Vulcan Centaur: towards end-to-end real-time perception in lunar rovers

J. de Curtó i Diáz and R. A. Duvall.
Iris Lunar Rover. Carnegie Mellon.

We introduce a new real-time pipeline for Simultaneous Localization and Mapping (SLAM) and Visual Inertial Odometry (VIO) in the context of planetary rovers. We leverage prior information of the location of the lander to propose an object-level SLAM approach that optimizes pose and shape of the lander together with camera trajectories of the rover. As a further refinement step, we propose to use techniques of interpolation between adjacent temporal samples; videlicet synthesizing non-existing images to improve the overall accuracy of the system. The experiments are conducted in the context of the Iris Lunar Rover, a nano-rover that will be deployed in lunar terrain in 2021 as the flagship of Carnegie Mellon, being the first unmanned rover of America to be on the Moon.

CCS Concepts:
- Robotics → Perception.

Additional Key Words and Phrases: Robotics, Lunar Rover, Perception, SLAM, VIO, Segmentation.

1 Introduction

Our aim is to present a novel pipeline to deploy state-of-the-art DL techniques in planetary rovers. With the advent of a new wave of planetary exploration missions, the need to call on generalizable perception and control systems that can operate autonomously in other worlds will become ubiquitous in the coming years.

2 Overall System

Following the design principles and the perception pipeline proposed in [Allan et al. 2019] in the context of the NASA Mission Resource Prospector, we put forward an improved technique for Visual Odometry (VIO) that could be exploited in a rover of the same characteristics. Although at the present time data from the Moon is scarce, there are already some open datasets available in analogue environments such as the POLAR Stereo Dataset [Wong et al. 2017] that includes stereo pairs and LiDAR information or [Vayugundla et al. 2018], that contains IMU, stereo pairs and odometry plus some additional localization data, all obtained on Mount Etna. Specifically for the task of semantic segmentation, Kaggle provides images from a rendered environment of the Moon and masks. More recently, as a benchmark for tasks of Computer Vision in the context of space exploration, a dataset containing PNG images and positioning information from the mission Chang E-4 to the Moon has been released [de Curtó and Duvall 2020], the data from CE4 consists on post-processed original files from the mission Chang’E1.

Our specific sensor suite, that will be on-board the Iris Lunar Rover [de Curtó and Duvall 2020], a project led by Carnegie Mellon that will deploy a four pound rover into the surface of the Moon by 2021 and that will be the first unmanned rover of America to explore the surface of the Moon, consists on IMU, two high-fidelity cameras and odometry sensors. Furthermore, it also has a UWB module [Alarifi et al. 2016; Ledergerber et al. 2015; Mueller et al. 2015; Xu et al. 2020] on-board to localize the rover with respect to the lander.

3 SLAM/VIO

Simultaneous Localization and Mapping (SLAM) and Visual Inertial Odometry (VIO) are defined as a function that transform raw data from the sensors into a distribution over the states of the robot. SLAM and VIO [Schneider et al. 2018; Usenko et al. 2019] have been for decades unparalleled problems in robot perception and state
estimation. Although typical dense SLAM systems are not differentiable, new approaches to solve this problem propose gradient-based learning over computational graphs to go all the way from 3D maps to 2D pixels [Murthy et al. 2020].

The first task to tackle in geometric computer vision, being SLAM [Engel et al. 2017, 2014; Newcombe et al. 2011], Structure-from-Motion (SfM) [Agarwal et al. 2009; Bloesch et al. 2018; Graham and Novotny 2020; Snively et al. 2006; Tateno et al. 2017; Z. Teed 2020], camera calibration or image matching, is to extract interest points [DeTone et al. 2018; Ono et al. 2018] from still images. We can define interest points as 2D specific locations in a given sample which can be considered stable and repeatable along different ambient conditions and viewpoints. The techniques used to traditionally attack this problem pertain to Multiple View Geometry [Hartley and Zisserman 2003], a subfield of mathematics that sets forth theorems and algorithms built on the assumption that those interest points can indeed be reliably extracted and matched across overlapping frames. Nonetheless, real-world computer vision operates on raw images that are far from the idealized conditions assumed in the proposed theory. Trading traditional modules with learning representations have lately been proven to be incredibly effective [DeTone et al. 2018; Tang and Tan 2019; Yang et al. 2020] as a way to bridge the gap between the conditions that we face in the real world and the assumptions made to design the algorithms. Plentiful of approaches also explore unsupervised learning of depth and ego-motion [Godard et al. 2017; Yin and Shi 2018; Zhou et al. 2017].

State-of-the-art approaches also deal with related problems such as object-level SLAM, that is, a system capable of optimizing object poses and shapes together with camera trajectory [McCormac et al. 2018; Sucar et al. 2020; Sünderhauf et al. 2017]. Although a SLAM system capable of incrementally mapping multi-object scenes seems not related to our task, its importance is revealed when we understand the fact that in many occasions the rover will localize itself with respect to the lander, which location is known; therefore a SLAM solution capable of optimizing the pose and shape of the lander along camera trajectory of the rover, would be distinctly adequate. With respect to this, we have to bear in mind that the principal technique that the rover will be using on-board to localize itself will be the UWB module [Xu et al. 2020], that will indeed use the lander as a way-station for data communication. The reason for this is that critical weight and power can be hugely saved using RF for communication and state estimation. Thus, SLAM and VIO computation will be done on-ground. Using the same philosophy, it seems natural also to rely on a technique that will jointly optimize pose and shape of the lander together with camera trajectories.

4 Shape and Pose of the Lander

We assume here that we have a segmentation mask of the lander that in our specific case is obtained by the use of semantic segmentation [Arbeláez et al. 2011, 2014; Chen et al. 2018a, 2017a,b, 2018b; Hariharan et al. 2015; He et al. 2017; Mostajabi et al. 2015; Yu and Koltun 2016; Zhou et al. 2017]. On some of these approaches, the segmentation process is guided by the use of a prior object detector [Girshick 2015; Huang et al. 2017; Liu et al. 2016; Redmon et al. 2016; Redmon and Farhadi 2017; Ren et al. 2015]. Specifically, we finetune our model building on DilatedResnet-101 [Zhao et al. 2017; Zhou et al. 2018] and UperNet-101 [Lin et al. 2017; Xiao et al. 2018] trained on ADE20K [Zhou et al. 2017]. Some examples of the mask given by our segmenter can be observed in Figure 2. To infer the shape and pose we will leverage existing techniques [Sarac et al. 2020] that given a depth image, full shape and pose is determined. These techniques normally address multi-object categories; where a previous classification step and object observation is necessary, however our approach is somewhat simpler in the sense that the only object under consideration will be the lander per se.

5 Temporal Interpolation between Subsequent Samples

In the absence of continuous data between adjacent temporal samples given by the camera and to mitigate the effects that this will incur in the algorithms used to localize the rover, we propose to adopt techniques from video frame interpolation. Although signal-breaking breakthroughs have been achieved by the use of recent deep convolutional neural networks, the quality of the resulting samples is often dubious due to object motion or occlusions. The main aim here is to synthesize non-existent frames in-between original samples to improve accuracy in the proposed VIO/SLAM approaches. Specifically for this purpose, we build on a recent depth-aware flow projection layer that achieves compelling upshots to synthesize intermediate sequences [Bao et al. 2019].

References

S. Agarwal, N. Snively, I. Simon, S. M. Setz, and R. Szeliski. 2009. Building Rome on a Day. ICCV (2009).
A. Alansari, A. Al-Salman, M. Alsaieh, A. Alnafessah, S. Al-Hadrami, M. A. Al-Annari, and H. S. Al-Khalifa. 2016. Ultra Wideband Indoor Positioning Technology: Analysis and Recent Advances. Sensors (2016).
M. Allan, U. Wong, P. M. Furlong, A. Rogg, S. McMichael, T. Welsh, I. Chen, S. Peters, B. Gerkey, M. Quigley, M. Shirley, M. Deans, H. Cannon, and T. Fong. 2019. Planetary Rover Simulation for Lunar Exploration Missions. IEEE Aerospace Conference (2019).
P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik. 2011. Contour Detection and Hierarchical Image Segmentation. TPAMI (2011).
P. Arbeláez, J. Font-Tuset, J. T. Barron, F. Marques, and J. Malik. 2014. Multiscale Combinatorial Grouping. CVPR (2014).
W. Bao, W.-S. Lai, C. Ma, X. Zhang, Z. Gao, and M.-H. Yang. 2019. Depth-Aware Video Frame Interpolation. CVPR (2019).
M. Bloesch, J. Czarnowski, R. Clark, S. Leutenegger, and A. Davison. 2018. CodeSLAM - Learning a Compact, Optimizable Representation for Dense Visual SLAM. CVPR (2018).
L. Chen, A. Hermans, G. Papandreou, F. Schroff, P. Wang, and H. Adam. 2018a. MaskLab: Instance segmentation by refining object detection with semantic and direction features. CVPR (2018).
L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. 2017a. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. TPAMI (2017).
L. Chen, G. Papandreou, F. Schroff, and H. Adam. 2017b. Refining Atrous Convolution for Semantic Image Segmentation. arXiv:1706.05587 (2017).
L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. 2018b. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. ECCV (2018).
J. de Curtó and R. Duvall. 2020. Cycle-consistent Generative Adversarial Networks for Neural Style Transfer using data from Chang-E-4. arXiv:2011.11627 (2020).
D. DeTone, T. Malisiewicz, and A. Rabinovich. 2018. SuperPoint: Self-Supervised Interest Point Detection and Description. CVPR (2018).
J. Engel, V. Koltun, and D. Cremers. 2017. Direct Sparse Odometry. TPAMI (2017).
J. Engel, T. Schöps, and D. Cremers. 2014. LSD-SLAM: Large-Scale Direct Monocular SLAM. ECCV (2014).
R. Girshick. 2015. FAST R-CNN. ICCV (2015).
C. Godard, O. Mac Aodha, and G. Brostow. 2017. Unsupervised Monocular Depth Estimation with Left-Right Consistency. CVPR (2017).
Figure 2. Segmentation of the Lander. Left: Image from CE4 [de Curtó and Duvall 2020]. In particular we are using color images from the panoramic camera of the rover of the mission to the Moon Chang’E-4. Middle: Generated mask given by model DilatedResNet-101 [Zhao et al. 2017; Zhou et al. 2018]. Right: Generated mask given by model UperNet-101 [Lin et al. 2017; Xiao et al. 2018].