AIAA SciTech Forum
6-10 January 2020, Orlando, FL
AIAA Scitech 2020 Forum

10.2514/6.2020-1377

Incremental Reward Learning for Robotic Exploration

Aastha Acharya[*][†], Shohei Wakayama[‡], Brian Hynek[§][¶], Bradley Hayes[‖], and Nisar Ahmed [**]
University of Colorado Boulder, Boulder, CO USA, 80309

Robotic exploration of other planets in the solar system continues to be a slow and lengthy process that requires close supervision by human operators on Earth. This is further exacerbated by the unavoidable communication delays and limitations, and ultimately results in long durations of time where the rover idly awaits commands from ground operators. To accelerate the process and increase rover efficiency, we propose incrementally training the rover throughout its traversal of a planetary surface to map its in situ observations to appropriate decisions that maximize the scientific return of the mission. By using basic classification and novelty detection techniques, the rover is able to process its surroundings to recognize new and interesting areas. Then, using a reward metrics based proxy between the observation space and decision space, it is shown that by allowing the rover to incrementally update its own estimated distribution of high-value science targets (i.e. estimated reward map), the capabilities of the rover can be increased to the point where it can be trusted to make informed, autonomous decision. Our proposed framework provides a conservative approach to autonomy that can fit with the current state-of-the-art and maximize the rover's capability to fully utilize captured in situ information. Furthermore, the framework also allows the rover to propose changes in its current plans to maximize the scientific gain while still completing its intended mission and maintaining a high level of safety.

I. Introduction

The rate of planetary exploration is very slow as it continues to require constant human supervision and is heavily impacted by communication delays as well as time- and bandwidth-limited communication windows. Given the time and resources needed to deploy exploratory robots to other planets in our solar system, the primary concern is rightfully the safety of these systems so that they can successfully complete their very specific scientific goals. This risk averse outlook brings advantages, such as extended longevity of missions resulting in prolonged exploration opportunities. However, there are also significant drawbacks, such as notable stretches of time where rovers remain idle while awaiting commands and decisions from human operators. As a result, the rovers that are currently exploring surfaces such as Mars can be viewed as suite of instruments that are utilized based on direct human command only. Ideally, to maximize the productivity of these rovers and expedite the currently laborious process of human decision-making, it would be desirable to equip them with autonomous decision making capabilities to opportunistically and safely explore new scientifically interesting planetary regions.

A typical Martian exploration mission centers around exploring a region that mission scientists deem to be of high value based on orbital data. These may include areas containing interesting rock layers or which have biologically significant chemicals, and their selection requires extensive preliminary work and field studies by the scientists. Once a scientifically valuable region has been chosen, further decisions are made regarding a safe landing site and specific areas of interest to be studied using the on-board instruments by the rover. As such, the on-surface operations of a Martian rover can be divided into two phases: traversal phase and scientific phase [1]. During the traversal phase, the rover travels between points of scientific interest along a planned path; during the scientific phase, it deploys instruments to study the sites designated by mission designers. Both of these phases require detailed planning and supervision by mission operators on Earth and are done at a very slow pace in order to be safety conscious. Therefore, to accelerate the rate of planetary exploration, autonomous decision making capabilities could be added during both of these phases so

[*] Graduate Research Assistant, Ann and H.J. Smead Aerospace Engineering Sciences, University of Colorado Boulder, AIAA Student Member
[†] Draper Fellow, The Charles Stark Draper Laboratory, Inc.
[‡] Graduate Research Assistant, Ann and H.J. Smead Aerospace Engineering Sciences, University of Colorado Boulder, AIAA Student Member
[§] Associate Professor, Department of Geological Sciences, University of Colorado, Boulder.
[¶] Research Associate, Laboratory for Atmospheric and Space Physics, University of Colorado, Boulder.
[‖] Assistant Professor, Department of Computer Science, University of Colorado Boulder.
[**] Assistant Professor, Ann and H.J. Smead Aerospace Engineering Sciences, University of Colorado Boulder, AIAA Member.

that the rover can safely explore regions of value along its traversal path, as well as make decisions regarding how to alter its path and which instruments to deploy based on its observations. To reach this level of autonomy, however, a conservative approach has to be taken that allows mapping from the in situ observations to decisions. In this paper, one such mapping from observations to such decisions is developed in the form of a science target reward map that can be learned and updated online by the rover, and we show how this can be used by the rover to propose alterations in its path to collect information of high value (i.e. high reward).

In our proposed method, the rover is initially equipped with estimated reward map which is generated using low resolution multispectral images obtained from satellites and encodes information regarding scientific value of a planetary surface region. As the rover begins its traversal of the planetary surface, it is able to gather detailed, high resolution observations using on-board multispectral cameras. These images are further abstracted into features, allowing for their classification into mineralogical signatures using the on-board classification system. The perceived mineralogical signatures are then directly converted into reward values based on the level of interest from the scientists for specific minerals. As a result, the rover is able to iteratively update its estimated reward map to add more details and correct any a priori estimations of the reward value, ultimately resulting in a reward map that can be used as a heuristic during its planning phase. In this paper, we focus primarily on the feature-to-reward encoding methodologies and leave the reward-to-decision mapping for future work. The feature-to-reward portion is particularly important because we anticipate the rover encountering features that it has not been previously trained with, which then have to be associated with existing or new reward values. To address this concern, this paper's main contribution is two-fold: 1) online reward shaping and learning using in situ observations and (if required) querying of a remote human operator and 2) utilization of novelty detection techniques online to recognize new features not initially in rover's training data. These additions enable an intelligent robotic autonomy that can:

- update the estimated reward map autonomously using on-board observations (thus fully utilizing the information captured in situ),
- detect new, previously unseen features in the traverse region autonomously (thus easing the human operator's workload of analyzing large volumes of data),
- prioritize data that is sent down to human operators (thus addressing the bandwidth limitation issue while downlinking data to Earth),
- perform incremental training on-board (thus constructively utilizing idle times of the rover and addressing bandwidth limitation issue that would arise if uplinking updated reward maps back to the rover from ground control).

To demonstrate our methodology, we simulate a rover that has just landed in a region of interest on a Mars analog site on Earth and is navigating to an area of high scientific value via predesignated waypoints. It is assumed that the ground operators on Earth have created an initial traversal plan for the rover based on mineralogical map generated using low resolution multispectral satellite imagery, and that the rover is equipped with coarse-grained reward map, feature-to-reward classification system, and multispectral imaging-based novelty detector. As the rover begins its traversal phase, the cameras on-board produce detailed multispectral images which are analyzed and used to update the rover's reward map. Furthermore, the rover is able to use its novelty detection ability to mark out any features not currently represented as known, which are then given priority when querying the operators on Earth. The human operator team can then provide feedback on the reward value, and the rover is able to perform incremental training of its feature-to-reward map using this information to perform supervised learning on-board, thus updating its estimated reward map online. Additionally, we maintain human-in-the-loop to provide mission team with the desired level of supervision over the duration of the mission. The human operators also have the ability to provide additional training data to augment or improve the rover's reward map at any time. We ultimately show that using the continuously updated reward map and information regarding its existing path, the rover can propose adding additional waypoints along its path in order to collect more scientifically valuable information while still maintaining its primary mission.

II. Background and Related Work

Any planetary exploration mission is a result of extensive amount of preliminary work by teams of scientists and engineers. Along with systematic decisions pertaining to design and deployment of the rover, large amount of study is dedicated to understanding the destination planetary surface and selecting a proper target for exploration and data gathering. However, scientists typically only have access to low resolution data captured by orbital instruments that may not necessarily be representative of the actual surface conditions. This is particularly true for Martian environment where much of the surface is covered in dust [2]. Hence, initial groundwork includes calibration of the satellite observed data with on-ground measurements, and these are usually performed on Mars analog sites on Earth. Baldrige et al [3] describe one such study of the Badwater Basin region where they compare the multispectral images gathered using MODIS/ASTER Airborne Simulator with the quantitative observations from ground to analyze the agreement between these two sources. While the initial calibration might still be necessary for the initial target selection, this process can be automated by allowing the rover to perform the calibration during its traversal phase by constantly updating reward map based on the observation. Since we propose to use supervised learning to perform the classification from observations to the reward map, the learning can continue even as the rover is performing its operations. This enables incrementally improved classification system for the rover, which will provide an improved basis for the decision making process. Once enough information gathering and training has been performed, the updated reward map can allow for more efficient and autonomous decision making by the rover in the future. An additional benefit of framing the feature-to-reward as a classification task is that the concept of transfer learning [4] can be used. In particular, the prospect of building and sharing feature and reward spaces amongst the planetary exploration robots is appealing as more systems are deployed onto the various planetary surfaces. This also allows for information across planetary surfaces to be shared and compared using the same metrics.

We recognize the importance to maintain human-in-the-loop even as we increase the autonomous capabilities of the rover in order to be risk averse and safety conscious. Therefore, when possible, we propose maintaining a collaborative human-robot framework where the rover is constantly querying the human for any important decision making process while also easing the workload of the operators by performing analysis on-board. This collaborative human-robot teaming to jointly perform a mission has been researched extensively within the field of Human Robot Interaction (HRI). For applications where robots are teleoperated by humans remotely, there exists a concept of using human operators as resources, both for their expertise as well to provide information that otherwise wouldn't be available to the rover. Fong [5] presented the idea of "collaborative control" where humans and robots are able to jointly perform a mission as they exchange information. More recent work by Burks et al. [6] focused on the concept of using "humans as sensors", where the humans can be queried by the robot to obtain useful information about the environment. Another research area that is receiving massive traction is the concept of using human experts to train the robot's reward models and policies within the framework of reinforcement learning. In particular, inverse reinforcement learning [7], which includes techniques such as apprenticeship learning [8] and learning from demonstration [9], is a field dedicated to learning from humans. In our method, we give human operators the role of an oracle so that they can corroborate or override the rover estimated reward map at any time. We expect high interaction with human operators initially as new and detailed observations are received, but expect human intervention to slow down as rover's classifier and novelty detector become well-versed in the region. However, constant querying of humans is still maintained as the rover has to receive the human's approval before making any changes to the current plans. One possible extension of this work can be to explore unreliable or delayed feedback provided by the human by having the rover perform further analysis on received information.

Communication limitation is a limiting factor in space robotics, but it is also studied in the context of terrestrial robotic applications. In particular, there have been studies performed on when and what to communicate in joint human-robot teams. Kaupp et al. [10] addressed this with the theory of Value-of-Information by providing a probabilistic framework to weigh the expected benefit of information received from the humans against the cost of obtaining it. Performing similar analysis in the domain of space robotics presents an interesting extension of this work and is left for future consideration.

The concept of novelty detection to detect outliers has been studied for many applications including fault detection, astronomy catalogues, and medical imagery [11]. It is also heavily studied in the field of robotics to recognize anomalies in the observation space of the robot. Methods of novelty detection for robots has included highly complex, biologically inspired method using Hopfield networks [12], neural network based techniques [13][14], and Principal Component Analysis (PCA) methods [15]. More recent work by Özbilge [16] showed the usage of recurrent neural networks (RNN) for online novelty detection. For our purposes, it is not practical to use highly complex or deep learning based online novelty detection techniques due to their computational expense, and we choose to use one-class SVM that is computationally efficient in performing our task.

To ground our proposed method within the current state-of-the-art, it is important to consider the advances made towards autonomy for planetary exploration. Arguably one of the most advanced autonomous system currently on-board a Martian rover is AEGIS (Autonomous Exploration for Gathering Increased Science) which allows autonomous deployment of the ChemCam instrument based on target selection on the Curiosity rover. The flow of operations for this system requires the scientists to specify the features of interest, such as large rock with high reflectance [17]. The rover then analyzes images captured by PanCam and NavCam during its normal operations to select a target at which a laser is fired by the ChemCam to study the vaporized material. In our method, we seek to bridge the gap between feature representation and decision making so that the scientists do not have to specify in advance what would be the desired features and the rover will be able to make this connection on its own. In addition, our method still provides scientists with the flexibility to alter their ranking by simply changing the reward values associated with the observed features.

Martian rovers are also equipped with AutoNav based on GESTALT (Grid-based Estimation of Surface Traversability Applied to Local Terrain) algorithm [18]. The stereoscopic images captured by the on-board cameras are used to create a goodness map based on the geometric information of the local terrain. The values in the goodness map denotes the ease of traversability, with high values indicating easily traversable terrain and low values indicating dangerous areas. This map is then used by the path planning algorithm to create the safest path for the rover. Our end vision is also formulated on the principle of mapping reward values to decisions, but we're further interested in expanding the decision spaces to include scientific actions based on low level textural, mineralogical, and chemical details on the observed terrain. Furthermore, before reaching this reward-to-decision encoding, we focus on the preliminary steps required to learn and convert new features to specific reward values.

III. Problem Setup

As a demonstrative example, we simulate a rover traversing Atacama Desert, a Mars analog site on Earth that is used extensively for many field robotic experiments [19]. To illustrate the algorithmic abilities, we disregard the communication delays for now and simply focus on the reward learning framework. We set up the simulation by assuming that only a satellite produced view of the desert is available. Analogous to TES or THEMIS instruments on Mars orbiters which have spectral resolution of 3x5 km/pixel or 100x100 m/pixel resolution respectively [3], high resolution multispectral images from Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) instrument of Terra satellite are collected from the USGS database [20] and downsampled to a lower resolution. Particularly, we consider the five available bands available in thermal infrared ($8 - 12$ um) from the Thermal Infrared Radiometer (TIR) of ASTER. Then, the ASTER Spectral Library [21] is used to produce the mineralogical map of the region using ENVI (Exelis Visual Information Solutions, Boulder, Colorado).

The data obtained from ASTER is used to generate two forms of maps: high resolution ground truth at 90 m/pixel that contains five mineralogical signatures (Fig. 1), and low resolution image at 500 m/pixel that contains four mineralogical signatures (Fig. 2). Following the generation of mineralogical maps for both sets of images, areas containing minerals of interests are marked with high reward values and the rest are ranked accordingly. The low resolution version serves as the initial feature-to-reward reward estimation map for the rover, where the features are the thermal emission values at each of the five bands. This information is used in a supervised learning process to train a classification system that is used by the rover throughout its operations. The ground truth version is assumed to capture the human operator's response once they are provided with full view of the terrain, and is therefore referred to throughout the operation for comparison and/or querying purposes.

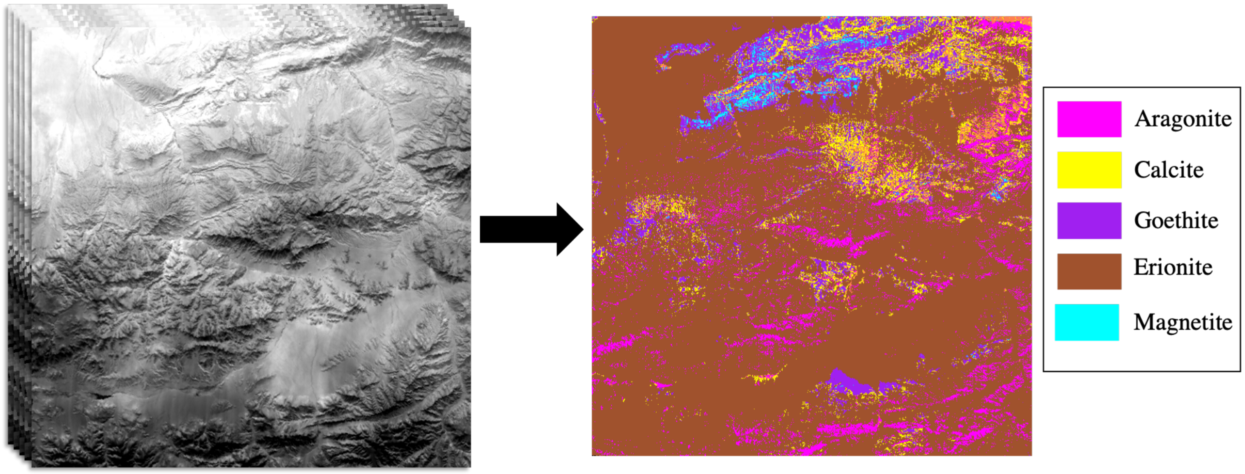

Fig. 1 High resolution (ground truth) multispectral images obtained from ASTER and their mineralogical signatures

There are five primary minerals that are recognized in Atacama Desert from the multispectral images as shown in

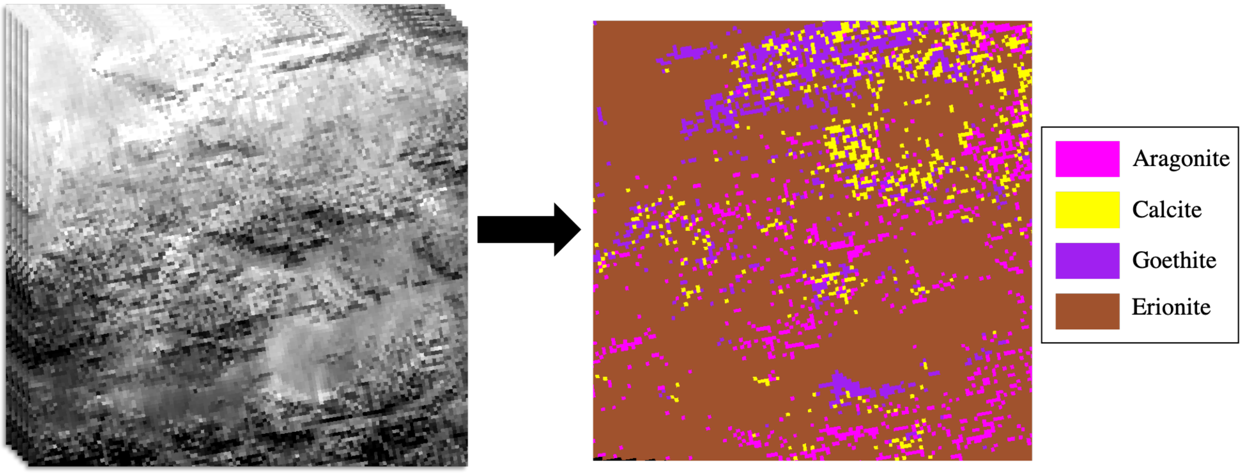

Fig. 2 Downsampled multispectral images from ASTER and their mineralogical signatures

Fig. 1 above. In generating the low resolution image, we only target four of the five minerals that we know are either widely available on the planetary surface or have been recognized by scientists as valuable signatures for exploration purposes. Goethite and calcite are selected as minerals of interest as both of these are considered to be representative of areas that may have contained water. To encode the importance of these regions as areas of high information, they are assigned high reward values. Areas that are not of much interest, such as those containing signatures of erionite or aragonite are assigned neutral reward values. The mineral that is not encoded in the low resolution map, magnetite, is considered to be the most valuable out of all the available mineralogical signatures and is left for the rover to discover on its own. The resulting details on the development of the reward maps as well as its usage to generate the initial path plan is described in Section IV below.

As the rover begins its traversal, it uses its sensors, notably its multispectral camera, to observe the terrain in close proximity. The field of view for the sensor is set at 360 degrees and 150 pixels around the rover, and captures data is high resolution. This may introduce features that are previously unseen by the rover and not included in its reward map. The rover performs two operations for any in situ observations: prediction of the associated rewards and novelty detection of the observed features. Reward predictions on the high resolution data are used to fill in fine-grained details in the rover's estimated reward map. Results from the novelty detection are considered particularly useful as they are used to trigger querying of the human operators. It is assumed that the human operators then use their detailed mineralogical maps and/or expertise to assign a reward value to the newly observed features, which is used by the rover for training on the novel region. Furthermore, the human operator can also access the rover's estimated reward map as well as the observation history at any time to manually verify and/or correct any reward values, and this additional information is also used by the rover for training purposes. These two methods of updating and correcting the reward map allow the rover to achieve incremental learning to better classify its observation and eventually make decisions which accordingly maximize science return.

IV. Methodology

A. Initial Training

The initial feature-to-reward based classification system for the rover is generated using the low resolution images. This provides a supervised learning training dataset which maps the features (thermal emission information from the five bands of multispectral image) to discrete reward values (labels). Our training set consists of N observations (pixels) represented in the feature space $X = \{x_i \in \mathfrak{R}^5; i = 1, \ldots, N\}$ which denotes the 5 available bands. The features are mapped onto science reward values represented as classes $Y = \{y_i \in \mathfrak{R}^{11}; i = 1, \ldots, N\}$ and are outputted from the classifier as a one-hot vector. Note that although the possible number of classes for rewards is set at 11 ranging from $\{0, 1, \ldots, 9, 10\}$, the initial training data will only contain features corresponding to 4 of the rewards ($\{0, 1, 3, 5\}$) associated with 4 of the minerals that are available at that time. Adding this buffer of extra classes at the beginning allows any newly discovered features to map onto these undefined classes.

Table 1 Mineral to Reward Mapping

Mineral	Reward
Erionite	0
Aragonite	1
Calcite	3
Goethite	5
Magnetite	9

After converting the collected satellite data into mineralogical signatures, each of the minerals are assigned a reward value based on their deemed worth by the scientists. Any number of minerals can be mapped to a single reward value, but we show a one-to-one mapping of the minerals to reward for demonstrative purposes. The minerals and their associated reward values are reproduced in Table 1 and are ranked according to properties such as their evaporative nature and abundance. The table also shows the associated reward value for magnetite, which is not part of the initial training set for the rover but will be learned later on. This mineral is assigned the highest reward value as its presence is not initially known by the scientists and because it provides good indication of presence of water on the surface. The resulting reward maps, both for the ground truth data and for the rover's partial view for initial training are reproduced in Fig. 3.

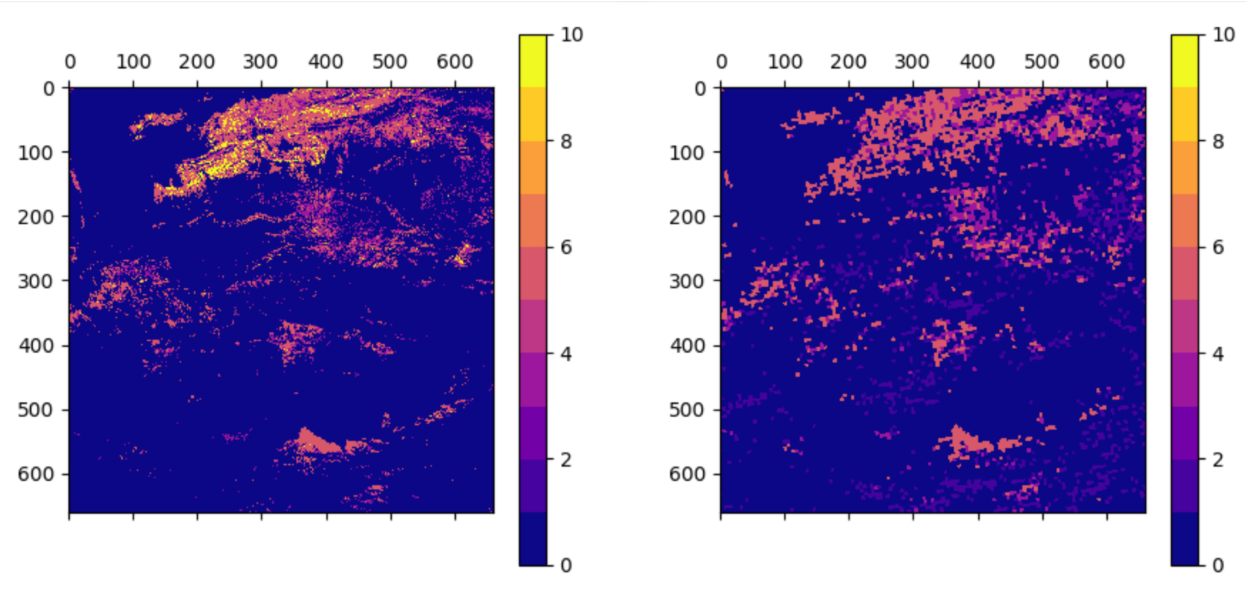

Fig. 3 Ground truth reward map (left) and rover's initial partial view reward map (right)

Using this information, support vector machine (SVM) classifier is trained to map the five-band thermal emission values to the reward classes. We choose SVM as our classification method because it adapts well to unbalanced classes, provides a unique solution due to its convex optimization function, can handle high dimensional feature spaces, and doesn't require extremely large training data. Furthermore, incremental SVM training has been shown to be practical method of online learning [22]. Using the satellite image as well the reward map, we frame the initial SVM training as a supervised learning problem. Since this is a multi-class classification problem, we train a linear one-versus-all classifier using stochastic gradient descent with L2 regularization. This trained, coarse-grained model is then used to generate reward predictions for all future rover observations, and it is updated incrementally as new and/or detailed feature instances corresponding to new and/or existing rewards are detected at higher resolution.

B. Path Planning

After the reward map is generated based on the multispectral image and the associated mineralogical information, we select the target area to be explored. For our purposes, we choose the region containing the densest areas of goethite as our final goal, and select four additional waypoints along the way at regions containing signatures of geothite and calcite. It is assumed that the landing site is the point farthest away from the goal location in the map. Using this information, we generate a list of coordinates (shown in Table 2) that is used to plan a path for the rover. Given the starting and the ending points, we use the A* algorithm [23] to perform path planning. We use the reward map as the heuristic for the planner to allow the rover to maximize the total reward that it collects along the path. It is assumed that the rover follows this path for the duration of the mission unless it recognizes areas of very high reward values and proposes a new waypoint to the human operators. In the scenario where a new waypoint is approved, A* is again implemented with the new starting and ending goal points, which allows the original goal point and intermediate waypoints to be reached unless directly altered by the human operator. An example initial path plan for the rover is shown in Fig. 4 (right) alongside the low resolution mineralogical map of the area (left). Note that the waypoint directly corresponds to areas containing signatures of geothite (purple) and calcite (yellow).

Table 2 Waypoints for Path Planning

Region	X-Coordinate	Y-Coordinate
Landing Site	10	650
Intermediate Point 1	390	550
Intermediate Point 2	375	385
Intermediate Point 3	90	300
Intermediate Point 4	400	200
Goal	430	30

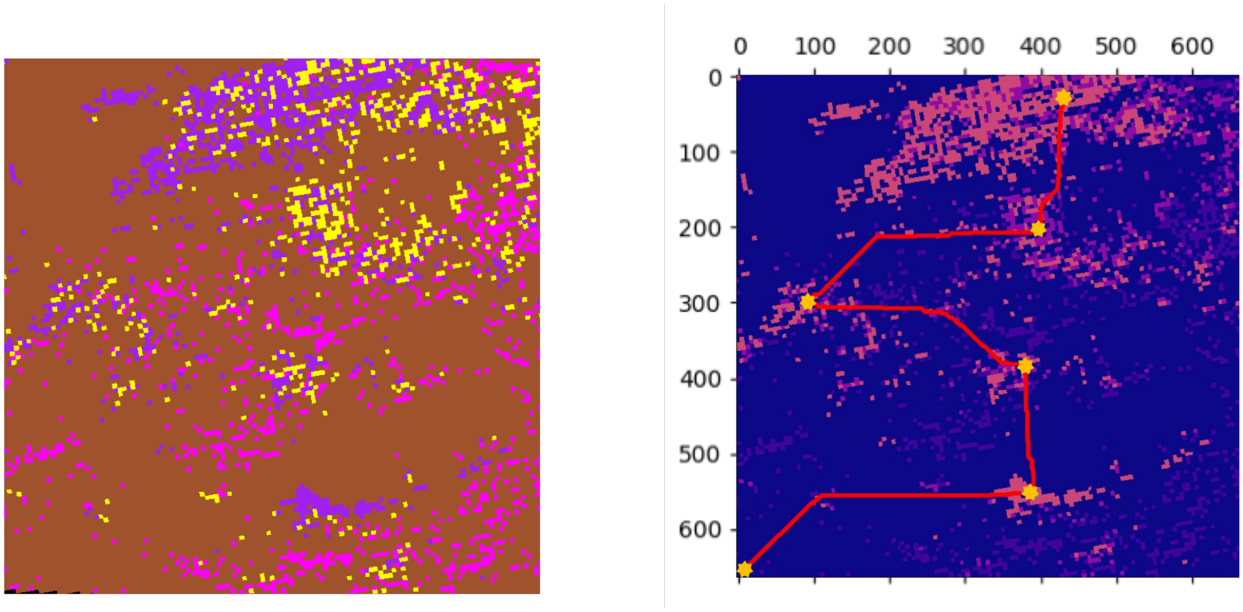

Fig. 4 Low resolution mineralogical map (left) and path plan for the rover with waypoints designated with yellow markers based on minerals of interest (right)

C. In Situ Classification

Once the rover is on the surface of the desert, the on-board sensors produce detailed data of the local terrain. The data collected in the same five bands as ASTER data are of particular interest, and the rover performs two operations in this feature space: reward prediction and novelty detection (described in Section IV.D). To generate a reward prediction, the observed data in the five bands are used as an input to the pre-trained SVM classifier described in Section IV.A, which then generates a reward prediction for the entire region. Inaccurate classification is expected at the beginning of the mission since the initial classification is only made on partial information obtained from the orbiters. The classification system is expected to improve as additional data is gathered, and the human operator is also able to intervene at any point to correct any mistakes the rover has made in classifying the terrain, thus providing additional training data for the rover. Eventually, we obtain an increasingly accurate reward map for the rover to perform its decision making on.

D. Novelty Detection

Novelty detection is performed on the images to determine if there are any new features that are not represented in the rover's initial estimated reward map. Our approach to novelty detection is based on one-class SVM classification [24]. This method groups all of the existing features that the rover has been trained on as one class, and any new features that are not represented are considered to be outliers. This approach requires mapping the existing data into a higher dimensional feature space and separating this representation from the origin with maximum margin. This generates one class for the existing features and any new feature that is mapped will produce either +1 (if it fits the original data set) or -1 (if there isn't a good fit). More formally, we can let Φ represent a feature map $X \longrightarrow F$ that maps the training data X to dot product space F. This allows us to use some kernel to evaluate the dot product in the image of Φ. For our purposes, we will be using a Gaussian, also known as radial basis function (RBF), kernel k as shown in Eq. 1.

$$k(x, y) = (\Phi(x)^T \Phi(y)) = \exp(\frac{-\|x - y\|^2}{c}) \tag{1}$$

To separate the hyperplane from the origin with maximum margin, the following dual problem is solved:

$$\underset{w \in F, \xi \in \Re^N, \rho \in \Re}{\text{argmin}} \frac{1}{2}\|w\|^2 + \frac{1}{\nu N}\sum_i^N \xi_i - \rho \tag{2}$$

$$\text{subject to} \begin{cases} (w \cdot \Phi(x_i)) \geq \rho - \xi_i \\ \xi_i \geq 0 \end{cases} \tag{3}$$

where w represents the normal vector used for regularization, ξ is a slack variable to measure the degree of misclassification of the data, ρ is the offset of desired hyperplane in the feature space, and $\nu \in (0, 1]$ is an upper bound on the fraction of training samples outside the decision boundaries and lower bound on the fraction of support vectors. For our purposes, we use $\nu = 0.0001$ to create the appropriate boundary for the five band thermal emission values in the initial training set.

E. Incremental Training

The detailed data gathered by the rover are used to update the reward map incrementally. The fine-grained pixel information provides more training data for the classification system, which helps in improving it throughout the traversal phase. The ground operators will have access to both of these information and can verify or intervene at any point in the process to modify the reward map themselves. We use the high resolution ground truth data to perform this step in our simulated framework. In addition to performing updates based on in situ observations, additional steps are taken if novel regions are detected within the field of view. Any novel observation becomes an automatic point of query to the human operator, and the answer is expected to be a reward value pertaining to the novel feature. If there are more than one type of novel feature and the associated reward value, this can also be easily communicated to the rover. When training the rover, information only on the novel region, rather than the entire map, is used, unless indicated otherwise by the human operator. This prevents the rover from re-training itself on information that it already has on the existing map and reduces the computational expense.

F. Performance Evaluation

To evaluate the performance of the rover and show the incremental learning of the classifier, we use precision-recall curve. Precision measures the ability of the classifier to identify only the relevant results, whereas recall measures the ability to find truly relevant data [25]. More specifically:

$$precision = \frac{true\ positive}{true\ positive + false\ positive} \tag{4}$$

$$recall = \frac{true\ positive}{true\ positive + false\ negative} \tag{5}$$

For multi-class classifier, each class is considered individually in a one-versus-all manner, and the results are averaged together. Precision for each reward class means the fraction of that detected reward value which is actually the true value. Recall refers to detecting all cases of that reward class within the map. We perform this evaluation over the entire map, comparing the ground truth data to the updated rover reward map after the rover has reached each of the waypoints.

V. Results

A. Initial Observations

It is assumed that the rover lands at coordinate point $(10, 650)$ and begins its operations here. The next two waypoints are located at coordinates $(390, 550)$ and $(375, 385)$. The ground truth reward map, the reward classification as predicted by the rover, and the novelty detection results at the three points are shown in Fig. 5. We note that the predicted classification doesn't match with the ground truth perfectly. This is expected since the initial classification training is performed based on the low resolution satellite data only, and the accuracy is expected to increase as more training data is gathered during the robot's traversal. The novelty detection results show that the novelty percentage is less than 1% of the observed region, so we can conclude that there are no significant novel areas to detect. Using the data produced at each of these three points and assuming that the observational data is downlinked to the operators to perform their own analysis, the classifier is retrained with the associated ground truth data at these points to simulate human operator increased involvement at the beginning of this mission. We also generate a precision-recall curve at each of the three points, which are shown in Fig. 6. The precision scores averaged across all classes are reproduced in Table 3 below. We can see an incremental improvement, albeit small, in the resulting precision score with each observation.

Table 3 Precision Score for Initial Observations

	Landing Site 1	Intermediate Point 1	Intermediate Point 2
Precision Score	0.64	0.65	0.67

B. Intermediate Point

The third waypoint that the rover visits is located at coordinate $(90, 300)$. The ground truth reward map, the rover predicted reward, and the novelty detection result at this point is shown in Fig 7, and the precision-recall curve is shown in Fig 8. The results from the novelty detection marks the areas that are unfamiliar in black. As the marked area is greater than 1% of the observed space, this is used to trigger the rover to immediately query the human operator and await their decision before proceeding.

In addition to sending the novelty detection results, the rover also sends a coordinate point centered at the observed novel region to the human operator as a waypoint suggestion in order to further explore the area. For this scenario, the chosen coordinate point by the rover is at $(175, 160)$. The human operator is then able to analyze the observed novel area to determine the proper mineralogical signature as well as the associated reward value. Given the waypoint suggestion from the rover, the operator can then approve, update, or reject the path alteration suggestion. Using our ground truth data, we see that the area marked as novel actually corresponds to signature of magnetite which is not initially captured in the rover's reward map. Hence, using the ground truth map, the area is assigned a reward value of 9, and given that this corresponds to the highest reward in the set, the waypoint is approved to further analyze the novel region. The regenerated path plan for the rover in shown in Fig. 9 (right) alongside the old path plan (left) for comparison. The new waypoint as well as the modified path are both shown in green. The rover then follows this new path to gather information from the novel region.

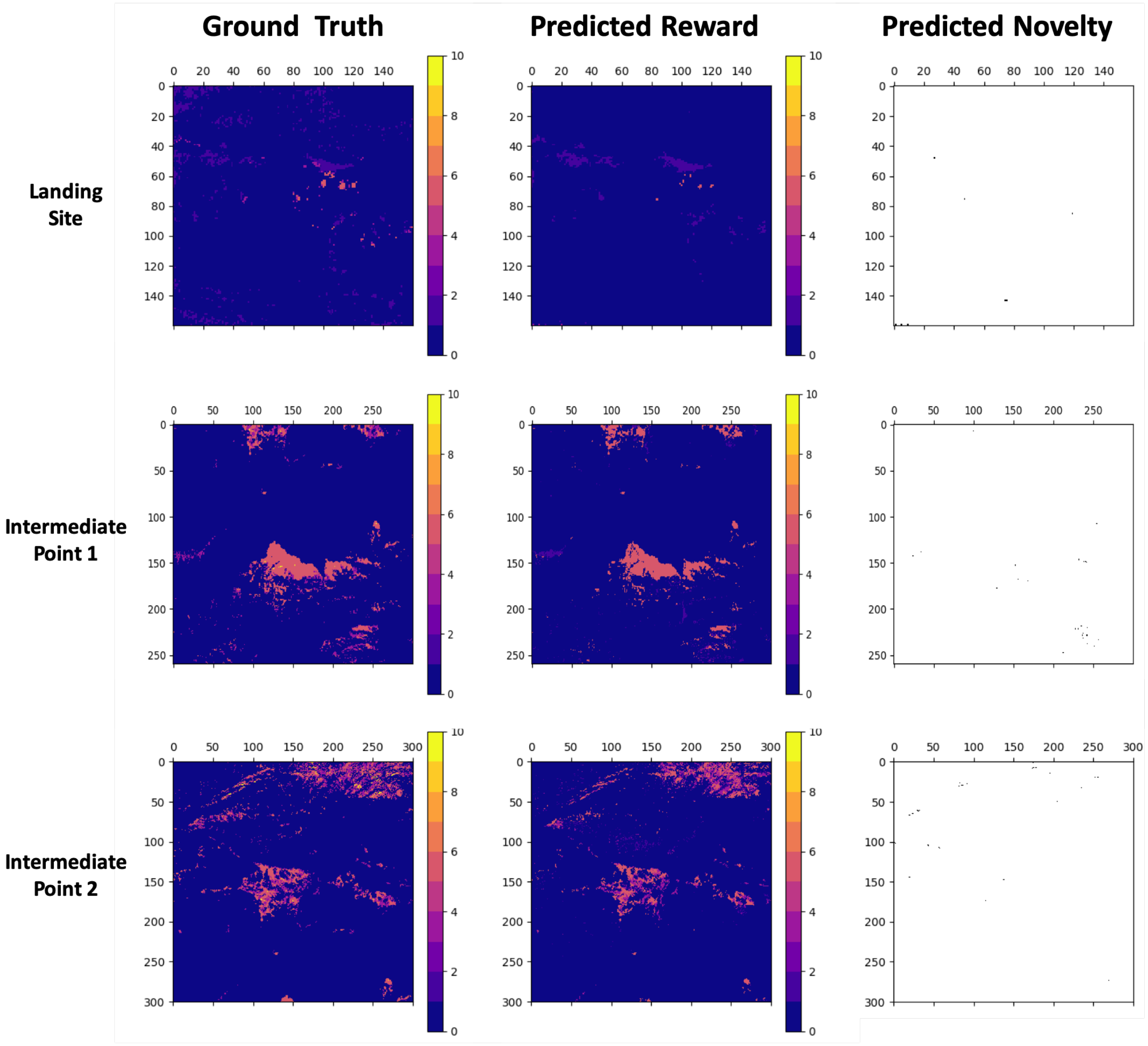

Fig. 5 Ground truth reward map, rover predicted reward map, and results from the novelty detection algorithm

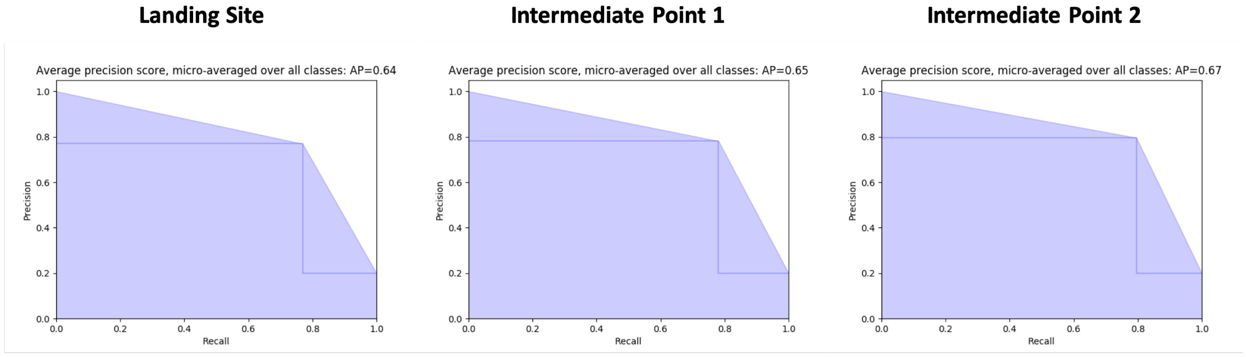

Fig. 6 Precision-recall curves for the three initial observations

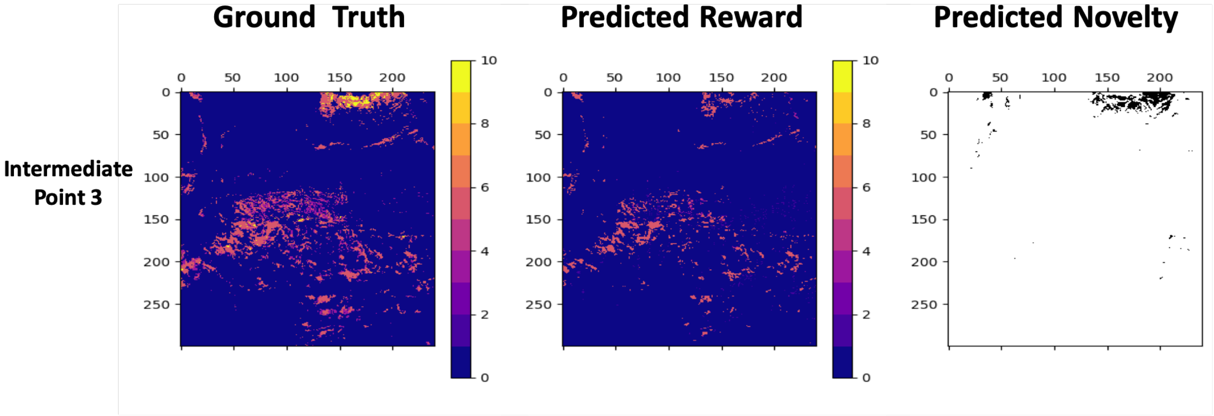

Fig. 7 Rover predicted reward map, ground truth reward map, and results from the novelty detection algorithm at intermediate point 3

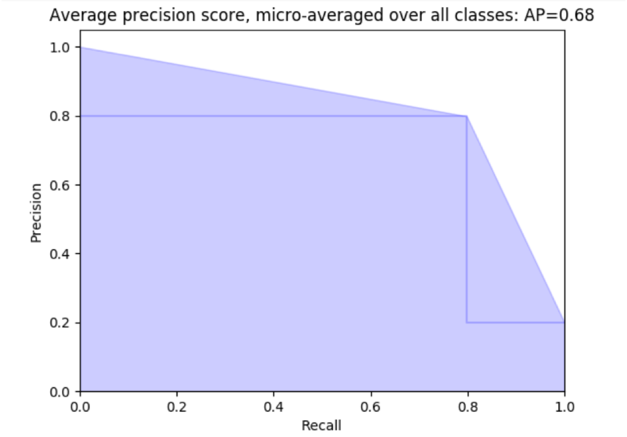

Fig. 8 Precision-recall curve at intermediate point 3

C. Novel Region

The region that we explore as a result of the novel observation has a dense magnetite signature. The ground truth reward map as well as the predicted reward map and the novelty detection for this region are shown in Fig. 11. Note the novelty detection returns almost all of the observed areas as known since the information from intermediate point in Section V.B is used to retrain the novelty detection algorithm. The resulting precision-recall curve for this region is shown in Fig. 11.

From the classifier results in Fig 11, we see a reward prediction map from the rover that is not completely representative of the ground truth data. This serves as a cue to the human operators to manually intervene in improving the classification knowledge for the rover. In this particular scenario, we are dealing with an unbalanced class as the classifier has not seen many instances of the novel feature. Recognizing this issue, the human operator provides additional training data that is relevant for that particular feature and the associated reward value so that the rover can be better prepared for observations related to that reward class in the future.

D. Goal Point

After exploring the novel region, the rover returns to its one remaining intermediate point and finally the target point. The ground truth reward map, the rover's reward map, and the novelty detection results for each of these points are shown in Fig. 12. The precision-recall curves at these points are reproduced in Fig 13, and a summary table for the precision scores for these two points and also at intermediate point 3 and the novel region are provided in Table 4.

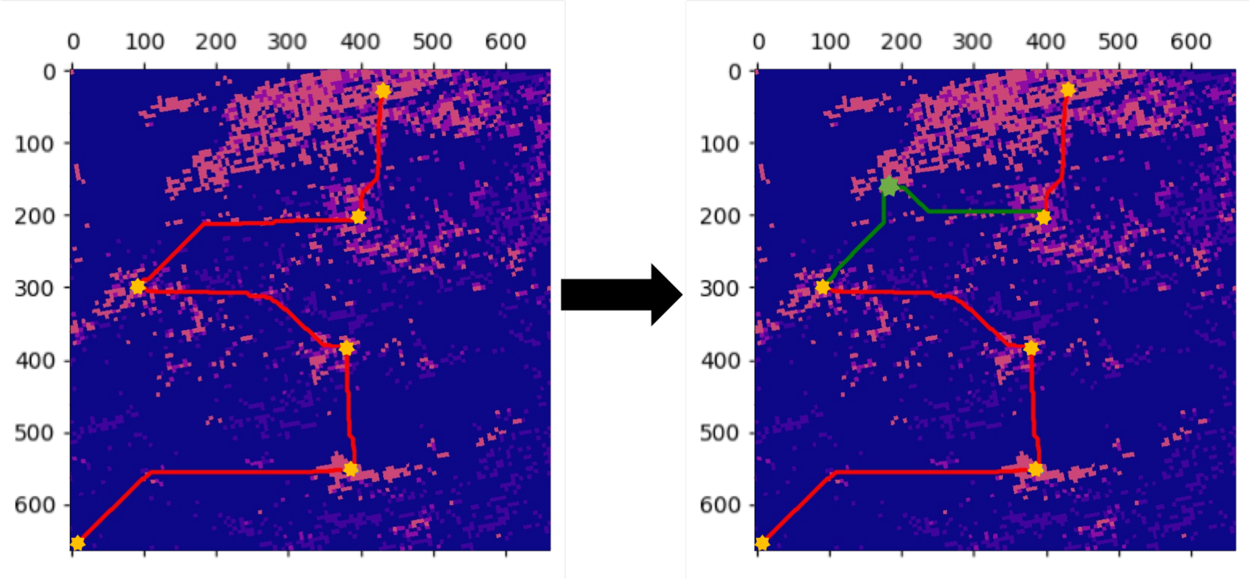

Fig. 9 Old (left) and new (right) path plan as a result of novelty detection and new waypoint selection

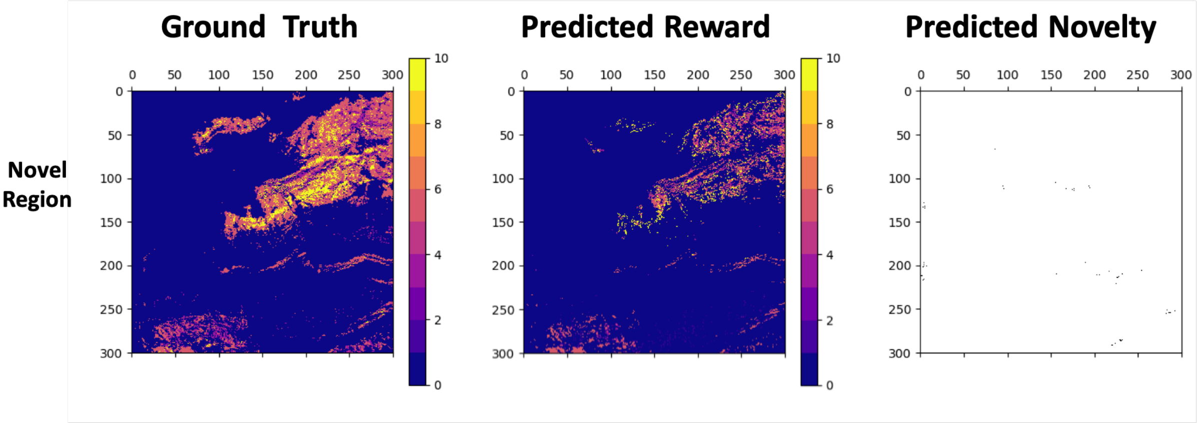

Fig. 10 Rover predicted reward map, ground truth reward map, and results from the novelty detection algorithm at the novel region

Following the results from the novel region (Section V.C), we are particularly interested in the classification results from the rover at these points after the human has provided additional data pertaining to the novel, high reward areas. As a result, we see an improved classification of the novel regions from the rover (shown in yellow in Fig. 12). This is obvious in the final reward map generated at the goal point where the rover succeeds in producing a detailed reward map based on its observation. The precision recall curve further corroborates this point as it shows a precision score of 0.75 at the final goal, which is the highest for the all the observations (Tables 3 and 4).

E. Final Results

After the rover has reached its end goal, we can compare the updated reward map with the initial. These results are shown in Fig. 14. As can be seen from the plots, the reward map has more details than was initially provided to the rover and it also marks out regions of novel areas not initially recognized as high reward areas based on the human communicated value. However, some areas of the final reward map are still coarse-grained, and these are regions that have not been explored by the rover. A comparison can also be made on the reward accumulated by the rover during its initial path plan as well as along its modified path plan. These values are calculated simply based on the rewards

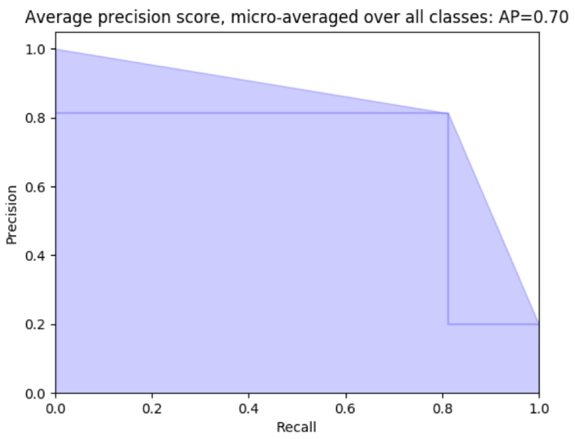

Fig. 11 Precision-recall curve at the novel region

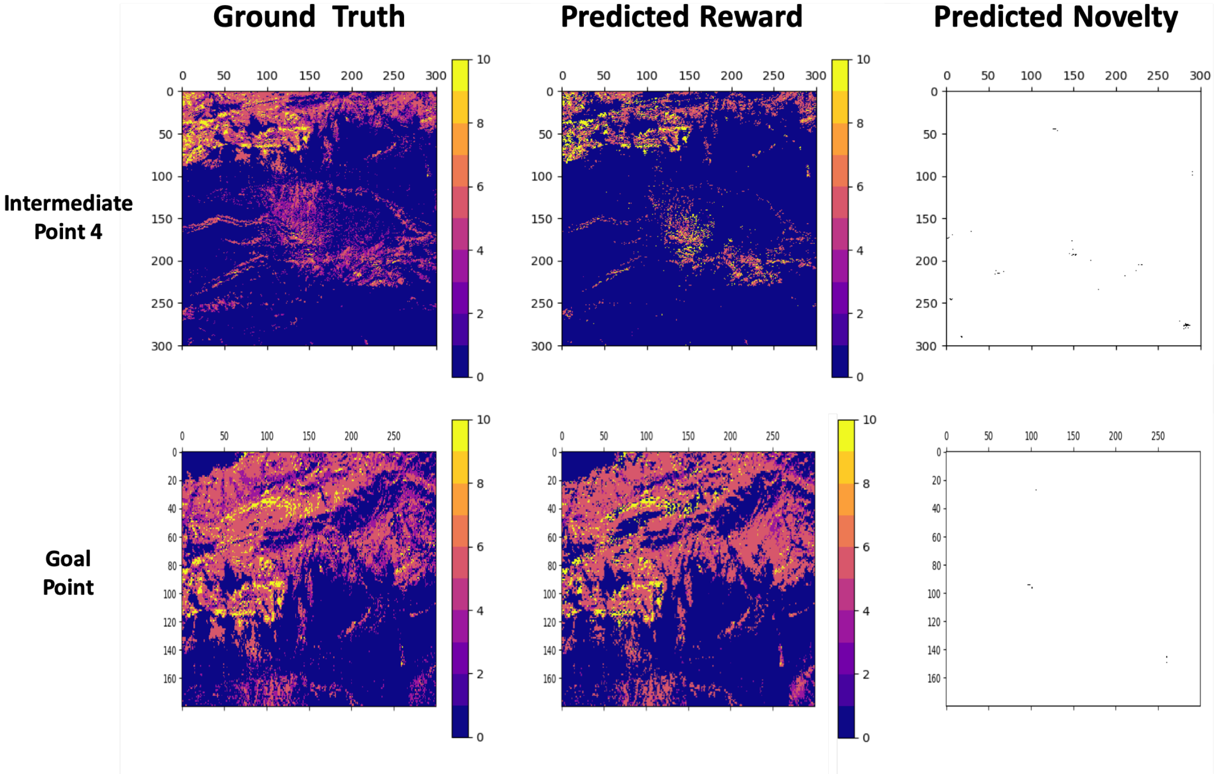

Fig. 12 Rover predicted reward map, ground truth reward map, and results from the novelty detection algorithm at intermediate point 4 and the final goal point

associated with visited pixel values. For the initial path path, we accumulate a reward of 1104 while for the modified path plan, we get 1198. This can further be compared based on the distance traveled by the rover during both of these phases. The reward per distance for the initial path plan is 0.73 whereas for the updated path plan, it is 0.85. Therefore, even though the distance travelled is increased, it is justified by the scientific information gain. A summary of the reward values is presented in Table 5 below.

We can further compare the precision-recall to evaluate the reward prediction by the rover against the ground truth. The results shown are for the results obtained at the landing site and eventually at the target site, and are shown in Fig 15. We can see that there is an improvement in the rover's classification of the terrain in comparison to its initial map,

Table 4 Precision Score for Latter Observations

	Intermediate Point 3	Novel Region	Intermediate Point 4	Goal Point
Precision Score	0.68	0.70	0.72	0.75

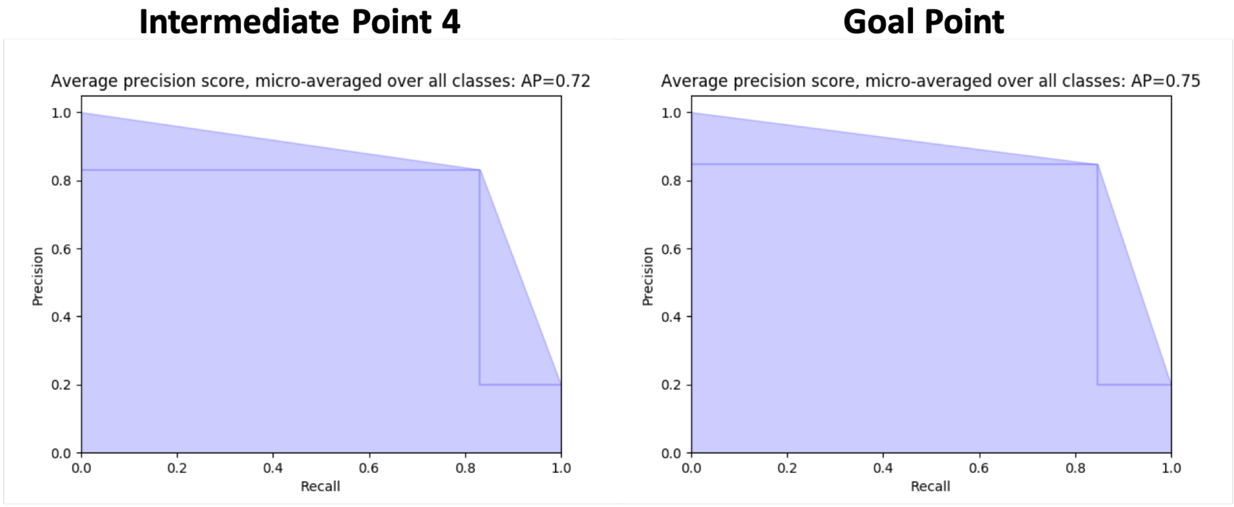

Fig. 13 Precision recall curve at intermediate point 4 and final goal point

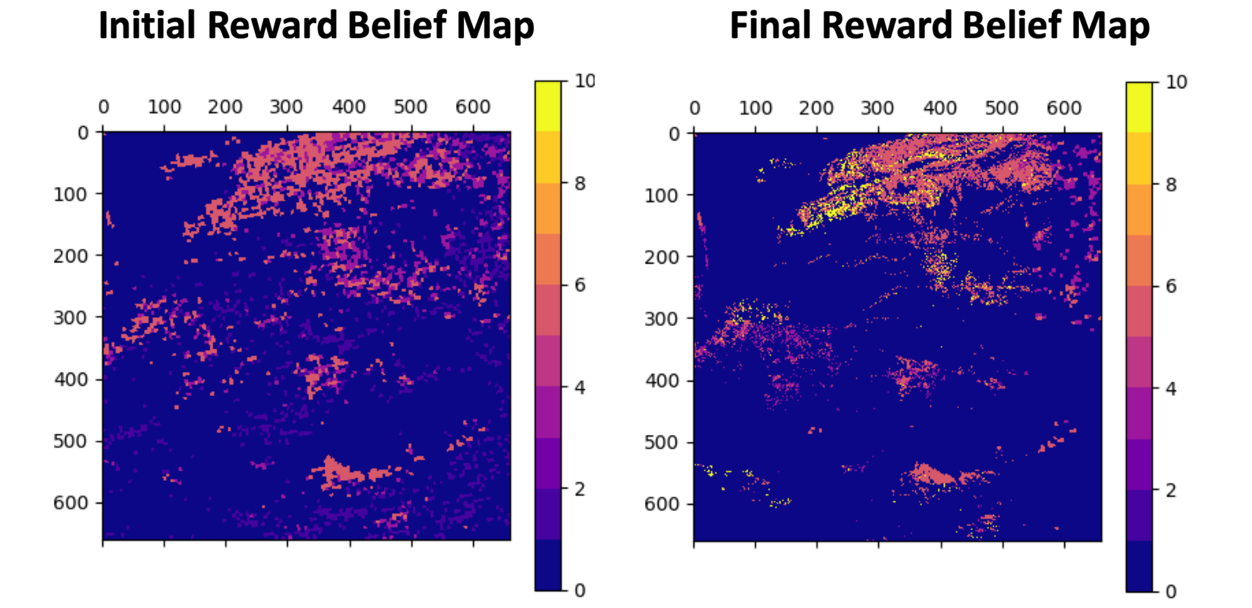

Fig. 14 Initial and updated rover's reward map

as expected. The results are simulated for less than 6 instances of training as the rover is traversing the environment. The precision score is expected to increase as the rover performs further incremental learning and/or trains at higher frequency. This is apparent from results shown in Tables 3 and 4 where we see the precision scores increasing after each observation and the training that follows. Furthermore, since the score is for the entire map where there are still regions

Table 5 Comparison of Gathered Reward Values

	Reward Value	Reward Per Pixel
Initial Path	1104	0.73
Modified Path	1198	0.85

not completely explored by the rover, the scores are comparatively low but expected.

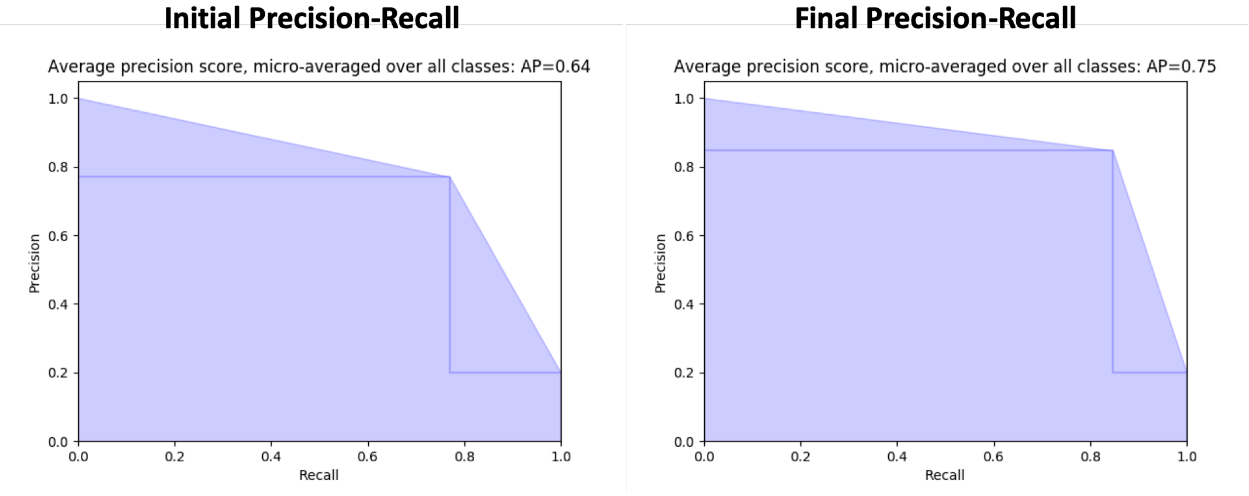

Fig. 15 Initial and updated precision recall curve.

VI. Conclusion and Future Work

In this paper, we showed how we can incrementally increase the autonomy and decision making capabilities of the rover within the current state-of-the-art by simply increasing the computation capabilities of the rover so that it can fully utilize its in situ observation by performing analysis on-board. Our approach to increasing the autonomy begins with providing a low resolution reward map to the rover that maps the observations to reward values. The rover can incrementally update its estimated reward map based on the high resolution observations that it gets on-board as well as by using novelty detection and querying the human operator. This reward map can then be used to make decisions such as those pertaining to traversal path or deployment of scientific instrumentation for further data gathering. To be conservative in our approach to autonomy for these expensive planetary exploration missions, our framework includes human-in-the-loop feedback so that any important decision has to be approved by the human operators on Earth. While we expect constant human supervision in the beginning of the mission phase to ensure proper training of the rover, we show that the number of queries made to the human is lessened with increased training of the rover, ultimately reaching a level where the rover can be trusted to perform the basic classification of its surrounding and propose appropriate actions. This helps to ease the workload of the human operators while still maintaining them as the ultimate decision makers in the mission.

This work has the potential to be extended in many different directions. For demonstrative purposes, we present and simulate a framework that allows the rover to change its path to capture scientifically valuable information. There is a potential to have multiple reward maps that can be used separately and/or jointly to capture various decision making criteria for robotic operations. This combined reward metrics based decision making for multiple phases in the mission is left for future work. We also focus on using one type of classification and novelty detection method throughout the paper as our focus on the framework rather than the analytic systems. This allows the framework to be adapted for different mission while additional studies can be performed to compare various classifiers and novelty detectors that are better suited for that particular mission. While we make the assumption that we are able to communicate with human operators despite the communication delays and limitation, this won't be true for actual missions. However, this is actually the strength of the framework we provide as the rover can perform efficient scouting operations while waiting to communicate with the human operator. Furthermore, using the concept of transfer learning, a rover's training in one planetary body can be used to augment the training of rover visiting another planetary body, therefore reducing the amount of training that has be to performed at the beginning for each new planetary body. Evaluation of how much training has to be performed, how much to trust the rover, and how much decision making capability should be left to the rover is left for future work.

References

[1] Ellery, A. A., “Robotic astrobiology – prospects for enhancing scientific productivity of mars rover missions,” *International Journal of Astrobiology*, Vol. 17, No. 3, 2018, p. 203–217. doi:10.1017/S1473550417000180.

[2] Christensen, P. R., “Regional dust deposits on Mars: Physical properties, age, and history,” 1986.

[3] Baldridge, A., Farmer, J., and Moersch, J., “Mars remote-sensing analog studies in the Badwater Basin, Death Valley, California,” *Journal of Geophysical Research (Planets)*, Vol. 109, 2004, pp. 12006–. doi:10.1029/2004JE002315.

[4] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C., “A Survey on Deep Transfer Learning,” *ICANN*, 2018.

[5] Fong, T., “Collaborative Control: A Robot-Centric Model for Vehicle Teleoperation,” Ph.D. thesis, Carnegie Mellon University, 2001.

[6] Luke Burks, L. B.-J. M. J. M. S. V., Ian Loefgren, and Ahmed, N., “Closed-loop Bayesian Semantic Data Fusion for Collaborative Human-Autonomy Target Search,” *International Conference on Information Fusion (Fusion 2018)*, 2018.

[7] Ng, A. Y., Russell, S. J., et al., “Algorithms for inverse reinforcement learning.” *Icml*, Vol. 1, 2000, p. 2.

[8] Abbeel, P., and Ng, A. Y., “Apprenticeship learning via inverse reinforcement learning,” *International Conference on Machine Learning (ICML 2004)*, ACM, 2004, p. 1.

[9] Schaal, S., “Learning from demonstration,” *Advances in Neural Information Processing Systems*, 1997, pp. 1040–1046.

[10] Tobias Kaupp, H. D.-W., Alexi Makarenko, “Human-robot communication for collaborative decision-making - A probabilistic approach,” *Robotics and Autonomous Systems*, 2010.

[11] Pimentel, M. A. F., Clifton, D. A., Clifton, L., and Tarassenko, L., “Review: A Review of Novelty Detection,” *Signal Process.*, Vol. 99, 2014, pp. 215–249. doi:10.1016/j.sigpro.2013.12.026, URL `http://dx.doi.org/10.1016/j.sigpro.2013.12.026`.

[12] Crook, P., Hayes, G., et al., “A robot implementation of a biologically inspired method for novelty detection,” *Proceedings of the Towards Intelligent Mobile Robots Conference*, 2001.

[13] Neto, H. V., and Nehmzow, U., “Real-time Automated Visual Inspection using Mobile Robots,” *Journal of Intelligent and Robotic Systems*, Vol. 49, 2007, pp. 293–307.

[14] Marsland, S., Nehmzow, U., and Shapiro, J., “On-line novelty detection for autonomous mobile robots,” *Robotics and Autonomous Systems*, Vol. 51, 2005, pp. 191–206. doi:10.1016/j.robot.2004.10.006.

[15] Neto, H. V., and Nehmzow, U., “Incremental PCA: An alternative approach for novelty detection,” *Towards Autonomous Robotic Systems*, 2005.

[16] Özbilge, E., “Experiments in online expectation-based novelty-detection using 3D shape and colour perceptions for mobile robot inspection,” *Robotics and Autonomous Systems*, Vol. 117, 2019, pp. 68 – 79. doi:https://doi.org/10.1016/j.robot.2019.04.003, URL `http://www.sciencedirect.com/science/article/pii/S0921889019300636`.

[17] Estlin, T. A., Bornstein, B. J., Gaines, D. M., Anderson, R. C., Thompson, D. R., Burl, M., Castaño, R., and Judd, M., “AEGIS Automated Science Targeting for the MER Opportunity Rover,” *ACM Trans. Intell. Syst. Technol.*, Vol. 3, No. 3, 2012, pp. 50:1–50:19. doi:10.1145/2168752.2168764, URL `http://doi.acm.org/10.1145/2168752.2168764`.

[18] Carsten, J., Rankin, A., Ferguson, D., and Stentz, A., “Global Path Planning on Board the Mars Exploration Rovers,” *2007 IEEE Aerospace Conference*, 2007, pp. 1–11. doi:10.1109/AERO.2007.352683.

[19] Thompson, D. R., and Wettergreen, D. S., “Robot Science Autonomy in the Atacama Desert and Beyond,” 2013.

[20] NASA LP DAAC, “ASTER Level 1 Precision Terrain Corrected Registered At-Sensor Radiance V003 [Data set],” `https://doi.org/10.5067/ASTER/AST_L1T.00`, 2015. doi:10.5067/ASTER/AST_L1T.003, distributed by NASA EOSDIS Land Processes DAAC.

[21] Baldridge, A., Hook, S., Grove, C., and Rivera, G., “The ASTER spectral library version 2.0,” *Remote Sensing of Environment*, Vol. 113, 2009, pp. 711–715. doi:10.1016/j.rse.2008.11.007.

[22] Laskov, P., Gehl, C., Krüger, S., and Müller, K.-R., “Incremental Support Vector Learning: Analysis, Implementation and Applications,” *J. Mach. Learn. Res.*, Vol. 7, 2006, pp. 1909–1936. URL `http://dl.acm.org/citation.cfm?id=1248547.1248616`.

[23] Hart, P. E., Nilsson, N. J., and Raphael, B., "A Formal Basis for the Heuristic Determination of Minimum Cost Paths," *IEEE Transactions on Systems Science and Cybernetics*, Vol. 4, No. 2, 1968, pp. 100–107.

[24] Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., and Platt, J., "Support Vector Method for Novelty Detection," *Proceedings of the 12th International Conference on Neural Information Processing Systems*, MIT Press, Cambridge, MA, USA, 1999, pp. 582–588. URL `http://dl.acm.org/citation.cfm?id=3009657.3009740`.

[25] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, Vol. 12, 2011, pp. 2825–2830.