Monocular Visual Pose Estimation via Online Sampling for Mars Sample-Tube Pickup

Bhoram Lee
GRASP Laboratory
University of Pennsylvania
Philadelphia, PA 19104
215-898-5814
bhorlee@seas.upenn.edu

Renaud Detry
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109
818-354-4729
detry@jpl.nasa.gov

Jasmine Moreno
University of California, Log Angeles
LA, CA 90095
310-825-4321
jmoreno@ucla.edu

Daniel D. Lee
GRASP Laboratory
University of Pennsylvania
Philadelphia, PA 19104
215-898-5814
ddlee@seas.upenn.edu

Eric Kulczycki
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109
818-354-4729
eric.a.kulczycki@jpl.nasa.gov

Abstract—**The 2011 Planetary Science Decadal Survey identified making significant progress towards Mars sample return (MSR) as a top priority for NASA, where a three-mission concept for MSR is currently being investigated. The Mars 2020 mission is intended to collect samples which will be sealed in tubes and left on the surface for potential return to Earth by a future mission. This paper studies the problem, autonomous pick-up of sample tubes using a monocular camera attached to a robot's end-effector. We estimate the pose of the tube and the gripper in a single image where both are visible, and compute incremental arm motions based on the relative transformation between the tube and the gripper. To estimate the pose of the tube in the tool-camera frame, we suggest an online sampling-based approach using image gradients in a coarse-to-fine framework. In order to reduce the search space, we employ a biased sampling strategy based on the target shape and projective geometry, and learn promising sampling regions on the fly. Our experiments demonstrate the effectiveness and efficiency of our strategy toward precise Martian sample retrieval.**

TABLE OF CONTENTS

1. INTRODUCTION

The NRC's Planetary Science Decadal Survey identified making significant progress towards Mars sample return (MSR) as the top priority long- term goal for NASA [1], and this was endorsed by NASA's Mars Program Planning Group (MPPG) [2]. The Jet Propulsion Laboratory (JPL) is currently investigating a three-mission concept for the potential MSR [3]. The Mars 2020 rover will core and collect about thirty promising rock and soil targets. These samples will be hermetically sealed in the tubes and deposited on the Martian surface for potential return to Earth. In the current concept, a potential future mission, with a sample retrieval lander (SRL), would collect the sample tubes and load them into an Orbiting Sample (OS) payload in a Mars Ascent Vehicle (MAV). The MAV would release the OS into Martian orbit, which could then be collected by a third mission, a Sample Return Orbiter (SRO). Once returned to Earth, these samples would be studied by scientists in laboratories with special room-sized equipment that would be too large to take to Mars. This would help them to answer fundamental questions concerning signs of past life or the modern Martian environments as habitats.

The presented work is a continuation of the efforts to demonstrate robust and repeatable retrieval of a sample tube from a Mars-like environment for potential MSR ([4] and [5]). While the study presented in [4] served as an initial overall proof-of-concept, the work by Papon *et al.*[5] focused on localizing the tubes in the visual field of a stereo camera fixed on the robot's chassis or mast. This study refines Papon's solution by parameterizing tube pick-up with images captured with a camera attached to the robot's end effector (toolcam). Given an approximate 3D location of the tube from the mast or chassis stereo camera, the toolcam can be posed to capture both the tube (on the ground) and the gripper (attached near the toolcam) in very near sight. Although 3D pose inference from monocular vision is inherently difficult, using a toolcam guarantees no occlusion by parts of the manipulator as well as minimal calibration errors.

In this paper, we present a study of 3D pose estimation of sample tubes via online sampling using a monocular camera attached near the end-effector. Unlike the conventional approach that prepares a large pre-defined set of templates or models associated with 3D poses, we exploit a geometric prior for constrained sampling online, and learn the distribution of promising poses on the fly for further improved sampling. Thus, our approach avoids problems that occur with offline training techniques, such as extensive computation and overfitting to the training data. To this end, we explicitly approach the problem in a coarse-to-fine fashion using multiple resolutions of image. The overall framework will be applicable to alternative implementations of details (for example, the choice of features) for potential extended studies on vision for grasping the sample tubes.

978-1-5386-2014-4/18/$31.00 ©2018 IEEE

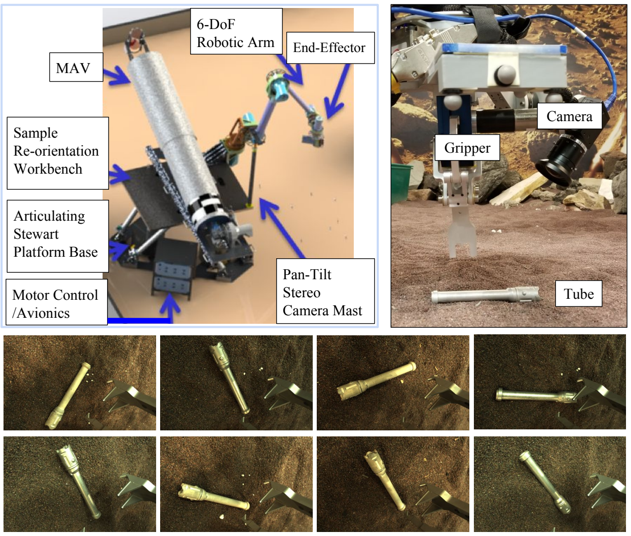

6-DoF Robotic Arm
End-Effector
MAV
Sample Re-orientation Workbench
Articulating Stewart Platform Base
Motor Control /Avionics
Pan-Tilt Stereo Camera Mast
Camera
Gripper
Tube

Figure 1. (Top-left) The MSTT platform, (Top-right) the toolcam at the end-effector, and (Bottom Rows) examples of tube images taken from the camera.

2. Mission Concept & Test Platform

As mentioned in [5], localizing and grasping sample tubes would require a considerable amount of time if accomplished by a remote human operator. To understand the time commitment, the current estimate of sample drilling operation is at least five Martian sols to execute since commands of the rover are only sent once per sol. If one assumes that retrieving a sample tube requires as much human intervention as a drilling campaign, retrieving thirty sample-tubes would require nearly half a year. If human intervention could be removed, then sample-tubes could potentially be recovered in a single sol. Automated caching could greatly accelerate the Sample Retrieval Lander (SRL) mission concept timeline and thus save other resources.

Localization of the rover and the sample-tubes is critical in enabling the automation. As in [5], we assume that the global positioning of the rover is obtained to one-meter-accuracy. Also, it is assumed that centimeter-accurate localization of tubes can be achieved by either a feature-based visual localization method or direct stereo observation as suggested in [5]. However, considering the commonly faced issue of camera-arm calibration errors and self-occlusions, accurate visual detection by a mast- or chassis-mounted camera may not lead to successful grasping. Hence, we see the need to investigate pose estimation using a toolcam.

Our problem is structured in the following way: First, the 3D models of the tube and the gripper are known since they are designed by the mission. Second, an initial pose estimate is given by the mast- or chassis-mounted stereo camera. Our aim is to build a visual pose estimation algorithm that exploit an image from the toolcam and a pose prior from the mast/chassis camera. The resulting output of this study will inform robotic grasp planning for sample tubes and can serve as a baseline for potential extended studies as well, which will be discussed in later section.

All data collection, developments, and experiments are done using resources of the Mars Sample Transfer Testbed (MSTT) platform at JPL (Fig. 1).

3. Image-based 3D Pose Estimation

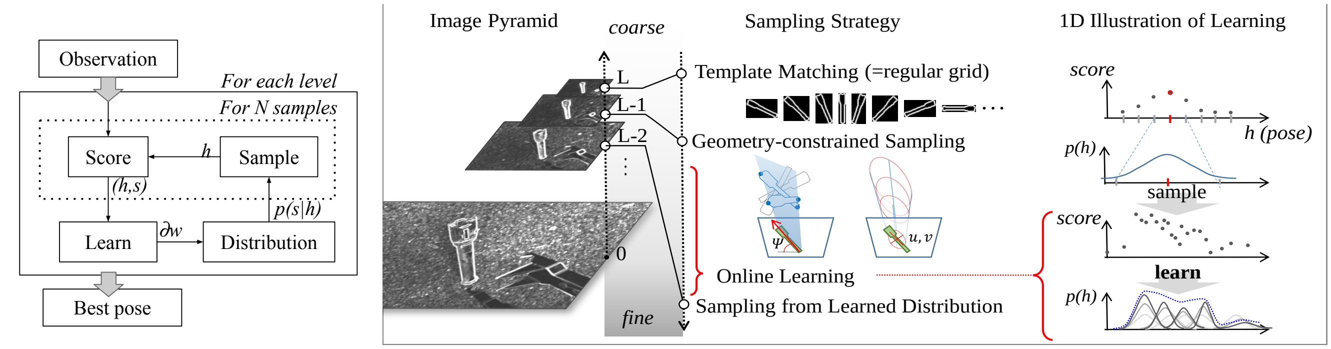
Contrary to our intuition that knowing 3D pose of a cylinder-shaped object would be easy, obtaining 3D information given only a single 2D image is an ill-posed problem. In addition, the bland shape of our target makes the problem challenging as the lack of 2D and 3D features adds ambiguity in determining 3D pose. With additional data such as range or depth observation, the difficulty could be alleviated. For example, there exist global algorithms to solve 3D pose estimation given enough time or computational resources [6].

Nevertheless, there have been much work on monocular 3D pose estimation since the early years of computer vision; there is strong motivation for this because stereo vision or depth information is much more expensive and not always available. Methods based on edge or line extraction [7] can be beneficial for theoretical studies or industrial settings where the background or light conditions can be controlled. However, it is hard to compute those features robustly. Pose estimation can be also regarded as a classification problem in some applications, where a precise pose is not required (for example, [8]). Yet, knowing pose in the 6DoF representation is critical for 3D interactive applications such as grasping.

Most recent studies on 6DoF pose estimation for general shapes learn a set of templates or a prediction model and find the best solution within the trained model at the test time ([9] [10] [11] [12] [13]). In general, off-line learning could easily fail to generalize to unseen data. In fact, it is even impossible to consider all feasible 3D poses, which is infinite, at the time of training. Powered by large-scale datasets, recent data-driven learning approaches look promising in dealing with certain applications. However, the idea of compiling large-scale data of sample-tubes on the Martian surface would not be appropriate at this time. Even if we were able to collect a reasonable amount of data, the problem of overfitting or bias to the training data is still inevitable. In order to avoid the inherent limitation of training a model, we suggest an online sampling framework in which our knowledge of the task can be encoded efficiently.

To this end, the feature selection will be critical as in any learning problem. In this paper, we chose image gradient, one of the most primitive image features. Despite its simplicity, the image gradient can be a reasonably good feature—it is very unlikely to have a random image on Mars that produces a high correlation with the boundary image of a tube. Having said that, the main contribution of the study lies on the high-level framework, which can be extended and tested with other sophisticated features.

Lastly, there have been efforts to solve vision and grasping in a tightly combined framework. Visual servoing techniques [14] [15] attempt to directly relate images with control inputs. However, they basically require known correspondences of a set of key-points defined on the target, which is hard to obtain for the tube because it is textureless and symmetric. An end-to-end learning approach for grasp control [16] would be powerful, yet this is still restricted to relatively simple lab environments under many assumptions. Also, it is not clear how to incorporate available priors of the mission into the approach. In this study, we focus on visual perception to estimate the 3D pose of a sample-tube, which will be followed by grasp planning to pick it up.



(a) (b)

Figure 2. (a) The hierarchical online sampling logic (b) Specific sampling strategies for each level: At the highest level ($l = L$), it relies on a small set of templates to find 2D location. At the next level, the sampling region is constrained based on geometrical priors such as camera projection and target shape together with the found 2D location. As more and more samples are evaluated, the distribution of high-scored points can be learned and used for improved sampling.

4. Pose Estimation via Online Sampling

This study explicitly exploits a hierarchy of visual resolutions in pose sampling. Our method starts from a coarse 2D template matching and continues to sample 3D poses for higher resolution images based on online information as well as available prior knowledge.

Technical details will be described later in this section, but let us give a sketch of how sampling is planned as the resolution level propagates. At the very initial search, we only look for 2D clues of the tube location, i.e. the pixel coordinates (u and v) and tilt (ψ) on the image plane. In a very low resolution, the two-dimensional information becomes dominant and any attempt to find details would be less meaningful. Considering the uncertainty and noise of the measurement, we model the 2D variables as Gaussians, which are used to sample them.

Given the 2D samples, we consider sampling 3D poses at the next level. A rough range of depth is known from visual detection by the stereo camera of the platform. Then we may sample z from a uniform distribution $U[z_{min}, z_{max}]$. The other 3D coordinates, x and y, of a pose sample can be determined by the projective relation with u and v, given camera parameters and the depth sample z. This allows us to sample more likely points that are consistent with the image observation. Also for orientation sampling, we take the advantage of the projective geometry together with the Euler-angle representation. The two Euler angles other than the angle around the symmetry-axis must be correlated to appear in a certain tilt angle ψ on the image. Thus, using the fact that one Euler angle is dependent on another given ψ, we can reduce one dimension for pose sampling .

After testing many samples, we may pick the best sample for the estimate. However, there exists a trade-off between the resolution and the number of samples: finer estimation can be achieved in a higher resolution but a larger number of samples is required, whereas fewer samples would work in a lower resolution but only provide a coarser solution. Our strategy for this is to learn the distribution of score over pose in order to revisit the regions of high score later. This process can be repeated for each level, and the sampling resolution can be increased accordingly.

The idea of our hierarchical online sampling is illustrated in Fig. 2. Now let us present a more technical description of our method.

Problem Statement

Our goal is to estimate the 3D pose of the tube, including rotation R and translation t, from an image observation I. Its 3D shape model and the camera parameters are available in advance, which allows us to render a tube image given a pose. Let us call a rendered image I_h using a pose hypothesis $h \in SE(3)$, a feature extraction function of an image $f(I)$, and a score function of two inputs $s(\cdot, \cdot)$. Then, we may write the problem as follows,

$$\hat{R}, \hat{t} = \arg \max_{R,t \in SE(3)} s(f(I_h), f(I)). \tag{1}$$

With this in mind, our implementation adopts the magnitude of image gradient G as the feature $f(\cdot)$, and the image correlation $corr(\cdot, \cdot)$ as the score function. Then, we can rewrite Eq. (1) as,

$$\hat{R}, \hat{t} = \arg \max_{R,t \in \mathcal{H}} corr(G_h, G). \tag{2}$$

where $\mathcal{H}$ is a subset of $SE(3)$. Since there is no analytic way to express the cost function, we evaluate the score by sampling. In theory, it is possible to obtain a solution close to the optimal as we infinitely increase the number of samples (as far as the score function reflects the actual score). However, in practice, we must restrict the search range to a certain set of hypotheses $\mathcal{H}$ as well as the number of samples. Thus, in this problem, it is important how to define the pool of sample poses $\mathcal{H}$. The rest of this section discusses our coarse-to-fine strategy to obtain a good set of $\mathcal{H}$.

Initial Search

We build a pyramid of gradient images $\{G^0, G^1, ...G^L\}$, where L represents the top level of the pyramid. For the highest level, $l = L$, we prepare a small set of templates $\{G_h^L\}$ from a predefined set of poses $\mathcal{H}^L$ for the initial 2D localization on the image. Specifically, $\{G_h^L\}$ are a set of cropped tube images where the target is centered and rotated 5 degrees around the optic axis of the camera while other dimensions are fixed, including depth. The image size is reduced up to 48×30 (pixels), and what is important at this coarse resolution is only u, v, and ψ on the image.

The best estimate is obtained by template matching based on the image correlation criteria:

$$(\hat{u}, \hat{v}, \hat{\phi}) = \arg \max_{g \in \{G_h^L\}} corr(g, G^L). \tag{3}$$

Then, we model the variables as Gaussians with a covariance corresponding to the resolution of image and templates.

$$\begin{aligned} \begin{bmatrix} u \\ v \end{bmatrix} &\sim N\Big(\begin{bmatrix} \hat{u} \\ \hat{v} \end{bmatrix}, R^\top(\hat{\psi}) \Sigma R(\hat{\psi}) \Big) \\ \psi &\sim N(\hat{\psi}, \sigma_\psi) \end{aligned}. \tag{4}$$

Geometry-constrained Sampling

After obtaining a rough 2D location of the tube (Eq. (4)), we sample poses within its neighborhood for the next level. Instead of naively sampling uniformly on $SE(3)$, we impose known projective geometric priors to bias the sampling.

First, $z \sim U[z_{min}, z_{max}]$, and x and y can be simply computed as $x = (u - c_x)z/f_x$ and $y = (v - c_y)z/f_y$ where f_x and f_y are the camera focal lengths of the x- and y-axis, and c_x and c_y are the image center.

Next, let $\boldsymbol{\theta} = [\ \theta_1 \quad \theta_2 \quad \theta_3\ ]$ be the Euler angle representation of the orientation of the tube, where θ_3 is the angle around the symmetry-axis. A derivation of the second angle θ_2 as a function of other variables can be found in Appendix. The resulting model can be written as,

$$\theta_2 \sim N(\hat{\theta}_2(\theta_1, \psi, [\ x \quad y \quad z\ ]) + s\pi, \sigma_\theta). \tag{5}$$

where $s \sim Ber(0.5)$ represents the flip between 0 or 1 according to the Bernoulli distribution, and σ_θ is the uncertainty parameter.

Online Learning for Improved Sampling

Having evaluated many samples, we get statistics of which regions may give higher scores. We use a radial basis function (RBF) network [17] with the squared exponential (Gaussian) nodes to capture this tendency of the score distribution over the pose domain. We chose this function in order to use it as a Gaussian mixture model to generate samples after learning. The i-th node of the network can be written as,

$$\phi_i(\mathbf{x}|\boldsymbol{\mu}_i, \sigma) = \exp(-\frac{||\mathbf{x} - \boldsymbol{\mu}_i||^{\mathbf{2}}}{\mathbf{2}\boldsymbol{\sigma}^{\mathbf{2}}}) \tag{6}$$

where $\mathbf{x}$ is the input of the network, and $\hat{s} = \sum_i w_i \phi_i(\boldsymbol{x})$ is the scalar output. The network is learned to minimize the squared loss with the L_1 regularizer,

$$\mathcal{L} = (\hat{s} - s)^2 + \lambda \sum_i |w_i|. \tag{7}$$

The means $\boldsymbol{\mu}_i$ are randomly initialized from the same distribution used for sampling, and weights are uniformly initialized ($w_i = 1/N$). The parameters are updated via stochastic gradient descent with a scheduled learning rate. The spherical variance parameter is used for σ and it is not learned to avoid overfitting. The regularization plays a role to force the network to have a sparse set of dominant nodes. After learning, nodes with small weights are removed and new nodes are regenerated from the major nodes with added noise. When used for sampling a pose, a Gaussian unit is randomly chosen by the normalized weight of the node and a pose is sampled from that Gaussian.

5. Experimental Results

We collected a set of tube images and also generated ground truth poses for a subset of it (24 images) by manually adjusting the 3D rendered model on the images for evaluation. The depth of the tube from the toolcam was kept within range of 18 to 25 cm and severe occlusions were not considered at this point of the study. Our implementations were written in C++ using OpenCV and OpenGL libraries. The Contrast Limited Adaptive Histogram Equalization (CLAHE) algorithm [18] was used in the preprocessing step to enhance the gradient quality under different light conditions. We extracted the binary boundary image of the tube mesh model rendered using the OpenGL shader for a sampled 3d pose. Lighting effects or other edge features are not considered for the rendering.

In the tests, we compared the following three methods: (1) naive uniform sampling, (2) geometry-constrained sampling (without learning), and (3) geometry-constrained sampling with learning. We examined the all scores of samples generated by each method, the scores of final estimates, and the pose errors of the estimates. As mentioned earlier, the image correlation is used as the score (OpenCV *matchTemplate()* with the option method="CV_TM_CCORR"). Since each sampling method includes random factors, all tests are repeated ten times and the statistical performance is reported. We also varied the number of samples to see the estimation performance.

All methods started from initial template matching with the image size reduced to 48 × 30 ($l = L$). The number of templates were 72 and the entire initial step took only 10 $msec$. After the initial search, the suggested method (method (3)) sampled initial 3000 images of 240×150 ($l = L - 1$)for learning the score distribution and then moved to 480 × 300 ($l = L - 2$) for the rest. Other methods (methods (1) and (2)) sampled images of 480×300 for a fair comparison. Although we considered three levels (L to $(L - 2)$), the learning-and-sampling scheme could be extended to higher resolutions and higher dimensions. Since it is usually difficult to deal with high dimensional data, we limited the dimension to four (x, y, z, and θ_1) in order to better observe the behavior and performance of the method. Hence, θ_2 was determined by the geometry-constrained sampling, and θ_3, which appeared ineffective to the score in most cases, by uniform sampling.

Score Distribution

Fig. 3 shows the histogram of scores of all test instances for each method, and it demonstrates that geometry-constrained sampling with online learning (Geo-OL) actually produced samples of better quality than the uniform sampling (Unif) and the geometric-sampling without online learning (Geo). The stretch of the tail of 'Geo-OL' toward higher scores results in higher chances of getting better samples in every trial. Note that the scores are normalized by the maximum correlation score of the corresponding image because the absolute value differs from image to image. Each trial returns a best estimate that has the maximum score among its sample pool, and Fig. 4 shows the result distributions of the best scores from all trials with $5k$ and $20k$ samples respectively. Obviously we observed that, as the number of samples is increased, it is more likely for a best score of one trial itself to be larger for all methods.

So far, we have validated that Geo-OL produces samples of higher score than the other two methods. Let us now investigate the performance in terms of 3D pose.

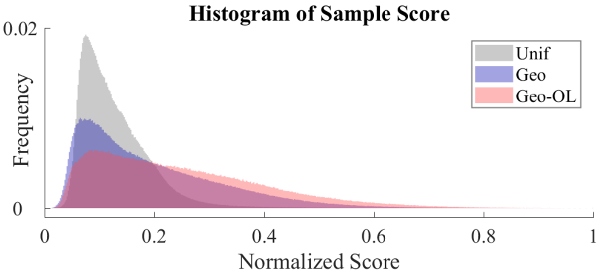


Figure 3. Histogram of Sample Scores for Each Method: This shows an example of sample score distributions for all test instances. (The y-axis is relative frequency.)

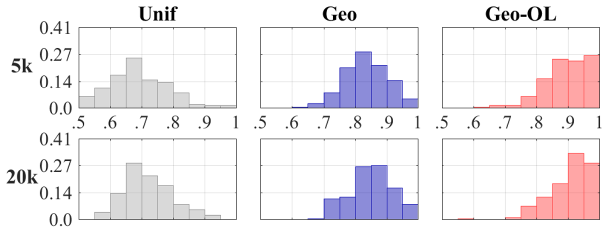


Figure 4. Maximum Score: The maximum sample score distributions for all trial with 5k samples (Up) and 20k samples (Down). (The x-axis is normalized score, and the y-axis is relative frequency.)

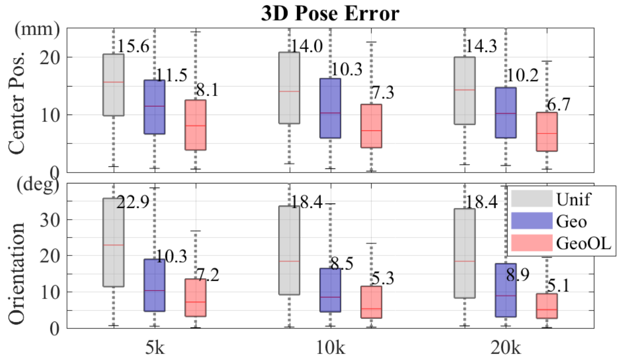


Figure 5. 3D Center Position and Orientation Errors: The numbers indicate the median values. (Note for boxplot: The red center line indicates the median, the box indicates 25 to 75%, and the whiskers indicate 1 to 99% of the data.)

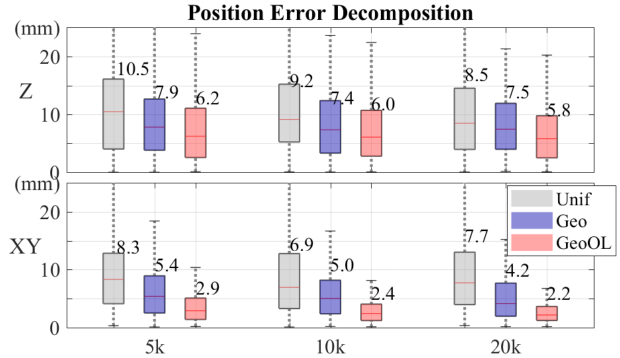


Figure 6. Position Error in Depth (Z) and X-Y: : The numbers indicate the median values.

Pose Accuracy

We evaluated the pose error of the estimates—the aim of this work is to estimate the 3D pose of the tube. In general, it is not easy to define what would be a proper measure for 3D pose error. In our application, the center of the tube can be be the target point of grasping. In addition, the angle around the major axis (θ_3) would not affect much for the grasping task. Therefore, we chose the squared distance of the center estimate and the ground-truth center, and the angular distance between the estimate of major axis and the ground-truth major axis of the target, for the evaluation of pose error.

Fig. 5 shows the boxplot of pose errors over all test images with the varied number of samples. The median values of error are indicated near the corresponding boxes. The statistics are from ten independent trials. We can see the suggested sampling method (red) resulted in the best performance no matter how many samples are used. In addition, the tendency of improvement is clear only for the suggested sampling method (red) when the number of samples is increased from $10k$ to $20k$; Both the uniform samples (gray) and the geometry-constrained sampling (blue) do not show this tendency, meaning that it is unlikely to obtain a significant improvement simply by having more samples. When examining the position error in depth (z) and 2D (x-y) as shown in Fig. 6, we found the main portion of position error is the depth component, which is not surprising for monocular vision. This 3D analysis of uncertainty, including the numbers and the directional tendency, can be used for grasp planning.

6. DISCUSSION AND FUTURE WORK

Some examples of the boundary image from a pose estimate are displayed in Fig. 7. Two test instances are shown, and each one is arranged in three rows: one for uniform sampling (row (a)), another for geometry-constrained sampling (row (b)), and the last for geometry-constrained and learned sampling (row (c)). Each column represents a different trial for the same image instance. Although the projected major axis is well aligned in most cases, we can see that the head and tail of the estimate are often off from the actual image. This sticks out in the cases of uniform sampling, (a)'s. From this we can visually see that the naive method results in poor samples. It is hard to tell that (c) shows better performance than (b) only from this small subset of arbitrary visual results, but we have seen from the quantitative analysis that the combination of geometric constrains and learning improved the probability of sampling better poses.

It will be interesting to extend the suggested method to multi-level progressive learning and sampling. Considering that only the first $3k$ samples are used for learning to obtain the results reported in the previous section, the performance could be improved if it continues to learn from all samples. Also when it becomes confident enough, a local optimization for refinement could be applied for best candidate samples.

There are many potential issues that we could not directly address in the study. For example, it did not cover strategies for adversarial background textures, but assumed that the boundary gradient is informative enough to discriminate the tube from the background. However, the conventional structure-from-motion method can be used to infer 3D information when there are distinctive textures in the image. This study is more for when no strong texture is available in the

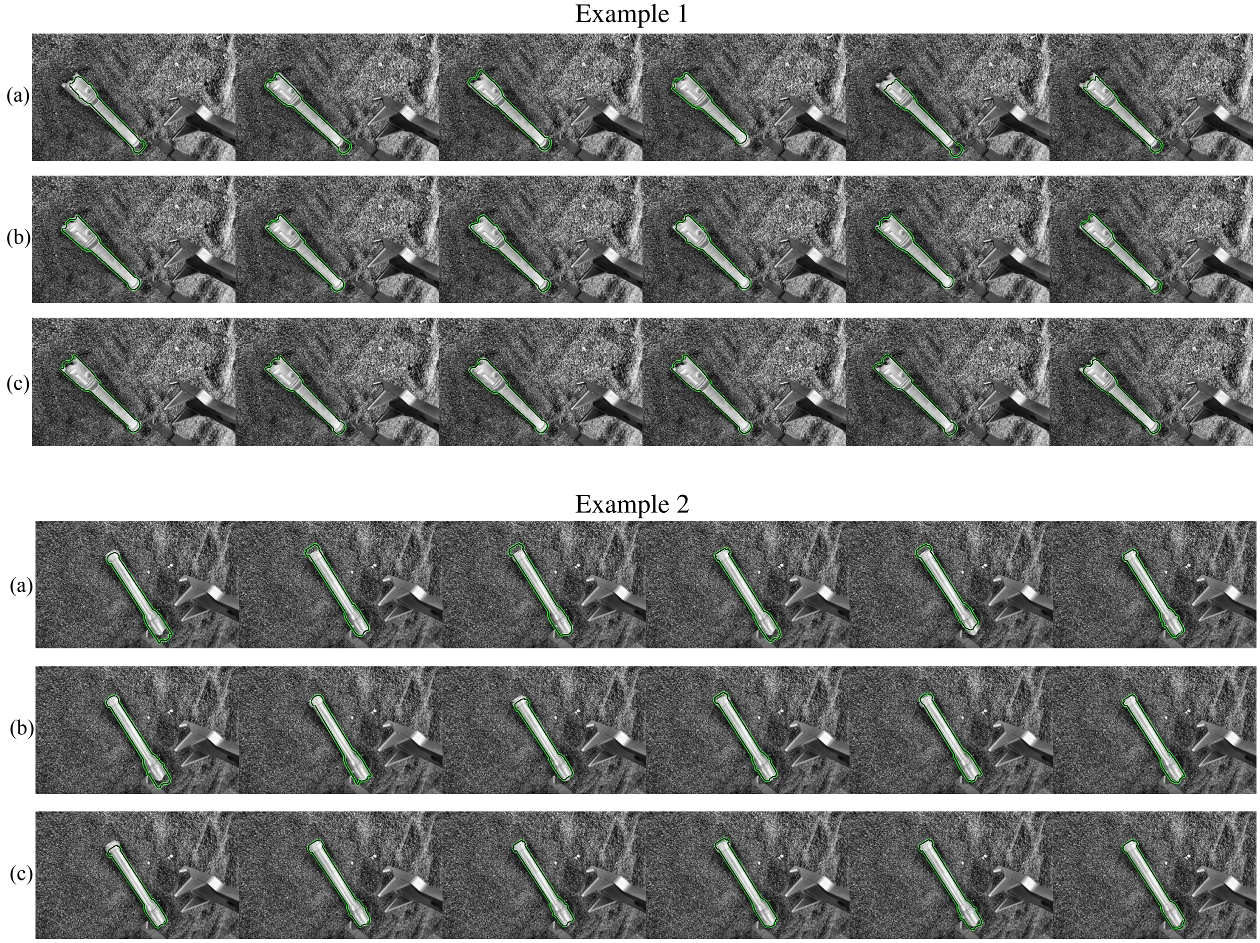


Figure 7. Visualized pose estimates for a subset of all trials using 10k samples for two arbitrarily chosen test examples: (a) Uniform sampling, (b) geometry-constrained sampling (without learning), and (c) geometry-constrained sampling followed by learned sampling. The rendered boundaries are highlighted in green-and-black double lines on the gray-scale images. (Best Viewed in Color)

scene. Also, there could be occlusions of the boundary by dust built over time. For this concern, a future study may exploit the segmentation information from the stereo camera in rendering of the boundary image, in order to reflect any partial visibility of the target.

Lastly, pose estimation of the end-effector is assumed to have done offline, and potentially manually. Since the end-effector pose will be fixed as intended, we see no need to involve an additional estimation problem for the task. In a preliminary study, we were able to test grasping using the end-effector pose obtained by the same way we generated the ground truth poses of the test images.

7. CONCLUSION

Autonomous grasping of sample tubes on the Martian surface would be a difficult task, but one that is desirable for recovering all samples with a reduced amount of time and other resources. In this work, we have presented a study on visual 3D pose estimation of the sample tube using a camera near the end-effector of a robotic arm. Instead of involving any pre-training models, we suggested and demonstrated a sampling-based approach using an image gradient-based score in a coarse-to-fine framework. In order to efficiently sample pose hypotheses, our method encodes geometric priors into the framework and employs an online-learning strategy. Our experiments demonstrated the effectiveness and efficiency of our strategy toward precise Martian sample retrieval.

APPENDIX

We summarize the derivation of $\hat{\theta}_2(\cdot)$ in Eq. (5). Let $\mathbf{z_b} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^\top$ be the unit vector of the tube along the major axis, and $\boldsymbol{\theta}$ be the Euler angle of the tube. We define a rotation $R(\boldsymbol{\theta})$ as follows,

$$R(\boldsymbol{\theta}) = R(\theta_1, 1)R(\theta_2, 2)R(\theta_3, 3). \tag{8}$$

Then the position of the vector's tip can be expressed as, $\boldsymbol{x}_l = \boldsymbol{x}_c + R(\boldsymbol{\theta})\boldsymbol{z}_b$, where $\boldsymbol{x}_c$ is the translation of the center.

Now consider the projection of $\boldsymbol{x}_c$ and $\boldsymbol{x}_l$ onto the image and let us denote them as $\begin{bmatrix} u_c & v_c \end{bmatrix}$ and $\begin{bmatrix} u_l & v_l \end{bmatrix}$ respectively. By the projective relation, $u_c = f_x \cdot x_c / z_c$ and $v_c = f_y \cdot y_c / z_c$, while u_l and v_l can be written as,

$$u_l = \frac{f_x(x_c + \sin\theta_2)}{z_c + \cos\theta_1 \cos\theta_2} \tag{9}$$

$$v_l = \frac{f_y(y_c - \cos\theta_2 \sin\theta_1)}{z_c + \cos\theta_1 \cos\theta_2}. \tag{10}$$

We define the 2D tilt angle ψ as $\tan\psi = (v_l - v_c)/(u_l - u_c)$, which can be rewritten using Eq. (9) and Eq. (10) as,

$$\tan\psi = \frac{-y_c \cos\theta_1 \cos\theta_2 - z_c \cos\theta_2 \sin\theta_1}{-x_c \cos\theta_1 \cos\theta_2 + z_c \sin\theta_2}. \tag{11}$$

From this, given θ_1, $\mathbf{x_c}$, and $\psi = \psi_0$, we can derive θ_2 as a function of the variables. Specifically,

$$\cos^2\theta_2 = \frac{z_c^2}{a^2 + z_c^2}, \tag{12}$$

where $a = x_c \cos\theta_1 - \tan\psi_0(y_c \cos\theta_1 + z_c \sin\theta_1)$. By taking a square root followed by arccos for both sides, we finally have,

$$\hat{\theta}_2 = \arccos\sqrt{\frac{z_c^2}{a^2 + z_c^2}}. \tag{13}$$

Note that in fact there exist four solutions of θ_2 in Eq. (12): $\pm\theta_2'$ and $\pm\theta_2' + \pi$ where θ_2' is a solution such that $0 \leq \theta_2' < \pi/2$. After first computing θ_2' from Eq. 13, we can decide the sign by checking the angle consistency with Eq. 11. Then we may determine whether to add the π term or not based on the Bernoulli distribution.

Acknowledgments

The authors wish to thank all of the members of the Mars Sample Transfer Testbed (MSTT) team for their long hours, hard work, and dedication to testing and improving the testbed and rover capabilities. A special thanks to Thirupathi Srinivasan, Wyatt Ubellacker, and John Mayo for their support with the STARM testbed operations to validate these approaches with actual hardware. Finally, we would like to acknowledge Joe Parrish and Chad Edwards of the JPL Mars Program Formulation Office for funding and directing this work. The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

The information in this paper about Mars sample return is pre-decisional and is provided for planning and discussion purposes only.

References

[1] National Research Council, *Vision and Voyages for Planetary Science in the Decade 2013-2022*. The National Academies Press, 2011.

[2] Mars Program Planning Group, "Summary of the final report," Presentation dated 25 September, 2012. [Online]. Available: https://www.nasa.gov/sites/default/files/files/MPPG-Summary_Report-9-25-12.pdf

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

[4] K. Edelberg, J. Reid, R. McCormick, L. DuCharme, P. Backes, and E. Kulczycki, "Autonomous localization and acquisition of a sample tube for mars sample return," in *AIAA SPACE 2015 Conference and Exposition*, 2015.

[5] 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*, 2017, pp. 1–9.

[6] J. Yang, H. Li, D. Campbell, and Y. Jia, "Go-ICP: A globally optimal solution to 3D ICP point-set registration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 11, pp. 2241–2254, 2016.

[7] D. G. Lowe *et al.*, "Fitting parameterized three-dimensional models to images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 5, pp. 441–450, 1991.

[8] J. Orozco, S. Gong, and T. Xiang, "Head pose classification in crowded scenes." in *BMVC*, vol. 1, 2009, p. 3.

[9] C. Choi and H. I. Christensen, "3D textureless object detection and tracking: An edge-based approach," in *IEEE/RSJ IROS*, 2012, pp. 3877–3884.

[10] S. Hinterstoisser, C. Cagniart, S. Ilic, P. Sturm, N. Navab, P. Fua, and V. Lepetit, "Gradient response maps for real-time detection of textureless objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 876–888, 2012.

[11] M. Zhu, K. G. Derpanis, Y. Yang, S. Brahmbhatt, M. Zhang, C. Phillips, M. Lecce, and K. Daniilidis, "Single image 3D object detection and pose estimation for grasping," in *2014 IEEE International Conference on Robotics and Automation*, 2014, pp. 3936–3943.

[12] E. Brachmann, F. Michel, A. Krull, M. Ying Yang, S. Gumhold *et al.*, "Uncertainty-driven 6D pose estimation of objects and scenes from a single RGB image," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3364–3372.

[13] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A convolutional network for real-time 6-DoF camera relocalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2938–2946.

[14] P. I. Corke, "Visual control of robot manipulators-a review," *Visual servoing*, vol. 7, pp. 1–31, 1993.

[15] D. Kragic and H. I. Christensen, "A framework for visual servoing," in *International Conference on Computer Vision Systems*, 2003, pp. 345–354.

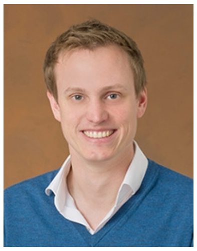
[16] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *Journal of Machine Learning Research*, vol. 17, no. 39, pp. 1–40, 2016.

[17] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[18] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Computer vision, graphics, and image processing*, vol. 39, no. 3, pp. 355–368, 1987.

BIOGRAPHY

Bhoram Lee *received her B.S. degree in mechanical and aerospace engineering in 2005 and a M.S. dgree in aerospace engineering in 2007 from Seoul National University, Korea. She is currently a PhD student at GRASP Lab, University of Pennsylvania. Before coming to Penn, she worked at Samsung Advanced Institute of Technology as a researcher. Her previous research experience includes visual localization of UAVs, sensor fusion for pose estimation, and mobile user interactions. Her current interest includes 3D vision, machine learning, and general robotics, and is focused on online learning for robot vision.*



Renaud Detry *is a research scientist at NASA JPL, and a visiting researcher in the Systems and Modeling Group (University of Liège, Belgium) and in the Computer Vision and Active Perception lab (KTH Kungliga Tekniska Högskolan, Stockholm, Sweden). He earned an engineering degree at the University of Liège in 2006, and a Ph.D. in robot learning from the same university in 2010. He subsequently earned Junior Researcher starting grants from the Belgian FNRS and from the Swedish VR. He alternated between KTH Stockholm and the University of Liège between 2010 and 2015, before joining the Robotics and Mobility Section at JPL in 2016. His research interests are in perception for manipulation, robot grasping, computer vision and machine learning.*

Jasmine Moreno *received her B.S. degree in Electrical Engineering in 2017 from the University of California, Riverside. Jasmine has played a significant role in UCR's Institute of Electrical and Electronic Engineers (IEEE) chapter, for which she currently serves as Chair and has previously served as Vice-Chair and Treasurer. Her efforts helped advance the organization with new activities such as ECE Day and LabVIEW Academy. She is currently a M.S student at the University of California, Los Angeles. She has interned at JPL for the past two summers. This past summer, she worked on the motion planning for the MSTT vision module.*

Daniel D. Lee *is the UPS Foundation Chair Professor in the School of Engineering and Applied Science, and the director of the GRASP Laboratory at the University of Pennsylvania. He received his B.A. summa cum laude in Physics from Harvard University and his Ph.D. in Condensed Matter Physics from the Massachusetts Institute of Technology in 1995. Before coming to Penn, he was a researcher at AT&T and Lucent Bell Laboratories in the Theoretical Physics and Biological Computation departments. He is a Fellow of the IEEE and AAAI and has received the National Science Foundation CAREER award and the University of Pennsylvania Lindback award for distinguished teaching. His group focuses on understanding general computational principles in biological systems, and on applying that knowledge to build autonomous systems.*

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 13 years.*