Concepts for Science Autonomy during Robotic Traverse and Survey

Trey Smith, Scott Niekum, David Thompson, and David Wettergreen
Robotics Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213, USA
{trey,sniekum,drt,dsw}@ri.cmu.edu

Abstract— Future Mars rovers will have the ability to autonomously navigate for distances of kilometers. In one sol a traverse may take a rover into unexplored areas beyond its local horizon. Naturally, scientists cannot specify particular targets for the rover in an area they have not yet seen. This paper analyzes what they *can* specify: priorities that provide the rover with enough information to autonomously select science targets using its onboard sensing. Several autonomous science operational modes and priority types are discussed. We also introduce a science priority language. A team of scientists was asked to use the language in specifying targets for a tele-operated rover, and qualitative results are reported.

TABLE OF CONTENTS

1. INTRODUCTION

Future Mars rovers will have the ability to autonomously navigate for distances of kilometers. At these scales, a day's traverse takes the rover into unexplored areas over its local horizon. Naturally, scientists cannot specify particular targets for the rover in an area they have not yet seen. This paper analyzes what they *can* specify: priorities that provide the rover with enough information to autonomously select science targets using its onboard sensing.

We aim to develop *science autonomy* (SA), the ability of a rover to reason about science goals and the science data it collects in order to make more effective decisions and improve the quality of science data return. Planetary rovers have limited resources for sampling their surroundings and limited downlink bandwidth for returning the data—we want every sample to count.

The simplest use of SA is *selective data return*, in which the rover moves and studies its surroundings as usual (without analyzing incoming science data in real time), and then makes decisions about what subset to download at its next communication window. Selective data return presents minimal mission risk—the SA technology only selects what data to download, and in case of errors any important data that was passed over may later be recoverable from the rover's onboard storage.

A more challenging and potentially more rewarding approach is *active science autonomy*, in which the rover reacts to incoming science data by selectively applying its sensors and possibly moving to features of interest. This approach makes new operational modes available to the science team. For instance, in the *science on the fly* mode the rover would opportunistically sample interesting features observed during long traverses. In the *intelligent site survey* mode the rover would characterize a site, choosing its coverage and sampling strategies in order to assemble a useful summary of what is present. We discuss these new operational modes in terms of their potential benefits and technical challenges.

In order for the rover to selectively sample features of interest, it must have some onboard representation of scientist priorities. Some types of priorities include *target signatures*, which prioritize a particular class of features (e.g. "sedimentary rocks"), and *representative sampling*, which prioritizes a variety of samples so that there is a sample to represent each class of feature [3]. One way to represent science priorities is using a value function that scores possible returned data sets, allowing them to be ordered from most to least preferred. We introduce a language for specifying this value function. A team of scientists was asked to use the language in specifying targets for a tele-operated rover, and qualitative results are reported.

2. SCIENCE AUTONOMY OPERATIONAL MODES

Expanded rover mobility has become technically feasible only recently, and operations strategies have not yet fully adapted. In some rover field tests, the science team has underused the mobility of the rover [2]. Nonetheless, a straightforward operations approach is emerging, which we call *baseline high-mobility operations*. Using this approach, the science team selects interesting sites in orbital imagery and can direct the rover to travel long distances and visit multiple

0-7803-8870-4/05/$20.00/©2005 IEEE
IEEEAC paper #1249

sites per day. The rover's time is split between the following modes:

- *Directed sampling:* The science team gives the rover specific local targets based on data from previous downlinks.
- *Periodic traverse sampling:* While traveling between interesting sites, the rover takes periodic samples.
- *Periodic site survey:* When the rover reaches an interesting site, it follows a coverage pattern and takes periodic samples. This data constitutes a preliminary survey, and the science team can follow up with directed sampling as necessary.

Our proposal for structuring active SA is to add new modes that extend baseline high-mobility operations. Some modes under consideration are the following (in order from least to most advanced):

- *Science on the fly:* While traveling between sites, the rover watches for potential science targets. It uses priorities provided by the science team to decide whether it is worthwhile to delay its traverse in order to perform follow-up observations, such as taking a spectrometer reading or visiting the target and using contact sensors.
- *Intelligent site survey:* When the rover reaches an interesting site, it moves around the site using an exploration strategy that balances coverage with selective sampling based on scientist priorities. This data constitutes a preliminary survey, and the science team can follow up with directed sampling as necessary.
- *Science-aware path following:* The rover follows a path defined by science features. Examples include the margin of a dry riverbed or crater, or evaporite deposits that mark an ancient shoreline.
- *Science-aware region mapping:* The rover identifies areas with uniform science properties and attempts to find their boundaries, in order to determine extent and generate a map. Example areas are geological units and habitats defined by the presence or absence of particular organisms.

In order to decide where to focus our development effort, we can measure each mode according to several criteria:

1. *Broad applicability:* It should be applicable under circumstances that are commonly encountered in mission operations.
2. *Efficiency enhancement:* It should provide more useful science data than the comparable baseline strategy given the same resource investment (in terms of energy, time, data volume, and operator attention).
3. *Ease of migration:* It should not be a "disruptive technology". It should integrate easily with existing systems and require minimal retraining for the science team.
4. *Testability:* There should be clear performance criteria, and it should be straightforward to compare performance with the baseline strategy.
5. *Feasibility:* It should be technically feasible to develop.

We believe that all four of the SA operational modes described earlier have broad applicability and good potential to enhance efficiency, although this is difficult to determine *a priori* and would best be analyzed through tests in the field.

However, science on the fly and intelligent site survey have a clear advantage in terms of ease of migration. They are essentially drop-in replacements for periodic traverse sampling and periodic site survey. The science team can specify daily activities just as they would in the baseline strategy, then "flip the switch" to use SA operational modes where appropriate. (Of course, there is additional effort, such as specifying priorities, which is discussed later.) The fact that these modes have corresponding baseline modes also improves testability. We can use existing performance metrics for the baseline modes, and easily compare results. For these reasons, our SA development focuses on science on the fly and intelligent site survey.

3. Science Priority Specification

In order for the rover to selectively sample features of interest, it must have some onboard representation of scientist priorities. We suggest an SA architecture that cleanly separates priority specification (a scientific question) from rover control strategies that implement the priorities (an engineering problem).

Separating priority specification from rover control strategy has advantages in terms of evaluation. Our overall goal in designing SA systems is to help the rover and science team together gain a more complete and accurate scientific understanding. The corresponding evaluation methodology would aim to measure how scientific accuracy varies as SA system parameters are tuned. For example, we could design an experiment in which two teams of scientists use a rover to study the same site, each time with different SA system parameters, and compare the accuracy of the two sets of conclusions (relative to ground truth).

This overall evaluation methodology directly addresses the goal of improving scientific understanding, but it is difficult to implement. First, how do we quantitatively measure the accuracy of scientific conclusions relative to ground truth? Because the conclusions are generally very unstructured, comparisons are difficult [9]. Second, experiments are difficult to control. The site must be continually varied because the science team can only get a fresh look at each site once; after that, the experiment is compromised by their prior knowledge. Finally, it is logistically difficult to assemble a group of scientists for repeated experiments. For these reasons, overall evaluation is complicated and likely to yield ambiguous results.

However, cleanly separating science priority specification from rover control strategy allows us to evaluate them separately. In particular, we can vary the rover control strategy and calculate the score of the returned data set according to the specified priorities. This experiment can be performed multiple times at the same site without repeated involvement from scientists.

Having motivated the notion of an explicit priority representation, we can ask what form it should take. In designing our language, we considered several design criteria. They are listed here in order from most to least important:

1. *Simple:* The language should be simple to interpret onboard the rover and easy for the scientists to use (with the appropriate supporting interface).
2. *Expressive:* The language should be flexible enough to express a range of scientist intentions, including:
 (a) *Target signatures:* Prioritizes particular classes of features, such as "white rocks".
 (b) *Novelty detection:* Prioritizes features that are unlike any previously encountered.
 (c) *Representative sampling:* Prioritizes a variety of data so that there is a sample to represent each class of feature present at a site.
3. *Maintainable:* It should be easy to understand, reuse, and modify priority specifications. Scientists should not need to start from scratch every time they want to tune the priorities.
4. *Autonomy compatible:* The specification should be compatible with the underlying science data understanding mechanism.

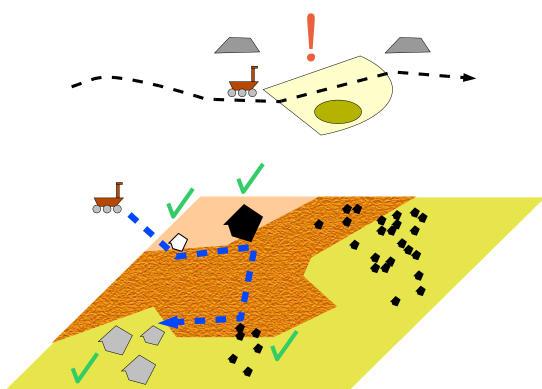

Figure 1. Operational mode sketches: (top) science on the fly with novelty detection, (bottom) intelligent site survey with representative sampling.

A number of schemes have been proposed for prioritizing science activities. During its Antarctic meteorite search, the Nomad robot focused on the single task of classifying rocks as meteorite/non-meteorite [8]. Active selection of the appropriate sensor to use on a rock relied on an estimate of information gain in the classification. The MISUS system simulated multiple rovers exploring an environment and classifying rocks into clusters based on their spectral characteristics [4]. Observations were preferred if they were likely to improve the accuracy of the clustering model. Neither of these representations offered the scientists detailed control over priorities. The closest representation to ours was developed by the OASIS project [3]. We have adopted their priority concepts (target signatures, etc.) as well as their approach of working closely with a science team. Our representation extends theirs by (1) providing a more flexible way to specify classes of features and (2) clarifying certain issues, such as how to use the various priority concepts together at the same time, and how to provide control over the bandwidth allocation.

Abstractly, we represent priorities with a value function that scores possible returned data sets, allowing them to be ranked from most to least preferred. A set of target signatures forms the core of the function representation. Each time a sample in the data set matches a target signature, the value of the data set is incremented. Novelty detection and representative sampling are discussed later—they are supported by a mechanism for automatically adding new target signatures to the set.

The representation must provide careful control over allocation of downlink data volume, since it is often a scarce resource. For example, given enough volume for 10 images, the science team might want 5 to target white rocks and 5 to target angular rocks. Unfortunately, allocation is complicated when the requested samples are to be found in unexplored areas. There might not be enough features of a given type to fill the request, or unexpected interesting features might need to be included, reducing space available to the original allocation. Lacking any special insight into how to handle this issue, we decided on a simple *quantity discount* mechanism. Scientists can specify a reward and a quantity cap for each target signature. Samples matching a high reward signature are preferred until the signature's quantity cap is reached.

Some samples in a data set are primarily useful for demonstrating the existence of a particular type of feature at a site. For this purpose, only one sample of that type of feature is needed; multiple samples at the same site would be redundant. To capture this idea, we provide the *neighbor discount*, which prevents the same target signature from matching two samples that are too close together.

Formally, each target signature t in the function representation consists of (1) a matching criterion $\mathrm{match}(t, \cdot)$ stating what samples in a data set match the signature, (2) a reward $\mathrm{reward}(t)$ to be added to the overall value of the data set when the signature fires, and (3) additional parameters $\mathrm{nthreshold}(t)$ and $\mathrm{qthreshold}(t)$ described below.

The overall value function V is defined to be

$$V(D) = \sum_{i=1}^{n} \sum_{t \in T} \mathrm{match}(t, d_i) V(t, d_i \mid d_1, \ldots, d_{i-1}), \quad (1)$$

where D is the overall data set consisting of samples $\{d_1, \ldots, d_n\}$. We number the samples (rather arbitrarily) in order by the time they were taken. T is the set of target signatures. $\mathrm{match}(t, d)$ ranges from 0 to 1, 1 meaning that signature t is a perfect match for sample d. Note that this formulation allows multiple signatures to match and fire for each sample. The value of each sample/signature pair is conditioned on prior samples as follows:

$$
\begin{aligned}
V(t, d_i \mid d_1, \ldots, d_{i-1}) \quad = \quad & \mathrm{qdiscount}(t \mid d_1, \ldots, d_{i-1}) \\
& \mathrm{ndiscount}(t, d_i \mid d_1, \ldots, d_{i-1}) \\
& \mathrm{reward}(t), \qquad (2)
\end{aligned}
$$

where one sees the two types of discounting:

- *Quantity discount:* Each signature is constrained to match only a certain proportion of the samples in the data set. The quantity discount factor $\mathrm{qdiscount}(t \mid d_1, \ldots, d_{i-1})$ is 0 if among the prior samples $d_1, \ldots, d_{i-1}$ there are k samples that match signature t, with $k/|D| \geq \mathrm{qthreshold}(t)$. The discount factor is 1 otherwise.
- *Neighbor discount:* The sample is discounted if it is not locally unique. The neighbor discount factor $\mathrm{ndiscount}(t, d_i \mid d_1, \ldots, d_{i-1})$ is 0 if one of the prior samples $d_1, \ldots, d_{i-1}$ both matches signature t and is within distance $\mathrm{nthreshold}(t)$ of the current sample d_i. The discount factor is 1 otherwise.[1]

Matching Samples to Target Signatures

This section discusses what constitutes a sample, and how we determine the quality of match between a sample and a target signature. Broadly defined, a sample could mean any kind of data: a spectrum, an image, a particular part of an image, or an assemblage from multiple sensors (such as an image mosaic with associated spectral information). Samples are matched to target signatures based on their measurable attributes, which fall broadly into three categories:

- *Meta-data attributes:* Attributes that describe the data product itself, such as the type of instrument used, the resolution, or the exposure time.
- *Viewing conditions:* Attributes such as ambient light levels or the distance of a camera from what it is viewing.
- *Target attributes:* Attributes such as the number and type of rocks present, distribution of soil types, or presence of microorganisms.

For the remainder of this paper (in accord with our experiments), we restrict a sample to mean a fixed assemblage of data products that provide a summary of the patch of terrain around a single rover location, roughly 5×5 meters. Because the same data products are always included in a sample, all samples have identical meta-data attributes and similar viewing conditions. As a result, our discussion focuses on target attributes, and our goal is to specify target signatures that match based on attributes of a patch of terrain. In particular, we focus on the rocks and soil present in a terrain patch. Each patch is broken up into discrete features such as individual rocks.

[1] Note that both types of discounting use a threshold function for simplicity; it may be that some class of smooth functions is more appropriate.

Formally, the matching criterion for each target signature consists of (1) a *feature class* defined in terms of measurable attributes of a feature (such as albedo of an individual rock), and (2) a threshold density of features in the patch that must be members of the feature class in order for the patch to match the target signature. For rocks, the threshold density can be measured in rocks per unit area or as a proportion of the total number of rocks present. For soil, the density is measured as a proportion of total area.

Membership of features in feature classes need not be all-or-nothing. We allow for partial membership (scores ranging from 0 to 1), and when matching a patch to a target signature, the sum of the membership scores in the signature's feature class is the figure of merit that we compare to the signature's density threshold. Although it is not necessary, we find it convenient to make the feature classes exclusive in the sense that the membership scores for a feature across all classes must sum to 1. This allows the scores to be thought of as a probabilistic classification.

Now, given a feature f, we want to define its membership score with respect to a class c. The decision must be based on what we know about f: we assume there is a set X of discrete or real-valued attributes, such that for every feature f and attribute x we have a measurement $\mathrm{value}(f, x)$. For example, a rock has measurable attributes like size, albedo, and angularity. We define a class c with a membership function $\mathrm{member}(c, f)$ that takes the measured attributes of f as inputs and outputs a score ranging from 0 to 1.

The membership function is represented as the conjunction of intervals. Each class c has an interval for every attribute x, with specified bounds $\min(c, x)$ and $\max(c, x)$. The interval can be open-ended on either or both ends if bounds of $\pm\infty$ are specified. The overall membership score is

$$\mathrm{member}(c, f) = \alpha \prod_{x \in X} \mathrm{member}(c, f, x), \qquad (3)$$

where $\mathrm{member}(c, f, x)$ is 1 when $\min(c, x) \leq \mathrm{value}(f, x) \leq \max(c, x)$, and 0 otherwise. α is a normalizing constant.

This scheme of specifying numerical bounds has strengths and weaknesses. It is most appropriate for what we call *hypothesis-driven science*, in which scientists command rover actions based on specific prior hypotheses about a site. For instance, before our recent field expedition, members of our science team hypothesized that we would find photosynthetic endoliths in large translucent rocks that are easily recognized because of their high albedo. This translates directly to a class of "all rocks with albedo $>$ 0.9 and size $>$ 10 cm." Although disjunctions such as "all rocks with size $>$ 10 cm *or* angularity $>$ 0.5" cannot be stated directly, they can be captured if necessary by specifying multiple target signatures.

However, numerical bounds are less natural for *data-driven science*, in which rover commands are based on interesting features in the data that may or may not be associated with specific hypotheses. An example is the mineral "blueberries" discovered by the MER rovers, considered to be evidence of past water [5]. Their existence was not predicted in advance, but once they were discovered, it was important to follow the data and search for them in subsequent observations. An effective SA system should make it easy to adjust priorities based on new data so that spontaneous ideas from the science team are more likely to be captured and used onboard the rover.

Data-driven classes of interest are more naturally expressed in terms of *in situ* examples: "rocks like these three that we observed yesterday." This is also convenient from the perspective of onboard science data understanding. Human subjective judgment is poor for specifying precise thresholds, and examples from the lab tend to differ from those in the field in a variety of ways, such as ambient lighting and the distribution of sample types. *In situ* examples should generally lead to more robust classification. Another advantage is that automatic clustering algorithms can naturally express their classes in terms of examples, making it easy to use manual and auto-generated classes side by side with the same representation.

In order to support both numerical and example-based representations, we extend the class representation in two ways: (1) we allow sets of positive and negative examples to optionally be specified, and (2) the interval for any attribute x can be left unspecified, allowing the SA system to automatically select a attribute membership function $\mathrm{member}(c, f, x)$ so as to optimize inclusion of positive examples and exclusion of negative examples. We intentionally leave the details of the optimization algorithm open so that the language is neutral across algorithms. The algorithm may optimize additional criteria, for instance minimizing overlap between classes, and the class of functions it considers need not be restricted to intervals. We present a companion paper at this conference that discusses Gaussian attribute membership functions and an expectation-maximization algorithm for optimization [10].

Finally, the priority language should be able to express novelty detection and representative sampling preferences. Our approach assumes the presence of an automatic clustering scheme that identifies new classes of features in the rover's science data.[2] We extend the specification by allowing scientists to specify *signature templates*, which state how to react when a new class is identified—a template creates a new implicit target signature and adds it to the signatures already specified by scientists. Specifying the template means specifying the parameters (priority, $\mathrm{nthreshold}$, etc.) of the signatures it should generate.

We provide the ability to specify two templates, one for novelty detection and one for representative sampling. The novelty detection template (which we expect to have a higher priority) creates a signature when a novel class is first identified onboard the rover. Later, after each interaction with the science team, the novel classes from that command cycle "graduate": the scientists can explicitly specify a signature involving that class, or, if they choose not to, the signature generated by the novelty detection template is replaced with one generated by the representative sampling template.

Overall, our priority language attempts to provide a balance across design criteria. In some places we have made specifications more maintainable at the cost of some expressivity. For instance, the effect of a rule is an additive term independent of other rules in the set. This independence limits the language but makes it easier to edit rules without concern for complex interactions. A range of priority concepts are supported, including target signatures, representative sampling, and novelty detection. And in support of autonomy integration, the extended class representation supports both manual tuning of classes and compatibility with automatic clustering algorithms.

4. Experimental Procedure

Our SA research is associated with the Life in the Atacama project, an effort to study the limits of life in the Atacama Desert of Chile, one of the driest places on Earth. By deploying a rover there we seek insights into how organisms adapt to extreme environments and how to study them with a rover under Mars-relevant operational constraints.

Fig. 2 shows our Zoë rover and its sensor configuration. Zoë is a solar-powered rover approximately 2 meters long and able to navigate autonomously over kilometer-scale distances at a maximum speed of about 2 km/hour. It has a suite of science instruments designed to support geology and biology studies:

- *Stereo Panoramic Imager (SPI):* A stereo triple of color cameras mounted on a pan/tilt unit. Each has a 21^o field of view and a 1024x768 CCD, providing angular resolution similar to the human eye. The SPI is primarily used to capture panoramic mosaics that provide the scientists with visual context for understanding a site.
- *Vis/NIR spectrometer:* A visible/near-infrared reflectance spectrometer. Light is gathered with a 1^o field of view fore-optic mounted on the pan/tilt unit and aligned with the SPI optical axis. The spectrometer is primarily used to characterize minerals.
- *NavCams:* A stereo pair of color cameras. Each camera has a 60^o field of view and a 1024x768 CCD. They are fixed to a mast rising from the front axle of the robot. Since the front axle can be steered left or right $\pm 15^o$, the camera pointing can be controlled in a gross sense. The tilt angle is fixed so that the cameras view an area about 1-10 m in front of the robot. The NavCams are primarily used for obstacle avoidance.
- *Fluorescence imager (FI):* The FI is an imaging system comprised of a flash lamp, color filters, and a cooled CCD. It can detect chlorophyll directly, and can also detect protein

[2] Automatic clustering schemes for SA are discussed in our companion paper.

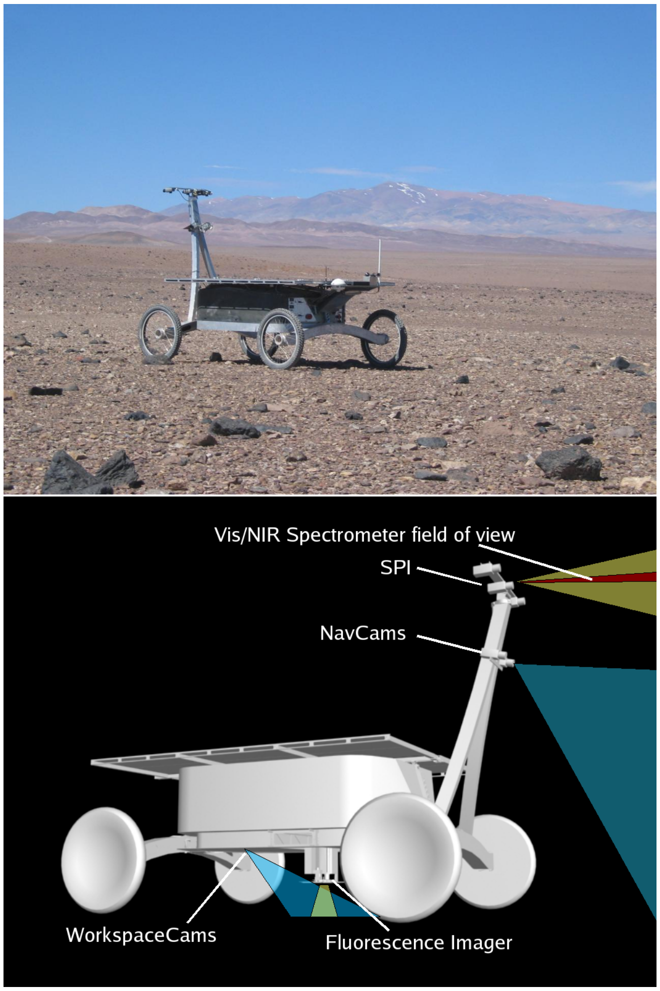

Figure 2. Zoë rover: (top) deployed in the Atacama desert, (bottom) sensor configuration.

and DNA when used with fluorescent indicator dyes. The FI points down from the belly of the rover with a 10×10 cm field of view. It is the rover's primary instrument for characterizing and unambiguously confirming the presence of life.

• *WorkspaceCams:* A stereo pair of color cameras mounted on the belly of the robot and aimed forward. Each camera has a 94^o field of view. The WorkspaceCams are primarily used to provide context that supports interpretation of FI images.

The Atacama investigation has three planned rover expeditions to the Atacama in 2003, 2004, and 2005. For the recently completed 2004 expedition, we did not have a complete SA system implementation, but we took advantage of the field deployment by conducting tests designed to guide the design of our SA system, which should be ready to use onboard in 2005.

This section describes one test in particular, the *human-onboard* experiment. A team of scientists was asked to use our language to specify their science priorities for a particular site. The priorities were then passed on to a human operator who was given fixed resource constraints (downlink bandwidth and distance traveled) and asked to choose rover actions so as to return the best possible data set according to the specified priorities. The idea is that the human operator simulates an onboard SA system in the intelligent site survey operational mode—the operator can freely examine all incoming rover data and can send commands in real time, but only a small subset of the data can be downloaded at the end of the run. human-onboard performance was compared to a simple periodic sampling strategy, and evaluated according to the specified priority value function. The human-onboard test was intended to achieve several goals:

1. *Priority language:* It should provide feedback from scientists as to whether our priority language is sufficiently usable and expressive. Also, any ambiguities in the language should be exposed by the human operator or during post-experiment performance evaluation.
2. *Sampling strategy:* It should inform us about how a person constrained to use rover sensing makes decisions about science actions. This information can guide SA software design and suggest changes to the rover configuration (e.g., camera position and field of view).
3. *Performance bound:* The human operator should perform better than any SA system that can be developed in the near term, so evaluating the human-onboard performance might be considered an upper bound on active science autonomy performance—with the caveat that performance is strongly affected by the specifics of our experiment, such as the selected field site and characteristics of the rover.

Our original intent was to allow the human operator to make decisions about sensing in the local area around the rover. For instance, using a SPI image to pick out an individual rock that matches a target signature, the operator could command a follow-up spectrometer reading using the pan/tilt unit to point the spectrometer, or the rover could be commanded to roll over the rock and deploy the FI. Unfortunately, the onboard software for precision pointing and motion was not ready to handle these tasks. Also, time constraints on the availability of the rover forced us to keep the operator's task as simple as possible.

As a result, we restricted the operator to make fewer decisions at a larger scale. Motion commands were constrained to work at a coarse resolution (on the order of 5 meters minimum distance), and sampling was reduced to a binary decision: at each rover location, collect either no data or a fixed set of data products. A preview command was also provided—this enabled the operator to quickly get a sense of what was present at a location and decide whether a sample there was justified.

We performed three human-onboard runs, designated A, B, and C. For run A, we asked the scientists to provide target signatures only, meaning no novelty detection or representative sampling signature templates were used. For runs B and C, we went to the other extreme, using *only* a single signature template that assigned a uniform priority to each new feature class as it was discovered. Fig. 3 provides basic statistics about the runs, and fig. 4 shows the rover's path and actions during an example run. For each run, the human operator was nominally allocated 300 meters of total rover travel and 6 samples, but all three runs were ended early due to time pressure.

Run	*Distance Traveled (m)*	*# Previews*	*# Samples*
A	200	8	2
B	226	8	4
C	260	9	4

Figure 3. Summary of experimental runs.

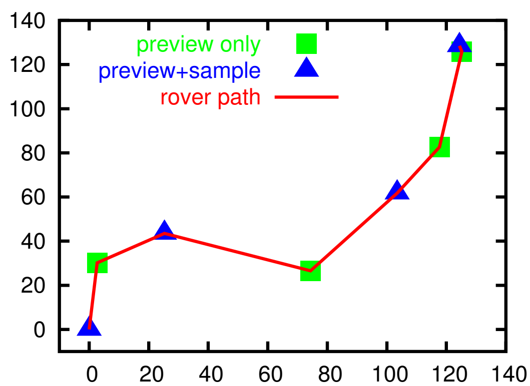

Figure 4. Rover path/actions for run B (distances in meters).

In order to provide a performance comparison, we also conducted a 300 meter periodic sampling transect at each human-onboard study site. For each periodic transect the rover was initially located at the same starting point used for the human-onboard run and pointed in an arbitrary direction that appeared traversable. From there the rover traveled forward in a straight line, taking a total of 6 samples evenly spaced at 60 meter intervals.

5. Experimental Results

The human-onboard experiment achieved the first two of its three goals. We obtained feedback from scientists about the usability of the priority language, and we gained a better understanding of human operator sampling strategy for this problem. However, the small number of runs and irregularities in the experimental protocol prevent us from making a meaningful quantitative comparison between human-onboard operation and periodic sampling. As a result, we focus on qualitative observations.

Scientist Priority Specification

Fig. 5 shows the rock target signatures specified for run A by our science team. The scientists generated the signatures in a discussion facilitated by one of us (Thompson). Overall they reported that the language made sense, and they were able to construct a fairly rich set of priorities. However, their difficulties with certain parts of the process are instructive.

First, the scientists expressed uncertainty about how to set the $\mathrm{nthreshold}$ and $\mathrm{qthreshold}$ parameters. Although they clearly understood the objective of allocating bandwidth, they were not comfortable selecting numerical values. In fact, for the $\mathrm{qthreshold}$ parameter, they refused altogether. Instead of absolute quantity thresholds they decided to specify relative thresholds, e.g., "We want the same quantity q of both of these signatures." It is unclear whether this is because relative thresholds are in fact a more natural representation, or because insufficient explanation of $\mathrm{qthreshold}$ led to misunderstandings during the discussion.

Second, the angularity attribute had not been formally defined before the discussion, so one of the signatures specifies "high" angularity without a numerical threshold value. This problem is clearly due to ambiguity in the task presented to the scientists.

Finally, during the priority specification process, the scientists had at their disposal substantial amounts of science data collected earlier at desert sites similar to the human-onboard sites. They could have used example features from the prior data in two ways: either by specifying example-based classes, or by measuring the properties of example features and using that information to manually tune threshold attribute values.

In fact, they elected to do neither. One of their target signatures was intended to capture light-colored quartz rocks that sometimes host photosynthetic endoliths. They had identified such rocks in earlier data sets, but they did not reexamine the prior data when selecting an albedo threshold of 0.9. Unfortunately, as far as we can determine there were no rocks that bright anywhere in the area. In a real SA application, this problem would be especially pernicious—if the scientists specify a class too strictly, the rover might pass over the very borderline features that would allow them to realize their mistake. This observation bears out our argument that use of *in situ* examples provides more robust classification.

Our overall conclusions from the scientist interaction are that (1) the priority language we used is a reasonable first step, (2) it is difficult to pin down usability problems because we did not invest enough time documenting the language and training the science team before the experiment, and (3) selecting numerical thresholds was a major source of problems that may be partially correctible through use of example-based classes.

Human-onboard Sampling Strategy

Human-onboard sampling seemed qualitatively far less effective than we anticipated before the experiment. Our analysis explains the problems we encountered and the discusses the extent to which they would also apply in an actual SA imple-

Rock attributes	*density*	*reward*	*nthresh*	*qthresh*
Size > 50 cm	> 1 rock	6	10	q
Size from 15-50 cm	> 1 rock	2	10	q
Angularity high	$> 50\%$	8	10	q
Albedo > 0.9	$> 5\%$	10	10	$2q$

Figure 5. Rock target signatures for run A.

mentation.

First, the preview images did not really provide the context needed to guide the operator. A good set of preview images should have wide coverage and high resolution, and it should be possible to collect it quickly. In practice our sensor configuration could support any two of these goals but not all three. If we had chosen to use the SPI, its narrow field of view would have made it necessary to assemble an image mosaic in order to get wide enough coverage, but that process would have been too time consuming. We elected to use the NavCams instead. Each preview consisted of three images from the left NavCam, pointed at azimuth -15^o, 0^o, and $+15^o$ so that they overlapped to cover a 90^o wedge in front of the rover. Fig. 6 shows an example image (the dark rock in the foreground measures approximately 10 cm).

Figure 6. Example preview image.

Unfortunately, the NavCam images suffered from two problems. First, we were forced to grab the images at 320×240 resolution in order to retain compatibility with the rover's stereo obstacle avoidance software. This low resolution over a 60^o field of view made interpretation difficult, particularly for terrain at distances of 5 meters and beyond. Second, the elevation angle of the NavCams was optimized for local obstacle avoidance, so they were not normally able to see terrain beyond about 10 meters distance. The result was that the operator was reduced to moving around more or less at random and using the preview data only to decide whether a sample at the current location was worthwhile. But note that this problem depends on the scale at which we were selecting samples—the area covered by the NavCam images would have been ideal if the task were to select local features for targeted spectrometer readings. This is not an isolated problem—sensing configuration issues similar to what we observed have also been identified in earlier SA systems [1].

Another problem was the time consumed by repeatedly stopping to grab a preview and decide on the next command, even using the high-speed NavCam-based preview approach. We define the *speed penalty* of an SA system to be the ratio of the average rover speed under periodic sampling to the average speed under SA. We were surprised to note that the speed penalty for all three of our human-onboard runs was greater than 2; we had not really attempted a speed analysis before the experiment. In a real application, a performance gap that large has the potential to more than offset any gains in the quality of data return when using active SA.

We conclude that the effectiveness of SA decisions is strongly affected by its sensing configuration, which determines the viewing area it covers, the size of resolvable features, and the rate at which data can be collected. The sensing requirements depend on the nature of the SA task (e.g., precision targeting of small features or large-scale motion decisions), and the speed penalty needs to be considered. The overall message is that requirements specific to SA should be considered early in the rover design process— retro-fitting an SA system onto a pre-existing sensing platform is likely to be problematic.

However, the good news is that modeling SA sensing requirements is very similar to modeling navigation sensing requirements, a topic that has received considerable attention [6], [7]. In each case, the overall system performance is affected both by inherent limits on the sensing technology and by computational limits on the rate at which data can be processed. In the future we hope to take advantage of this connection to produce more effective SA system designs.

6. CONCLUSIONS

This paper sketched a broad conceptual framework for science autonomy and analyzed new operational modes in terms of potential benefits and technical challenges. We introduced a language for specifying science priorities, and applied the language in a rover field experiment with a team of scientists. Although results to date are only qualitative, we made several important observations that will guide our ongoing design of a full SA system.

7. ACKNOWLEDGMENTS

The authors thank the Zoë operations team for supporting our rover experiments, and the science team for specifying priorities. Discussions with James Dohm and Kimberly Warren-Rhodes were especially helpful. This research was supported in part by NASA under grants NNG0-4GB66G and NAG5-12890 with program executive Michael Meyer.

REFERENCES

[1] Dimitrios Apostolopoulos, Liam Pedersen, Benjamin Shamah, Kimberly Shillcutt, Michael D Wagner, and William Red L. Whittaker. Robotic antarctic meteorite search: Outcomes. In *Proc. of IEEE International Conference on Robotics and Automation*, pages 4174–4179, May 2001.

[2] N. A. Cabrol, G. Chong-Diaz, C. R. Stoker, V. C. Gulick, R. Landheim, P. Lee, T. L. Roush, A. P. Zent, C. Herrera Lameli, A. Jensen Iglesia, M. Pereira Arrerondo, J. M. Dohm, R. Keaten, D. Wettergreen, M. Sims, L. Pedersen, A. Bettis, G. Thomas, and B. Witzke. Nomad rover field experiment, Atacama Desert, Chile, 1, science results overview. *J. Geophys. Res.*, 106(E4):7785, 2001.

[3] R. Castaño, R. C. Anderson, T. Estlin, D. DeCoste, F. Fisher, D. Gaines, D. Mazzoni, and M. Judd. Rover traverse science for increased mission science return. In *Proc. of IEEE Aerospace Conference*, Big Sky, Montana, March 2003.

[4] Tara A. Estlin, Tobias P. Mann, Alexander G. Gray, Gregg Rabideau, Rebecca Castano, Steve Chien, and Eric D. Mjolsness. An integrated system for multi-rover scientific exploration. In *Proc. of AAAI*, 1999.

[5] K. E. Herkenhoff, S. W. Squyres, R. Arvidson, D. S. Bass, III J. F. Bell, P. Bertelsen, N. A. Cabrol, L. Gaddis, A. G. Hayes, S. F. Hviid, J. R. Johnson, K. M. Kinch, M. B. Madsen, J. N. Maki, S. M. McLennan, H. Y. McSween, Jr. J. W. Rice, M. Sims, P. H. Smith, L. A. Soderblom, N. Spanovich, R. Sullivan, and A. Wang. Textures of the soils and rocks at Gusev Crater from Spirit's microscopic imager. *Science*, 305(5685):824–826, 2004.

[6] A. Kelly and A. Stentz. Rough terrain autonomous mobility - part 1: A theoretical analysis of requirements. *Autonomous Robots*, 5:129–161, 1998.

[7] In So Kweon. *Modeling Rugged Terrain by Mobile Robots with Multiple Sensors*. PhD thesis, Robotics Institute, Carnegie Mellon University, 1990.

[8] L. Pedersen. *Robotic Rock Classification and Autonomous Exploration (tech. report CMU-RI-TR-01-14)*. PhD thesis, Robotics Institute, Carnegie Mellon University, November 2000.

[9] Geb Thomas, Mark Reagan, A. E. Bettis, N. Cabrol, and A. Rathe. Analysis of science team activities during the 1999 Marsokhod rover field experiment: Implications for automated planetary surface exploration. *J. Geophys. Res.*, 106(E4):7775, 2001.

[10] D. Thompson, S. Niekum, T. Smith, and D. Wettergreen. Automatic detection and classification of features of geologic interest. In *Proc. of IEEE Aerospace Conference*, Big Sky, Montana, 2004.

***Trey Smith** is a Ph.D. student at the Carnegie Mellon University Robotics Institute, studying under a GSRP Fellowship with NASA Ames Research Center. His research interests include probabilistic planning, rover navigation and science operations, and multi-robot systems. He received B.S. degrees in computer science and math from Carnegie Mellon in 1999.*

***David Thompson** is a Ph.D student at the Carnegie Mellon University Robotics Institute. His research interests include machine learning and its applications in exploration robotics. He received a B.A. in computer science from Carleton College in 2001 and an M.Sc. in informatics from the University of Edinburgh in 2002.*

***Scott Niekum** is an undergraduate computer science and cognitive science student at Carnegie Mellon University. His research interests include computer vision, machine learning, parallel probabilistic problem solving, and efficient knowledge representations. He expects to graduate in May 2005 and hopes to continue on to a computer science Ph.D program.*

***David Wettergreen** is an Associate Research Professor at Carnegie Mellon University Robotics Institute. His research in space robotics advances capabilities for autonomous navigation and exploration. A current focus is enabling robots to reason about science objectives and to modify behavior based on science observations. Dr. Wettergreen is the Principle Investigator for the ASTEP Science on the Fly project.*