ capture configurations, totalling 4160 images annotated with ground-truth camera (left optical frame) and tube poses. In practice, 8 images were corrupted during our experiments, such that 4152 images remained usable for further analysis.

Averaging Tube Poses

We recorded the pose of all MoCap constellations (i.e., camera setup and tube) every time we captured camera images. However, we did not physically move the tube every time we moved the cameras. Instead, the tube stayed at one location for all data captured under the sparse rock settings, and at a second location for all data captured under the dense rock settings. Thus, within each setting, we could refine the tube pose by computing an average ${}^{O}\mathbf{T}_{T}$ as follows. For the translation component of ${}^{O}\mathbf{T}_{T}$, we simply took the arithmetic average of the translation component of the tube poses ${}^{O}\mathbf{T}_{T_i}$ recorded over all the configurations i under a given rock distribution setting. However, the orientation component of ${}^{O}\mathbf{T}_{T}$ could not be computed by direct arithmetic averaging of the rotation components of the corresponding ${}^{O}\mathbf{T}_{T_i}$. Instead, we computed an average rotation as the eigenvector corresponding to the largest eigenvalue of the 4×4 matrix $\mathbf{M} = \sum \mathbf{q_i} \mathbf{q_i^T}$, with $\mathbf{q_i}$ the rotation component of ${}^{o}\mathbf{T}_{T_i}$ as quaternion, following [5] (see [6] for a more extended overview on rotation averaging).

5. ROCK-SAND-TUBE SEGMENTATION

Since there is a period of 6 to 8 years between Mars 2020 dropping the tubes and the Fetch rover retrieving them, we anticipate the possibility that sand may move during that time, possibly covering certain tubes (partially or fully) while also making sand patterns unusable for localization. However, rock surfaces that are $5\,\mathrm{cm}$ above the ground or higher are likely to remain unaffected by drifts. We partially captured this constraint in our localization benchmarks, by denying the use of visual features issued from image pixels that correspond to sand. We limited our localizer to using rock features. Thus, we preprocessed our dataset to mask tube and sand areas from every image, keeping only the rocks. In preliminary experiments, we measured that it took about $30\,\mathrm{min}$ on average to mask out a single image. Instead, we trained a deep neural network to perform rock segmentation, specifically a Fully-Convolutional Neural Network (FCNN) [7] that takes as input a full-size 2448×2048 image from the dataset and directly produces an image of the same size where every output pixel models the probability of being part of a rock, tube or sand, which in turn can be thresholded to mask out sand and tube.

Layer	Kernel size	Strides	Filters
Convolution 1	24×24	4×6	256
Convolution 2	8×8	1×1	196
Convolution 3	4×4	1×1	128
Deconvolution 1	2×2	2×2	196
Deconvolution 2	4×6	4×6	256
Deconvolution 3	4×4	4×4	3

Table 1. Neural network layer parameters.

Segmentation: Training Dataset

We constructed a training dataset starting from a set of 16 images chosen across various capture configurations (rock density, camera viewpoint, stereo pair, illumination). We hand-labeled each of these 16 images at the pixel level, by drawing masks indicating what parts of the image are rocks, sand or tube, respectively depicted in blue, green and red in Fig. 3a. We further extended by mirroring (vertical, horizontal, and both), yielding 64 training annotated images in total. In order to be able to use a single network for both Flea and Grasshopper camera images, we trained it on grayscale images only (i.e., direct Flea images or converted Grasshopper images).

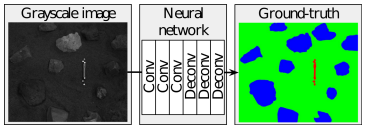
Neural Network

We built our FCNN with three convolutional layers, each followed by a maxpool layer of size 2×2 and stride 2×2, and three deconvolutional layers [7], of size summarized in Table 1. Passing as input a $2448 \times 2048 \times 1$ grayscale image of the scene, the neural network thus outputs a $2448 \times 2048 \times 3$ tensor where each element of coordinates (i, j) is a 3-element vector of the probabilities that pixel (i, j) in the input image belongs to the sample tube, sand or rocks. We then trained the neural network using a weighted cross-entropy loss over the pixels $\mathcal{I} = [1, 2048] \times [1, 2448]$ and available classes $\mathcal{C} = \{\text{rock, sand, tube}\}$:

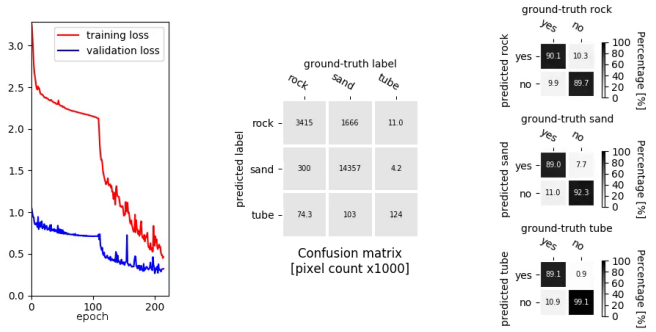
$$\mathcal{L} = \sum_{(i,j)\in\mathcal{I}} \left\{ -\sum_{k\in\mathcal{C}} \frac{1}{N_k} p_k^{(i,j)} \log q_k^{(i,j)} \right\}, \tag{4}$$

with $p_k^{(i,j)}$ and $q_k^{(i,j)}$ the respective ground-truth and predicted probabilities that pixel (i, j) is of class k, and N_k the number of pixels of class k in the training set. The $\frac{1}{N_k}$ weighting factor helped with the dataset being highly unbalanced, e.g., containing many more sand than tube pixels. We trained the neural network with 8-fold validation by minimizing the loss of Eq. (4) with the Adam stochastic optimization method [8] and a 10^{-4} initial learning rate.

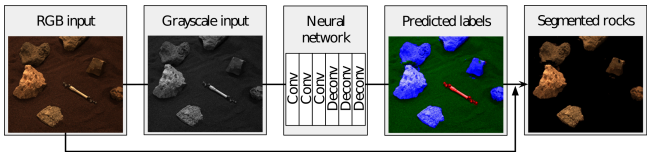
During training, we monitored the following three metrics (a) the training (and validation) loss, (b) the confusion matrix for per-pixel labelling, i.e., how many pixels were predicted to be a certain class versus their ground-truth class, and (c) the percentage of true/false positives/negatives for each class. While (a) was the loss function to be optimized, it was really a proxy for (c), the quantity we are interested in for image segmentation. We experimentally observed that extended training could lead to lower values for the loss function but worse classification accuracy. Thus, we interrupted training if the classification accuracy, here chosen as the average percentage of true positives and true negatives across all classes, stopped improving. We depict these three training metrics in Fig. 3b over 217 epochs, with the maximum



(a) Training image with ground-truth labels.



(b) From left to right: training and validation losses over 217 epochs, confusion matrix, percentage of true/false positives/negatives per class.



(c) Rock-sand-tube segmentation pipeline at inference time.

Figure 3. Neural network for image segmentation.

accuracy network being obtained at epoch 197. At inference time, we simply passed a given image through the neural network to obtain a per-pixel label mask and only kept the pixels corresponding to the class we were interested in, here the rocks, see Fig. 3c. We executed this procedure on the entirety of the dataset described in Section 4 in order to evaluate the accuracy of terrain-relative localization using rock features only. Note that when deploying terrain-relative localization on the actual Fetch rover, it sufficed to compare newly acquired images by Fetch (including sand) to images captured by the Mars 2020 rover and manually segmented on Earth between the two rover missions.

6. Terrain-Relative Localization

In this section we describe our proposed indirect localization method. At the mission level, this involves three steps: First, the Mars 2020 rover regularly collects images along its path within each tube depot, and several images of tubes and surrounding rocks after each tube release. Second, ground operators build a sparse feature map of each depot, containing the pose ${}^{O}\mathbf{T}_{T}$ of each tube $\{T\}$ with respect to rock features in the environment $\{O\}$, together with the pose ${}^{O}\mathbf{T}_{C_{\text{M2020}}}$ from which each image was taken by the Mars 2020 camera $\{C_{\text{M2020}}\}$. Finally, the Fetch rover launches, and the map is copied to its onboard computer. Upon reaching a depot, Fetch leverages the map to navigate the depot and pick up tubes.

We assumed that the Fetch rover is able reach the entry point of each depot with a 1 m accuracy using ground-in-the-loop localization [9], which allows it to localize itself with respect to the first M2020 image taken at the entrance of the depot. Fetch follows M2020's path until the first tube, and finally localizes itself with respect to a M2020 image of the tube site. Matching features between a M2020 image and a Fetch image yields ${}^{C_{\text{Fetch}}}\mathbf{T}_{C_{\text{M2020}}}$, the spatial transformation between the Mars 2020 $\{C_{\text{M2020}}\}$ and Fetch $\{C_{\text{Fetch}}\}$ rover cameras. We finally obtain the transformation from Fetch to the tube ${}^{C_{\text{Fetch}}}\mathbf{T}_{T}$ by transiting through the tube ${}^{O}\mathbf{T}_{T}$ and Mars 2020 camera ${}^{O}\mathbf{T}_{C_{\text{M2020}}}$ poses obtained previously:

$$ {}^{C_{\text{Fetch}}}\mathbf{T}_{T} = \underbrace{{}^{C_{\text{Fetch}}}\mathbf{T}_{C_{\text{M2020}}}}_{\text{feature matching}} \underbrace{{}^{C_{\text{M2020}}}\mathbf{T}_{O}}_{\text{known}} \underbrace{{}^{O}\mathbf{T}_{T}}_{\text{known}} . \tag{5} $$

The problem of terrain-relative tube localization thus boils down to that of estimating the relative transform between two camera viewpoints. In the following, we evaluate it over the multiple capture variations of the dataset described in Section 4.

Pose Estimation

We estimated transformations between stereo pairs by feature matching using the approach of [10] and the associated LIBVISO2 library[2] modified to use SIFT [11] as feature detector and descriptor, as well as "deep descriptors" [12] predicted by a convolutional neural network from image patches around SIFT keypoints. The algorithm works as follows. Consider two left and right rectified image pairs $(\mathbf{I}_{\mathbf{i}}^{\mathbf{L}}, \mathbf{I}_{\mathbf{i}}^{\mathbf{R}})$ and $(\mathbf{I}_{\mathbf{j}}^{\mathbf{L}}, \mathbf{I}_{\mathbf{j}}^{\mathbf{R}})$ taken under different but compatible capture configurations i and j (that is, two different combinations of camera viewpoint, stereo pair and lighting condition, but on the same depot, i.e., rock distribution). We computed four-way feature correspondences, using SIFT as keypoint detector, together with either SIFT or deep feature descriptors. We performed this correspondence search in a "circular" manner:

$$ \begin{array}{ccc} \mathbf{I}_{\mathbf{i}}^{\mathbf{L}} & \leftrightarrow & \mathbf{I}_{\mathbf{i}}^{\mathbf{R}} \\ \updownarrow & & \updownarrow \\ \mathbf{I}_{\mathbf{j}}^{\mathbf{L}} & \leftrightarrow & \mathbf{I}_{\mathbf{j}}^{\mathbf{R}} \end{array} , \tag{6} $$

and only kept matches such that starting from one image feature and following the circle led back to that same feature. To accelerate the correspondence search, we required matches between left and right images within the same stereo pair to be within 1 pixel of one another along the vertical axis (epipolar constraint). We depict the resulting matches between left images taken from two different viewpoints in Fig. 4, after rock segmentation. We denote by $\{C_i\}$ and $\{C_j\}$ the optical frames at capture configurations i and j, respectively. The transformation ${}^{C_i}\mathbf{T}_{C_j}$ between the two camera poses was computed by minimizing the reprojection error of all features via Gauss-Newton optimization on inliers estimated within a RANSAC scheme (20,000 iterations) when at least 6 circular matches were available.

Localization Accuracy Metrics

Let us denote by ${}^{C_i}\mathbf{T}_{C_j}$ the ground-truth (i.e., motion-capture) transformation between C_i and C_j and by ${}^{C_i}\mathbf{T}_{\widehat{C}_j}$ the corresponding transformation estimated through feature matching by considering C_i as the reference viewpoint (e.g., captured by Mars 2020 and for which operators annotated camera and tube poses). The notation $\widehat{C}_j$ indicates that the viewpoint estimated from feature matching could differ from

[2] http://www.cvlibs.net/software/libviso/

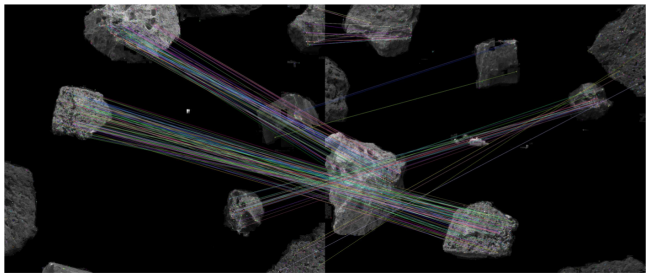


Figure 4. Matching features between segmented rocks.

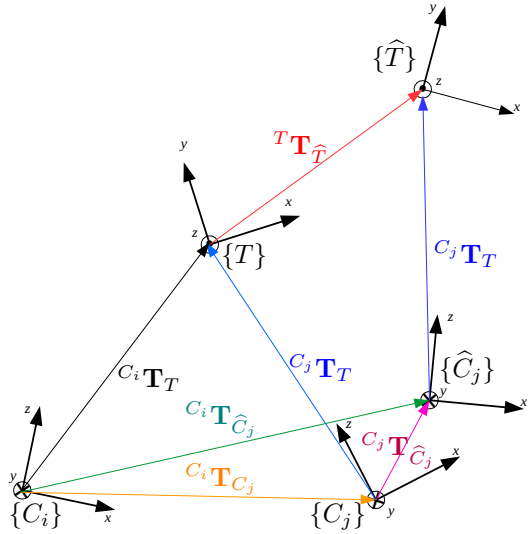


Figure 5. Composing the transformation from feature matching ${}^{C_i}\mathbf{T}_{\widehat{C}_j}$ with the ground-truth camera-to-tube transformation ${}^{C_j}\mathbf{T}_{T}$ results in an error transformation for the tube ${}^{T}\mathbf{T}_{\widehat{T}}$.

the real one. The transformation error is given by:

$$ {}^{C_j}\mathbf{T}_{\widehat{C}_j} = \underbrace{{}^{C_j}\mathbf{T}_{C_i}}_{\text{ground truth}} \underbrace{{}^{C_i}\mathbf{T}_{\widehat{C}_j}}_{\text{feature matching}} . \tag{7} $$

The optical frame transformation error was however not an adequate metric for the problem of picking up a tube: intuitively, a null camera translation error with an orientation error of a few degrees could lead to a large position error for the tube. To satisfactorily evaluate applicability to tube pick-up, we defined our metric as the position-orientation distance between (a) the pose of the tube in Fetch frame, assuming a ground-truth M2020-Fetch transformation, and (b) the pose of the tube in Fetch frame computed via visual localization. Instead of evaluating camera localization errors alone, we evaluated their effect on tube localization. To do so, we first decomposed the relative transformation between a camera frame $\{C_i\}$ and the tube frame $\{T\}$ by transiting through another camera frame $\{C_j\}$:

$$ {}^{C_i}\mathbf{T}_{T} = {}^{C_i}\mathbf{T}_{C_j}\,{}^{C_j}\mathbf{T}_{T}. \tag{8} $$

We then defined a "virtual tube" $\widehat{T}$ as that obtained by blindly following the ground-truth transformation between the real viewpoint and the real tube ${}^{C_j}\mathbf{T}_{T}$, but starting from $\widehat{C}_j$ the viewpoint obtained from rock feature matching. That is, we set:

$$ {}^{\widehat{C}_j}\mathbf{T}_{\widehat{T}} \coloneqq {}^{C_j}\mathbf{T}_{T}. \tag{9} $$

We depict the corresponding frames and transformations in Fig. 5. Compared to Eq. (8), instead of transiting through the ground-truth transformation ${}^{C_i}\mathbf{T}_{C_j}$, we computed the virtual tube pose by composing the transformation between cameras from terrain-relative localization with the ground-truth camera-to-tube transformation:

$$ {}^{C_i}\mathbf{T}_{\widehat{T}} = {}^{C_i}\mathbf{T}_{\widehat{C}_j}\,{}^{\widehat{C}_j}\mathbf{T}_{\widehat{T}} = {}^{C_i}\mathbf{T}_{\widehat{C}_j}\,{}^{C_j}\mathbf{T}_{T}. \tag{10} $$

The resulting error transformation at the level at the tube was then expressed using Eq. (10) as:

$$ \begin{aligned} {}^{T}\mathbf{T}_{\widehat{T}} &= {}^{T}\mathbf{T}_{C_i}\,{}^{C_i}\mathbf{T}_{\widehat{T}} && (11)\\ &= ({}^{T}\mathbf{T}_{C_j}\underbrace{{}^{C_j}\mathbf{T}_{C_i})({}^{C_i}\mathbf{T}_{\widehat{C}_j}}\,{}^{C_j}\mathbf{T}_{T}) && (12)\\ &= \left({}^{C_j}\mathbf{T}_{T}\right)^{-1}\,{}^{C_j}\mathbf{T}_{\widehat{C}_j}\,{}^{C_j}\mathbf{T}_{T}. && (13) \end{aligned} $$

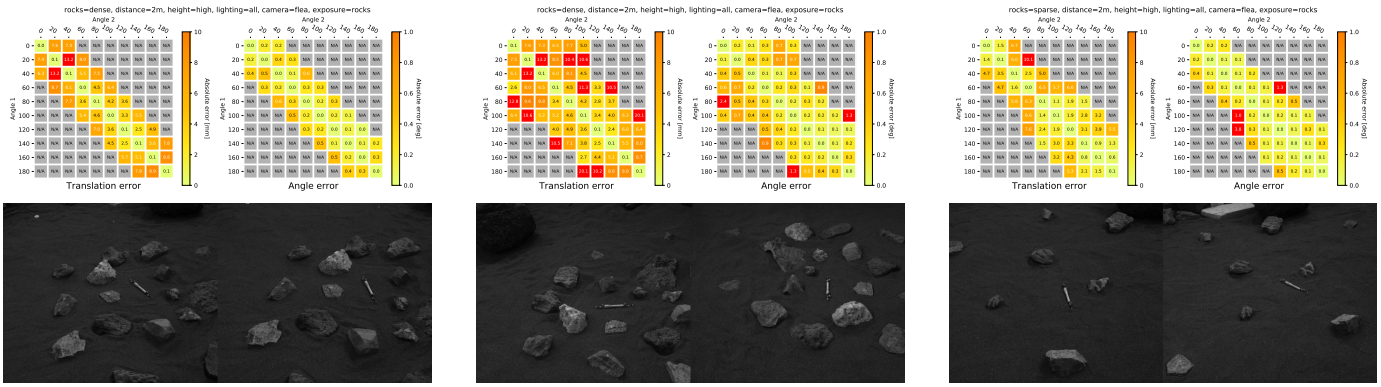
The tube error transformation ${}^{T}\mathbf{T}_{\widehat{T}}$ was thus solely expressed as function of the camera error transformation ${}^{C_j}\mathbf{T}_{\widehat{C}_j}$ and a ground-truth camera-to-tube transformation ${}^{C_j}\mathbf{T}_{T}$. In practice, we could use Eq. (13) to estimate the effect of camera error ${}^{C_j}\mathbf{T}_{\widehat{C}_j}$ for any object P placed at ${}^{C_j}\mathbf{T}_{P}$ with respect to the camera. In the rest of the paper, we estimated errors at the level of the ground-truth tube pose measured by the motion capture system but could also repeat our methodology for other locations in the depot and refine workspace constraints thusly. Finally, let us decompose the transformation error ${}^{T}\mathbf{T}_{\widehat{T}}$ into a translation vector ${}^{T}\mathbf{t}_{\widehat{T}}$ and a rotation matrix ${}^{T}\mathbf{R}_{\widehat{T}}$. We defined the tube translation error as the L^2 norm of ${}^{T}\mathbf{t}_{\widehat{T}}$, and the tube orientation error as the absolute angle in the axis-angle representation of ${}^{T}\mathbf{R}_{\widehat{T}}$.

7. Experiments

We now apply our terrain-relative localization pipeline (see Section 6) to estimate transformations between all possible capture configurations within our dataset (see Section 4), i.e., 270,920 pair-to-pair transformations. In the following, we report resulting translation and rotation errors at the level of the tube when varying capture configuration. Note that as we require at least 6 matches to estimate camera transformations from correspondences in our current implementation, if only 5 matches or less are available, then we mark the transformation calculation as unsuccessful and depict it on gray background in the error tables reported next. In our experiments, successful transformation estimates only yielded angular errors below 2° (most of them being below 1°), thus we focus on translation errors as success criterion for tube pickup. For every error table, we thus depict the viewpoint comparison producing the worst-case (largest) translation error among the successful transformation calculations, which we require to be below $5\,\mathrm{mm}$ for tube pickup based on the current Fetch rover concept being equipped with a $1\,\mathrm{cm}$-wide parallel jaw gripper. For the sake of readability, we depict the left image from both stereo pairs before image segmentation but we run our experiments using rock features only.

Preliminary Experiments: Descriptors and Rock Density

First, we report the tube estimation errors using both CNN-based (see Fig. 6a) and SIFT (see Fig. 6b) descriptors, when



(a) Deep descriptors, dense rock distribution

(b) SIFT descriptors, dense rock distribution.

(c) SIFT descriptors, sparse rock distribution.

Figure 6. Tube localization results when varying one parameter at a time between feature descriptors and rock density. Top: translation and rotation errors. Bottom: viewpoints producing worst-case translation errors over successful transformations, (a) 13.2 mm, (b) 20.1 mm, (c) 10.1 mm. Unsuccessful transformations are on gray background.

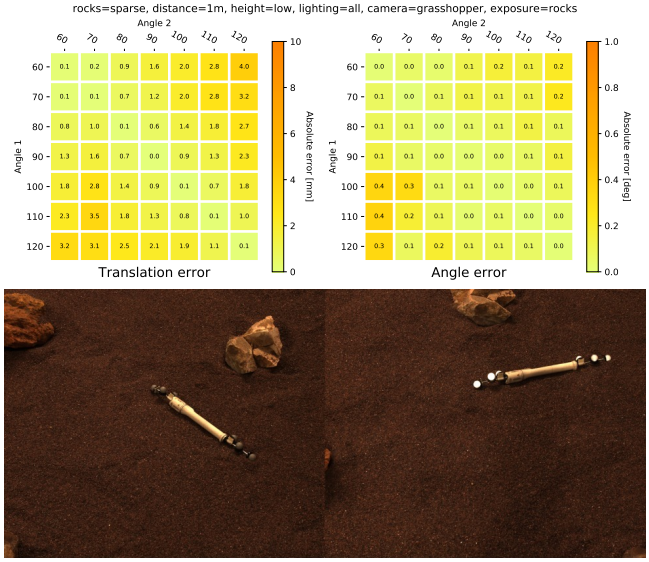


Figure 7. Top: translational and angular errors when viewing the scene from different angles. Bottom: left camera images for worst-case translation error (4.0 mm).

comparing stereo taken from two different viewpoint angles between 0° and 180° on the dense rock distribution, all other parameters being equal. We observe from Figs. 6a and 6b that transformations obtained from each are of comparable accuracy, but SIFT descriptors are able to yield transformations on a wider range of viewpoint changes. We also depict in Fig. 6c the accuracy results when using SIFT features on the sparse rock distribution and observe (as it can be expected) that the more rocks being available facilitates feature matching across larger viewpoint variations. As we observe that these two trends is consistent over different capture configurations, we choose for the rest of this study to present our results using SIFT features (i.e., the better method we can choose) on the sparse rock distribution (i.e., the more challenging distribution, which we may not be able to effect on Mars).

View Angle

We observe from Figs. 6b and 6c that pose estimation consistently works for up to 60° angle differences on the dense rock distribution and 40° on the sparse rock distribution. With these images captured from 2 m away from the tube, these angle differences correspond to, respectively, 2 m and 1.37 m distances between viewpoints. As the Fetch rover is able to position itself globally within 1 m of a chosen location, it is thus not a problem if tube localization using only rock features fails beyond that. Furthermore, while the 10.1 mm worst-case translation error for Fig. 6c is insufficient for blind grasping, it can be used to drive the rover closer to the tube, capture new images, then perform a new round of feature matching in preparation for grasping. We depict in Fig. 7 the tube localization errors for view angles between 60° and 120°. Captured 1 m away from the tube, this 60° maximum angle difference also corresponds to a 1 m maximum distance between viewpoints. We observe that all viewpoint transformations can be calculated under these capture configurations, with a worst-case translation error at the tube level of 4.0 mm, which is sufficient for tube pickup. Note that, as described previously, it is always possible to iteratively drive the Fetch rover closer to the Mars 2020 reference viewpoint, until achieving sub-millimeter tube localization accuracy.

Distance to Tube

In the previous experiment, we assessed the effects of viewpoint variations while keeping the camera setup 1 m away from the tube. In this experiment, we compare viewpoints taken from the same view angle but either 1 m or 2 m away from the tube, i.e., both viewpoints are on the same radial line with the tube but at different distances. We represent the resulting transformation errors in Fig. 8, illustrating also the scale difference between how rocks appear, and how many can be seen. We report a worst-case translation error of 8.1 mm, which is slightly too large for grasping but we can, similarly to the previous case, use to navigate closer to the tube before recalculating our pose.

Camera Height

Having assessed the effect of transformations on the horizontal plane, we now assess the effect of transformations along the vertical axis. For different view angles 1 m away from the tube, we compare stereo pairs captured when the tripod was lowered to 0.80 m above the ground and when it was raised to 1.6 m, all other capture parameters being the same. We report a worst-case translation error of 3.3 mm, which is sufficient for tube pickup.

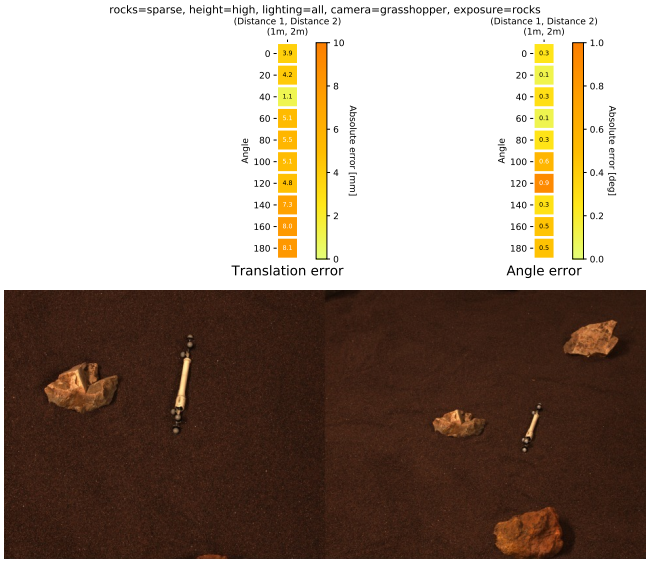


Figure 8. Top: translational and angular errors for $1\,\mathrm{m}$ vs. $2\,\mathrm{m}$ distance to the tube over different view angles, all other parameters being constant. Bottom: left camera images for worst-case translation error ($8.1\,\mathrm{mm}$).

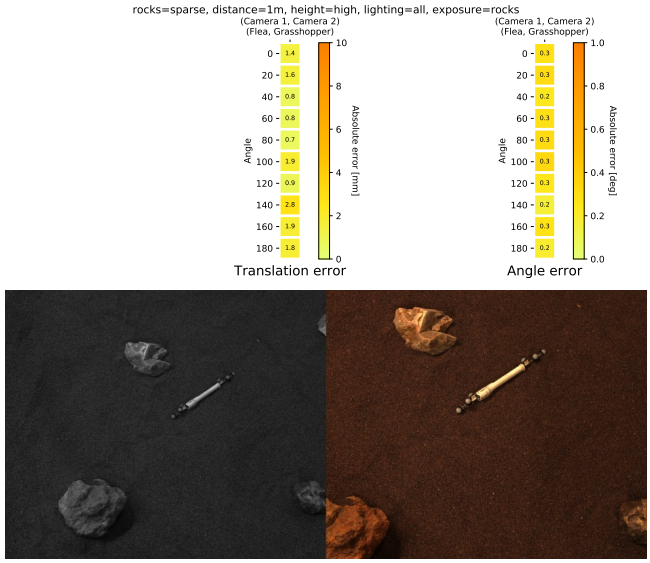


Figure 10. Top: translational and angular errors for Flea vs. Grasshopper cameras over different view angles, all other parameters being constant. Bottom: left camera images for worst-case translation error ($2.8\,\mathrm{mm}$).

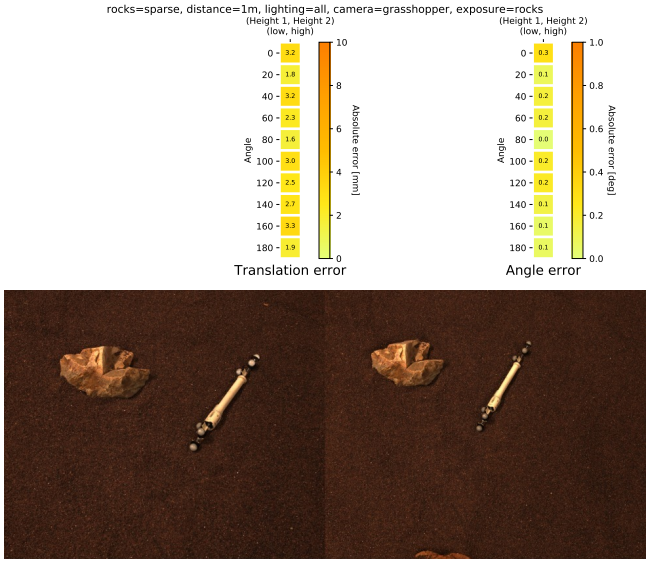


Figure 9. Top: translational and angular errors for low ($0.8\,\mathrm{m}$) vs. high ($1.6\,\mathrm{m}$) camera heights over different view angles, all other parameters being constant. Bottom: left camera images for worst-case translation error ($3.3\,\mathrm{mm}$).

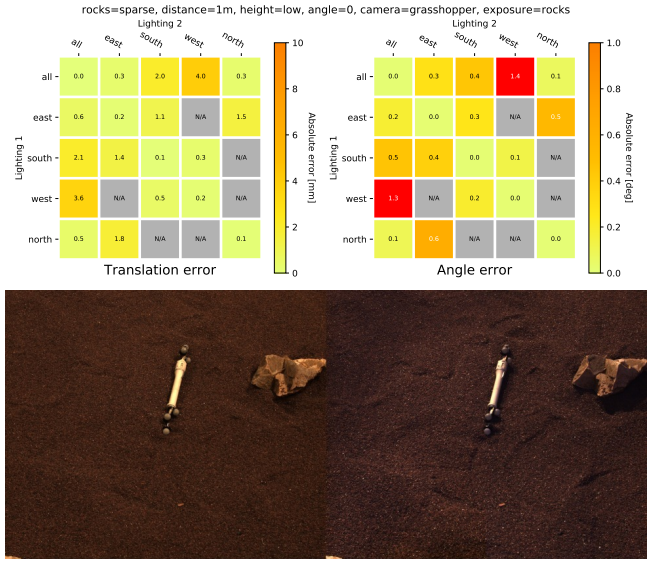


Figure 11. Top: translational and angular errors for varying lighting conditions (all lights on, or only North, East, South, West), but same viewpoint. Bottom: left camera images for worst-case translation error ($4.0\,\mathrm{mm}$).

Camera Variation

We now assess the effect of capturing images using two different sets of stereo cameras on tube localization. To do so, we take images captured by the Flea cameras as the first set of stereo pairs, and as images captured by the Grasshopper cameras as the second set. Both stereo pairs are captured at the same instant, such that the transformation between the two is constant, determined by how they were mounted onto the aluminium camera plate and calibrated, as described in Section 3. We estimate this transformation by rock feature matching, and repeat this experiment for different view angles $1\,\mathrm{m}$ away from the tube. Note that both stereo pairs are still captured at the same time, just from different test locations (i.e., their global poses change but their relative pose is constant). We report in Fig. 10 a worst-case translation error of $2.8\,\mathrm{mm}$, which is sufficient for tube pickup. Together with our previous results on camera height variations, this suggests that it is possible to leverage both types of cameras together (namely, hazcams and navcams) on the Mars 2020 and Fetch rovers, even if their parameters differ across missions.

Lighting Variations

All the results presented so far were obtained between stereo pairs captured under the same lighting conditions. This is not an assumption that can be easily enforced in practice. Instead, it is necessary that the Fetch rover can operate at

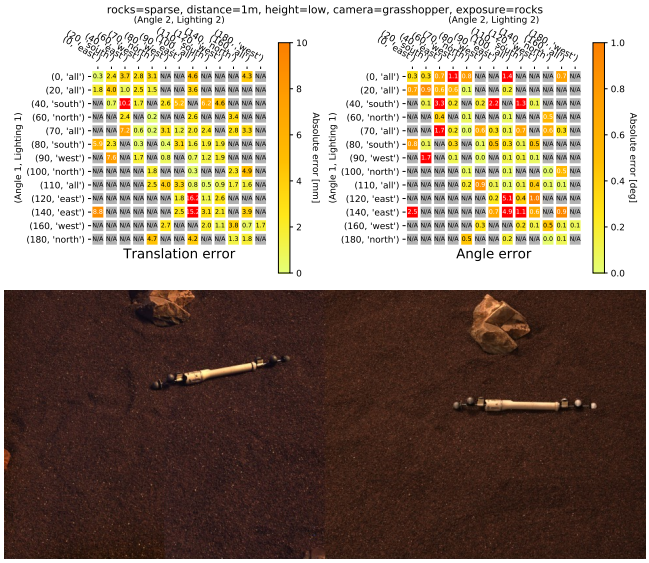


Figure 12. Top: translational and angular errors for varying lighting conditions and varying viewpoints. Bottom: left camera images for worst-case translation error ($16.2\,\mathrm{mm}$).

different times of the day and of the year, that may not overlap with the times Mars 2020 captured depot images. In order to evaluate the accuracy of terrain-relative localization under changing light conditions, we captured images from one single viewpoint under five settings. With independently controlled lights at the cardinal points (North, East, South, West), we either turned all four of them on together, or only one at a time. As all images are captured under the same viewpoint, we should ideally estimate a zero transformation when comparing stereo pairs. Instead, we obtained configurations where transformation estimation failed (i.e., less than 6 matches were found between the two stereo pairs), which we did not observe previously, even in cases with $1\,\mathrm{m}$ actual motion. We depict these results in Fig. 11. We also report a worst-case translation error of $4.0\,\mathrm{mm}$, which is comparable to the errors obtained when the viewpoints actually changed. While this level of error still permits tube pickup when sufficient matches are found, we experimentally observe that feature matching fails when comparing opposing lighting directions.

View Angle and Lighting Variations

From Fig. 11, feature matching seemed to only fail under opposing lighting conditions, i.e., between South/North and East/West. As the Fetch rover will only be expected to function over part of the day, this challenge could be partially mitigated by choosing a time range to minimize the lighting difference. However, performing terrain-relative localization on both varying lighting and varying viewpoints show that the problem is not simply solved by avoiding extreme light changes. We indeed observe in Fig. 12 that transformation estimation can fail between adjacent lighting conditions even within a 10° angle difference (i.e., a $0.17\,\mathrm{m}$ distance between viewpoints), e.g., 100° view angle, North light vs. 90° view angle, East light. We thus identify that changes in lighting conditions are a major challenge for terrain-relative localization.

8. Conclusion

The autonomous recovery of sample tubes, occurring years later after having been dropped, is a challenging problem due to the uncertainty from both how the environment might change over time and the sparsity of terrain features we can rely on for localization. In this paper, we have presented a complete localization and testing pipeline including: a new image dataset spanning multiple capture modalities with ground-truth poses from a motion capture system; a data-driven model for rock-sand-tube segmentation to enforce the consequences of time over the usable image features (namely, sand moving and covering the tube); and a terrain-relative localization algorithm leveraging ground-truth pose annotations from Mars 2020 operators with feature matching with respect to Fetch images using only rocks.

Our experiments showed that our approach is then robust to varying camera viewpoints (namely: view angle, distance to tube and height) as well as camera type (different baselines and fields of view, representative of different camera setups between Mars 2020 and Fetch hazcams/navcams), enabling tube localization within $5\,\mathrm{mm}$ directly in most cases (sufficient for tube pickup). We also showed that when tube localization accuracy is insufficient for direct pickup, it remains in the cm-scale, which lets us leverage the imperfect transformation estimate for precision re-positioning before performing a new iteration of terrain-relative localization.

However, we determined that changes in lighting conditions were a considerable challenge in terrain-relative localization. We indeed observed that transformation estimation could fail under large changes of light direction, even when the camera viewpoint did not change at all. Furthermore, we observed it failing under small lighting changes combined with small viewpoint changes. As it is not reasonable to constrain the Fetch rover to only operate when its lighting conditions perfectly match that of Mars 2020 many years earlier, it is critical to develop new localization methods that are robust to lighting changes specifically. To this end, we are currently evaluating the use of shadow removal and synthetic relighting techniques [13]. This would enable, for example, the generation of a new appearance model of the depot captured by Mars 2020, but under Fetch lighting conditions. The generated model could then be used in combination with localization methods that compare synthetic renderings of a textured mesh to the current view [14], [15].

Some uncertainties remain beyond the sensitivity of our techniques to lighting variations. First, terrain-relative localization is ultimately contingent on the possibility of having an accurate map of the depot from Mars 2020 images. Thus, as future work, we would like to assess the effect of pose annotation uncertainties in terrain-relative localization, and possibly other sources such as camera calibration errors, noise in the MoCap system, stochasticity in the localization algorithms, etc. While we were able to extract feasibility assessments from select slices of a 6D configuration space (rock density, view angle, distance to tube, camera height, camera type, lighting direction), we would also like to develop a systematic way to explore the space and, e.g., automatically extract configuration sets satisfying chosen accuracy requirements. We are currently constructing an outdoor dataset to demonstrate robustness to natural lighting conditions. We are also defining an end-to-end test campaign that uses a rover prototype to show hardware tube pickup. Finally, we are constructing failure recovery procedures that address failed terrain-relative grasps, with either ground-in-the-loop input or, in cases where the failed grasp freed the tube from occluding sand,

by directly segmenting the tube in Fetch's camera images.

ACKNOWLEDGMENTS

The research described in this paper was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. The information presented about potential Mars Sample Return architectures is provided for planning and discussion purposes only. NASA has made no official decision to implement Mars Sample Return.

REFERENCES

[1] R. Mattingly and L. May, "Mars sample return as a campaign," in *2011 Aerospace Conference*. IEEE, 2011, pp. 1–13.

[2] J. Papon, R. Detry, P. Vieira, S. Brooks, T. Srinivasan, A. Peterson, and E. Kulczycki, "Martian fetch: Finding and retrieving sample-tubes on the surface of mars," in *2017 IEEE Aerospace Conference*. IEEE, 2017, pp. 1–9.

[3] J. Maki, C. McKinney, R. Sellar, D. Copley-Woods, D. Gruel, D. Nuding, M. Valvo, T. Goodsall, J. McGuire, and T. Litwin, "Enhanced engineering cameras (eecams) for the mars 2020 rover," in *3rd International Workshop on Instrumentation for Planetary Mission*, vol. 1980, 2016.

[4] P. Corke, *Robotics, Vision and Control: Fundamental Algorithms In MATLAB® Second, Completely Revised*. Springer, 2017, vol. 118.

[5] F. L. Markley, Y. Cheng, J. L. Crassidis, and Y. Oshman, "Averaging quaternions," *Journal of Guidance, Control, and Dynamics*, vol. 30, no. 4, pp. 1193–1197, 2007.

[6] R. Hartley, J. Trumpf, Y. Dai, and H. Li, "Rotation averaging," *International journal of computer vision*, vol. 103, no. 3, pp. 267–305, 2013.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[8] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[9] T. Parker, M. Malin, F. Calef, R. Deen, H. Gengl, M. Golombek, J. Hall, O. Pariser, M. Powell, R. Sletten *et al.*, "Localization and'contextualization'of curiosity in gale crater, and other landed mars missions," in *Lunar and Planetary Science Conference*, vol. 44, 2013, p. 2534.

[10] A. Geiger, J. Ziegler, and C. Stiller, "Stereoscan: Dense 3d reconstruction in real-time," in *2011 IEEE Intelligent Vehicles Symposium (IV)*. Ieee, 2011, pp. 963–968.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[12] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 118–126.

[13] J. Philip, M. Gharbi, T. Zhou, A. A. Efros, and G. Drettakis, "Multi-view relighting using a geometry-aware network," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, p. 78, 2019.

[14] G. Pascoe, W. Maddern, A. D. Stewart, and P. Newman, "Farlap: Fast robust localisation using appearance priors," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 6366–6373.

[15] K. Ok, W. N. Greene, and N. Roy, "Simultaneous tracking and rendering: Real-time monocular localization for mavs," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 4522–4529.

BIOGRAPHY

Tu-Hoa Pham *is a Robotics Technologist at the NASA Jet Propulsion Laboratory, Caltech Institute of Technology, currently working on machine vision for Mars Sample Return. He holds a Diplôme d'Ingénieur in Aerospace Engineering from ISAE-SUPAERO (2013), an M.Sc. in Applied Mathematics from Université Paul Sabatier (2013) and a Ph.D. in Robotics from Université de Montpellier (2016), which he conducted at the CNRS-AIST Joint Robotics Laboratory on the topic of force sensing from vision. Prior to joining JPL in 2018, he spent two years as a research scientist at IBM Research Tokyo, where he worked on deep reinforcement learning for robot vision and manipulation in the real-world.*

William Seto *is a Robotics Technologist at NASA's Jet Propulsion Laboratory. He joined JPL in 2017 after receiving his M.S. in Robotic Systems Development from Carnegie Mellon's Robotics Institute. He develops software to enable autonomous capabilities in maritime and terrestrial environments. His outside interests include soccer and chicken tenders.*

Shreyansh Daftry *is a Robotics Technologist at NASA Jet Propulsion Laboratory, California Institute of Technology. He received his M.S. degree in Robotics from the Robotics Institute, Carnegie Mellon University in 2016, and his B.S. degree in Electronics and Communication Engineering in 2013. His research interest lies is at intersection of space technology and autonomous robotic systems, with an emphasis on machine learning applications to perception, planning and decision making. At JPL, he has worked on mission formulation for Mars Sample Return, and technology development for autonomous navigation of ground, airborne and subterranean robots.*

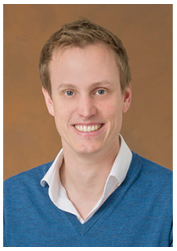
Alexander Brinkman *received his M.S in Robotic Systems Development from Carnegie Mellon's Robotic Institute, then joined the Robotic Manipulation and Sampling group at Jet Propulsion Laboratory in 2017. He develops manipulation software and autonomous capabilities to enable future sampling missions to Europa, Enceladus, Mars, and comets.*

John Mayo *is a robotics mechanical engineer in the Robotic Climbers and Grippers Group at JPL. John received a Bachelor of Science in Mechanical Engineering from Texas A&M in 2014 and Master of Science of the same from the Massachusetts Institute of Technology in 2016. As part of his graduate studies, John worked on hardware for the HERMES humanoid robot, developing a hybrid hand-foot device under direction of Sangbae Kim. Additionally, John co-founded and led the MIT Hyperloop Team to design and build a magnetically levitated vehicle and participated as a mentor in the new student-led shop, MIT Makerworks.*

Yang Cheng *is a principal member of the Aerial System Perception Group at JPL. Dr. Cheng is a leading expert in the areas of optical spacecraft navigation, terrain relative navigation, surface robotic perception and cartography. He has made significant technical contributions in technology advancement in the area of structure from motion, 3D surface reconstruction, surface hazard detection and mapping for spacecraft landing site selection, stereo sub-pixel interpolation, wide area surveillance, landmark recognition, map projection etc. Dr. Cheng is the key algorithm developer of MER descent image motion estimation system (DIMES) and Mars2020 lander vision system (LVS), which will be part of Mars 2020 EDL system.*

Curtis Padgett *is currently the Supervisor for the Aerial Perception Systems Group and a Principal in the Robotics Section at the Jet Propulsion Laboratory. Dr. Padgett leads research efforts focused on aerial and maritime imaging problems including: navigation support for landing and proximity operations; automated, real-time recovery of structure from motion; precision geo-registration of imagery; automated landmark generation and mapping for surface relative navigation; stereo image sea surface sensing for navigation on water and image based, multi-platform contact range determination. He has a Ph.D. in Computer Science from the University of California at San Diego, and has been an employee of JPL since graduating in 1995.His research interests include pattern recognition, image-based reconstruction, and mapping.*

Eric Kulczycki *received a dual B.S degree in Mechanical Engineering and Aeronautical Science and Engineering from the University of California, Davis, in 2004. He received his M.S. degree in Mechanical and Aeronautical Engineering also from the University of California, Davis in 2006. He is a member of the engineering staff at the Jet Propulsion Laboratory, California Institute of Technology, where he is currently involved in Mars sample transfer chain technology development, sampling technology development for extreme environments, Mars 2020 Sample Caching Subsystem, and mechanical design of various mobility platforms. He has worked at JPL for over 15 years.*



Renaud Detry *is a research scientist at NASA JPL, and a visiting researcher at ULiege/Belgium and KTH/Stockholm. Detry earned Master's and Ph.D. degrees in computer engineering and robot learning from ULiege in 2006 and 2010. Shortly thereafter he earned two Starting Grants from the Swedish and Belgian national research institutes. He served as a postdoc at KTH and ULiege between 2011 and 2015, before joining the Robotics and Mobility Section at JPL in 2016. His research interests are perception and learning for manipulation, robot grasping, and mobility. At JPL, Detry is machine-vision lead for the Mars Sample Return technology development program, and he conducts research in autonomous robot manipulation and mobility for Mars, Europa, Enceladus, and terrestrial applications.*