A Field Demonstration of an Instrument Performing Automatic Classification of Geologic Surfaces

Dmitriy L. Bekker *   David R. Thompson*   William J. Abbey*

Nathalie A. Cabrol †   Raymond Francis ‡   Ken S. Manatt*
Kevin F. Ortega*   Kiri L. Wagstaff*

March 28, 2014

Abstract

This work presents a method to automate simple aspects of geologic image analysis during space exploration. Automated image analysis onboard the spacecraft can make operations more efficient by generating compressed maps of long traverses for summary downlink. It can also enable immediate automatic responses to science targets of opportunity, improving the quality of targeted measurements collected with each command cycle. In addition, automated analyses on Earth can process large image catalogs, such as the growing database of Mars surface images, permitting more timely and quantitative summaries that inform tactical mission operations. We present TextureCam, a new instrument incorporating real time image analysis to produce texture-sensitive classifications of geologic surfaces in mesoscale scenes. A series of tests at the Cima Volcanic Fields in the Mojave Desert, California demonstrate mesoscale surficial mapping at two distinct sites of geologic interest.

1   Introduction

To date, ongoing Mars rover missions have accumulated an archive of just under 380,000 microscopic and macroscopic images: 309,680 over 10 years for the Mars Exploration Rover mission (MER) and 68,131 for the Mars Science Laboratory (MSL) in a year. This is the result of increased mission survivability, mobility, and autonomy. It can be assumed that archives will grow even larger in future missions due to continuing

*Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA.

†NASA Ames Research Center / SETI Institute, Moffett Field, CA 94035, USA.

‡Centre for Planetary Science and Exploration (CPSX), Western University, London, Ontario, Canada.

progress in technology and instrumentation [Bakambu et al., 2012, Furgale and Barfoot, 2010], as well as in science operation productivity [Deans et al., 2012]. Standard now on Mars rovers [Squyres et al., 2003, Grotzinger et al., 2012], and planned for the Mars 2020 mission [Mustard et al., 2013], multi-scale imaging allows a science team to generate powerful nested imagery to characterize morphological, mineralogical, and compositional domains to sub-meter resolution, while simultaneously mapping them on a scale of tens of kilometers. However, analysis bottlenecks, such as limited downlink bandwidth and limited analyst time, can prevent missions from fully utilizing these data for tactical operations planning. The present work describes image analysis algorithms and instrument hardware that can automate rudimentary aspects of geologic analysis both on the ground and onboard spacecraft. Simple automatic visual processing of these images can improve the efficiency of mission operations as well as the ultimate science yield.

A primary challenge for Mars surface mission operations is the limited time the science team can afford to spend daily on data analysis. The team must quickly analyze data downlinks, while simultaneously deciding on the next science targets and designing traverse plans - all within the handful of hours before the next uplink window. Both targets and traverses are chosen based on the highest priorities of the science team. In most cases, a science team will preferentially collect and analyze the data that best support the primary mission science objectives (e.g., for high impact publications and communications), and will only later have time for an in-depth review of the archived data. As a result, it often takes many years for data archives to be fully investigated. Some observations may never be fully analyzed, creating a potential gap in mission productivity. The earlier and more rigorous the quantitative analysis, the better informed later exploration decisions can be. Timely analysis is especially critical for rover missions, since there is little chance of revisiting sites once the mission is over. Further, missions rarely return to previously investigated targets during mission operations, due to a general perception that a mission only makes progress if it goes forward — perhaps to explore as many diverse environments as possible in a limited, unknown mission lifespan. This is, so far, a clear distinction between robotic exploration and terrestrial fieldwork, where revisiting sites and resampling are not only common but often critical.

In such context, the use of efficient autonomous analytical image techniques can increase the scientific impact of large archives during operations. It could be particularly valuable for a mission such as a Mars 2020 mission, a central goal of which will be to seek signs of early life on Mars and to identify a high-value subset of samples to prepare for a later Sample Return mission. More than for any previous mission, this objective requires (a) an especially careful contextual mapping of large traversed areas in order to understand the stratigraphic context of samples and discoveries; (b) the ability to investigate and interpret as critical or non-critical numerous sedimentary landforms and their remnants (from residual mesoscale features to rocks and particulates), whether layered or non-layered, and in situ or not; (c) the capability to identify textural groups that may bear evidence of potential fossil biogenic constructs from others which may not; and (d) the analytical power to investigate the textural, morphological, and compositional elements of a scene that can provide information on the energy of transportation and on depositional environments in the explored region.

Automated image analysis can also improve exploration efficiency by performing simple data interpretation onboard the spacecraft [Gulick et al., 2001, Woods et al., 2009, Woods et al., 2011]. Missions such as the Mars 2020 mission may spend a considerable fraction of the mission performing long drives and reconnaissance, passing through terrain which has only been seen previously from orbit. It is important that missions retain the flexibility to respond to targets of opportunity, deploying targeted remote sensing or contact instruments as needed without stalling the mission or incurring the delay of an extra command cycle. Onboard image analysis allows the spacecraft to respond immediately to these unanticipated targets, for example by directing high resolution cameras or targeted spectroscopy at novel rocks or stratigraphic features [McGuire et al., 2010]. This idea was the core of the AEGIS system, which provided computer vision for automatic science feature selection by the Mars Exploration Rover Opportunity [Estlin et al., 2012]. A simple visual analysis can also allow more effective targeted measurements upon arrival at a destination outcrop. Specifically, automatic mesoscale maps can indicate the principal units of material for a representative remote measurement. The same analysis can also show good locations for contact sensing that are free of dust or fractures. These simple image analyses can save critical command cycles and latency required for Earth operators to micromanage each approach, surface evaluation, and placement maneuver. Finally, surface classification products can be downlinked to provide information about terrain the rover has passed when bandwidth constraints prevent downlinking complete images. In these ways, automated image analysis offers a path to address key operations bottlenecks and improve the science return of the mission over its finite operational lifetime.

The following sections present TextureCam, an instrument that incorporates built-in automatic image analysis. The techniques we describe can improve the analytical power and decision making of future missions by allowing (1) surficial mapping over much larger area than currently possible, including maps produced for tactical decisions during science operations, and (2) onboard automated mapping that will make autonomous planetary rovers more efficient. To address the second goal we have built this capability into embedded sensor hardware. The result is a camera system that can classify surfaces in real time - a *smart instrument* that we demonstrated in field trials at relevant geologic sites in the Cima Volcanic Fields of the Mojave Desert, CA. These tests show the feasibility of real-time image analysis and provide a rough picture of expected system performance for some simple surface classification tasks. This work focuses on real-time image analysis in mesoscale scenes, which we define as containing features generally greater than 64 mm in size (anything smaller falls into "gravel" or "sediment" category). Automatic classification of rocks on the microscopic scale has also shown good results [Mlynarczuk et al., 2013].

2   Approach

This section describes our technical approach, beginning with a brief description of the image analysis algorithm and then explaining our instrument hardware implementation.

2.1 Algorithm

We formulate the problem of geologic image analysis as *pixel classification* in which the system categorizes image regions based on predefined labels. This does not capture all geologic image analysis tasks, but we find it to be a general and flexible formulation that can accommodate many or most of the stated objectives. The pixel categories may represent geologic content such as layering [Thompson et al., 2011, Wagstaff et al., 2013], rock surfaces [Wagstaff et al., 2013, Francis et al., 2014], and particle sizes. Alternatively they could involve non-geologic attributes like illumination and shadow [Foil et al., 2013] or operations objectives such as identifying physical surfaces in good condition for contact sensing [Wagstaff et al., 2013]. This formulation is therefore broad enough to encompass a wide range of autonomous activities and image analyses in support of long-range traverse science. For example, simple post-processing applied to pixel maps has been used for selective data return [Thompson et al., 2008], target selection for opportunistic instrument deployment [Estlin et al., 2012], and region-of-interest compression [Wagstaff et al., 2004].

Formally, the class of a pixel is represented by an unknown variable y. Given a vector of image attributes $\mathbf{x}$ in the vicinity of a pixel, the goal of the analysis is to determine the probability distribution of classes, $P(y|\mathbf{x})$. Attributes could be raw pixel values, average intensities of image subregions or even the responses of a filter bank [Shotton et al., 2008]. Here, we wish to capture local spatial context and texture, so we form attributes based on the statistics of small image patches. Specifically we apply simple arithmetic operations to pixels in the local neighborhood of the pixel to be classified. Similar to work by [Shotton et al., 2008], each attribute is the result of an operation applied to two *test locations* at given spatial positions relative to the *classified pixel*. In this work, the window is 31 pixels wide. Legal operations include the sum, difference, absolute difference, and ratio. Many thousands of attributes are possible using this recipe, and together they provide a rich description of the visible and spectral appearance of a small image patch. The goal of the analysis is to determine the posterior class probability given these features. Applying this process to each location in the image produces a map of the geologic surfaces in the scene.

We use a statistical machine learning approach, estimating $P(y|\mathbf{x})$ from a representative set of training images using a *random forest* model [Breiman, 2001, Shotton et al., 2008]. The random forest is a collection of independent weaker models known as *decision trees*. Each decision tree is trained using a separate random subset of data and produces a separate classification solution. After all decision trees have given an answer for the pixel, these answers are averaged to form the final classification prediction.

Each decision tree is represented as a branching binary graph of nodes (Figure 1). Each node performs a threshold test on a specific attribute $\mathbf{x}_i \in \mathbb{R}$ of the pixel to be classified. These attributes are generally simple arithmetic operations applied to two other pixels at specific relative locations. For example, an attribute might be a sum or ratio of two locations within a small rectangular neighborhood. One compares the result of this operation to a predetermined threshold τ. Based on the result, control then descends left or right down the decision tree. The process repeats at each new node, ultimately leading to a classification decision at a leaf node. Since the decision

is a composite of many attribute tests, the tree can model the complex visible structure and texture of the image patch.

Each tree finds a different solution. The majority classification from all trees forms the overall class prediction. Additionally, pixel posterior class distributions are stored at each leaf, so that the system can produce posterior classification probabilities $P(y|\mathbf{x})$. This indicates the system's confidence in the classification result, and can be used to exclude low-confidence or ambiguous regions. It can also flag anomalous areas that are different from materials seen in the training data [Foil et al., 2013].

The training procedure requires an analyst to supply labeled data in the form of designated image regions for each class. One strategy is to use an image editing program to "paint" over the pixels, where different colors signify different surface classes. It is not necessary that the entire image be labeled; usually a few representative exemplars are sufficient, and the entire process requires just a few minutes. We can duplicate the training images and labels using different rotational transformations to grow the training set and improve the invariance of the result to orientation. Then, an automated decision forest learning procedure assigns a subset of training pixels to each decision tree and grows each tree from the root down. At each iteration, the training procedure defines the test at each node to be the one which splits the remaining dataset into the most well separated populations, with respect to the pixel labels. The specific splitting criterion we employed maximizes information gain with respect to the posterior class distribution. In this study, we used a single labeled image in our training episodes, thus demonstrating the performance of the algorithm with a small training set.

The random forest model has several beneficial properties. First, each tree's calculation is independent so the algorithm is highly scalable, and degrades gracefully with the number of trees and the number of nodes per tree. This provides a convenient means for designers to trade between execution speed and accuracy. Second, each tree requires just a few arithmetic operations so the algorithm can be made quite fast. Third, training the trees on separate random subsets of the data promotes good generalization performance, because it avoids the idiosyncrasies of any single subset. We did not experience any problems with local minima, and classifier performance was robust across a wide range of training parameters. For further details on the forest training process, we refer the reader to previous work on decision trees [Breiman, 2001, Shotton et al., 2008, Wagstaff et al., 2013].

2.2 Instrument Architecture

The physical system instantiates this algorithm in a hardware testbed capable of real time automatic surface mapping. The TextureCam instrument consists of three main components: the cameras, the data processing unit (DPU), and a computer (Figure 2). The cameras and the computer are unmodified commercial products, while the DPU is based on a commercial field programmable gate array (FPGA) development board with custom-built interface electronics and packaging.

We use two 3CCD Hitachi HV-F202SCL color cameras (the second camera will be integrated in the future). Each camera has three internal image sensors with filters for measuring red, green, and blue wavelengths. This configuration results in images with better precision than cameras in a 1CCD configuration as each pixel is a true sense of
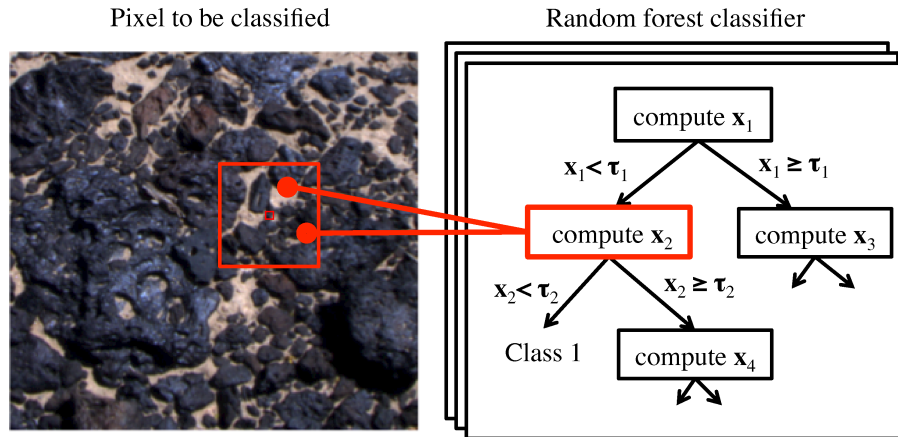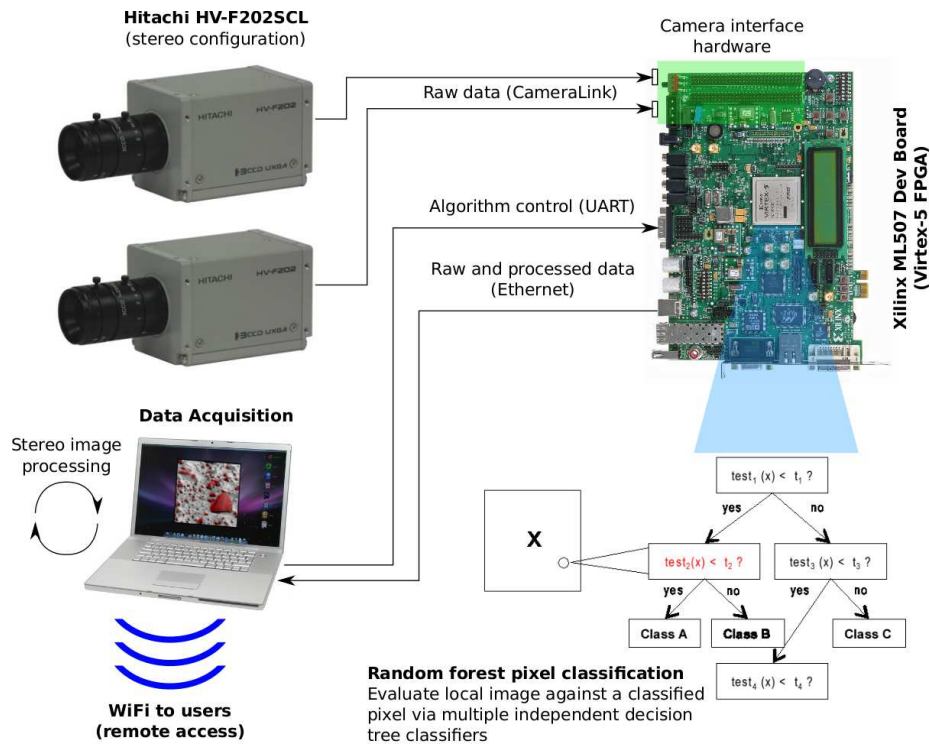
Pixel to be classified

Random forest classifier

compute $\mathbf{x}_1$

$\mathbf{x}_1 < \boldsymbol{\tau}_1$ $\mathbf{x}_1 \geq \boldsymbol{\tau}_1$

compute $\mathbf{x}_2$ compute $\mathbf{x}_3$

$\mathbf{x}_2 < \boldsymbol{\tau}_2$ $\mathbf{x}_2 \geq \boldsymbol{\tau}_2$

Class 1   compute $\mathbf{x}_4$

FIG. 1. Pixel classification algorithm concept

Hitachi **HV-F202SCL**
(stereo configuration)

Camera interface
hardware

Raw data (CameraLink)

Algorithm control (UART)

Raw and processed data
(Ethernet)

Xilinx ML507 Dev Board
(Virtex-5 FPGA)

Data Acquisition

Stereo image
processing

$test_1(x) < t_1$?

yes   no

X

$test_2(x) < t_2$? $test_3(x) < t_3$?

yes   no   yes   no

Class A   **Class B**   Class C

$test_4(x) < t_4$?

WiFi to users
(remote access)

Random forest pixel classification
Evaluate local image against a classified
pixel via multiple independent decision
tree classifiers

FIG. 2. TextureCam instrument concept. During the 2013 field test, we used only one camera; the second camera (for stereo vision) will be integrated for our next field test in 2014.
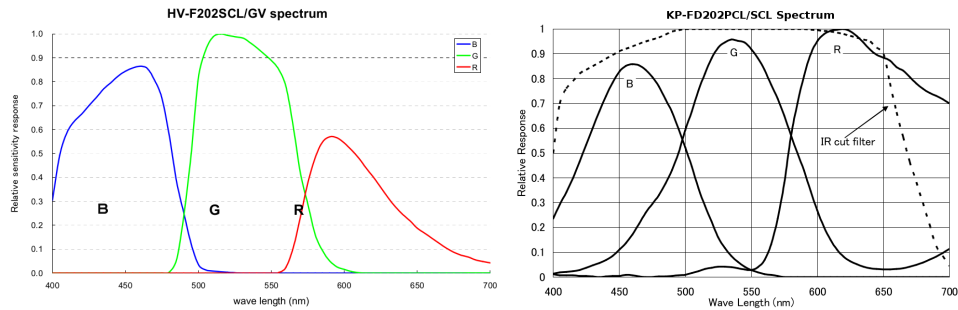
FIG. 3. Spectral response of the Hitachi 3CCD HV-F202SCL camera (left) as compared to a similar 1CCD camera, the KP-FD202SCL (right). The individual color bands have sharper cut offs in the 3CCD configuration [Hitachi, 2013].

its corresponding primary color. The high-fidelity images ensure the algorithm receives a very accurate representation of the scene to be classified. In contrast, a 1CCD camera has a single image detector with, for example, a Bayer mosaic filter. Images recorded in this configuration need to be passed through a demosaicing algorithm to extrapolate the red, green, and blue values. The spectral response of the Bayer mosaic filter typically shows greater overlap in the color bands, resulting in more cross talk between the pixels and a lower dynamic range than a 3CCD camera (Figure 3).

The cameras connect to the DPU via a Mini Camera Link interface, a high speed, serialized, low-voltage differential signaling (LVDS) standard commonly used in machine vision applications. This interface adapter is attached to a Xilinx ML507 FPGA Development Board. The data acquisition logic, the command/control of the cameras, and the random forest classifier are implemented on a Virtex-5 FPGA[1]. Our initial prototype performs the classification using an embedded Power-PC processor (inside the FPGA), but the random forest is well suited for an efficient logic-only implementation since it is highly parallelizable: each pixel can be classified independently of all others in the image. Time-sensitive portions of the algorithm will be ported to the FPGA logic for the next field test. The goal is to achieve a classification output rate of 1 frame per second.

The output from the DPU is a data stream with both raw images and the classification results. This is sent over Ethernet as UDP packets in a direct connection to a computer which captures, displays, and stores this data. A timestamp along with GPS coordinates is also recorded. While the first field test only utilized one camera, the next expedition will add a second camera and stereo image processing (performed on the computer). A WiFi network is used to connect field users to the computer in order to control the instrument, view raw and processed data, as well as perform training for new classifiers. This multi-user collaborative environment has proven to be very efficient in our first field test.

---

[1]This FPGA was selected because a radiation-hardened version in the same family is available, which will ease the transition to a flight instrument implementation for a future planetary rover.

FIG. 4. The TextureCam single-camera field instrument consists of a tripod-mounted camera with sunshade (upper inset) and a backpack-mounted DPU and control computer (lower inset).

2.3 Field Mobility

We aim to evaluate this system in a range of analogue field environments, which demands a portable, rugged and fully self-sufficient prototype. To ease in transportation and for increased field mobility, the instrument is mounted on a pair of hard-frame backpacks (Figure 4). One backpack holds a gasoline generator, providing power for all of the electronics and user laptops. The other backpack holds the DPU, computer, cabling, cameras, and spare gasoline canisters. Once a field site is selected, the backpacks are placed on the ground, the generator activated, connections are made to the instrument, and the camera is placed on a tripod. The whole process takes about five minutes. An aluminum shield with a solar-powered fan protects the camera from overheating in direct sunlight, and an umbrella provides a fashionable dust cover and shade for the instrument electronics.

2.4 User Interface

All control of the TextureCam instrument is handled via a graphical user interface running on the computer and visible to all field users over a WiFi network connection (Figure 5). The interface allows the user to select which classifier to use from a set of classifiers pre-loaded on a memory card in the DPU. Commanding the instrument to acquire an image instructs the DPU to record a snapshot from the cameras, run the
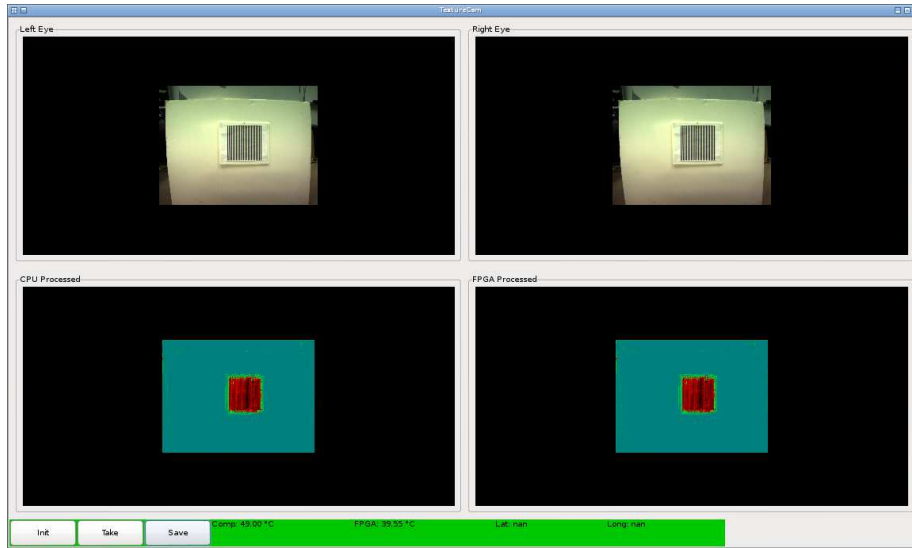
FIG. 5. The instrument is controlled through a graphical user interface over a wireless connection. The top row in the interface shows the raw image from the cameras looking at a striped training target. The bottom row shows the CPU-processed result (left) and FPGA-processed result (right). The system recognizes the striped texture of the training target, with warmer colors corresponding to a higher probability for that class. A control bar, with buttons and indicators for temperature and GPS coordinates, is present at the bottom of the frame.

classifier on the raw image, and send the raw and processed image to the computer, which displays them in the user interface window. The computer (CPU) also runs the classification processing on the received raw image, allowing for an immediate visual comparison between CPU processing and FPGA processing. The "save" command instructs the DPU to transmit the probability maps and the computer to save acquired (and self-processed) image data, as well as to record the timestamp and GPS coordinates of the current location. The temperature of the DPU and the computer is also displayed, prompting the user to shut down the instrument if operating conditions become too hot for the electronics.

3 Field Experiment Methods

We deployed our instrument prototype on a field expedition in the second year of the project. While the algorithm development and the bulk of the algorithm refinement occurred in a laboratory setting (with artificial geological and lighting conditions), there was no better test of performance and operational logistics than in a relevant environment as may be seen on a future planetary rover. We've identified a number of implementation bugs and design improvements that would have been hard to uncover
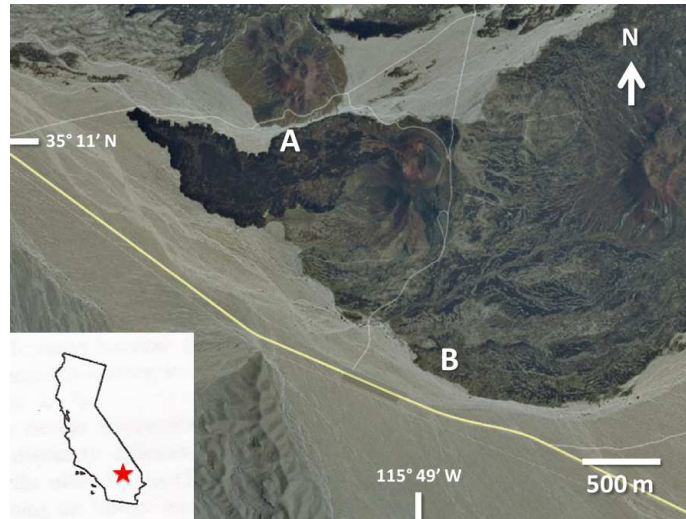
FIG. 6. Southern edge of the Cima Volcanic Field. Located ˜25 km east-southeast of Baker, CA in the Mojave National Preserve. Individual experiments were performed at: (A) a north-facing wall, cut by erosion, exposing multiple volcanic layers; and (B) the upper surface of the southern-most lava flow. Image modified after MapQuest.

in the laboratory.

3.1 Site

Our field tests took place at the southern edge of the Cima Volcanic Field (Figure 6). This field is composed of a series of basaltic lava flows that lie in the Mojave Desert region of Southern California, ranging in age from Late Miocene (6-7 Ma) in the north to Late Pleistocene (20-500 ka) in the south [Dohrenwend et al., 1984, Wells et al., 1995, Wilshire, 1992]. This entire region, averaging less than 1.5 cm of precipitation a year, has been declared a Mars analogue site by the Terrestrial Analogs Panel of the NSF-sponsored Decadal Survey [Farr, 2004]. More specifically, Cima was selected for this particular study due to the great diversity of surfaces observed in these volcanics, including not only textures relevant to basaltic rocks, but also those applicable to unconsolidated regolith and stratified sediments [McFadden et al., 1987, Wells et al., 1985]. Individual field experiments to evaluate the real time performance of the system were designed for the following locations: (A) a highly eroded wash containing a north-facing wall that exposes multiple volcanic events interbedded with immature paleosol layers (Figure 7) and (B) the upper surface of a volcanic lava flow that has developed a mature desert pavement interspersed with patches of clay and loose volcanic clasts (Figure 8). The two sites were labeled as Site A and Site B, respectively. The field experiment took place from April 29 to May 2, 2013, under clear skies and good mid-day illumination.

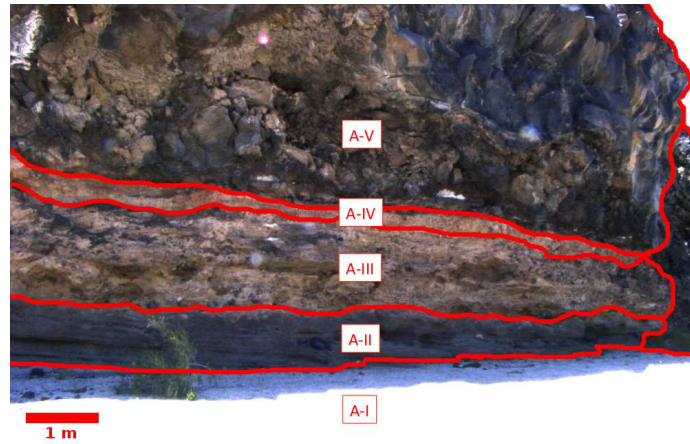We defined visually distinct textural classes in each scene. Site A exhibited at

FIG. 7. Layered units at Site A. I: Channel Bed. II: Sorted volcanic gravel and lava bombs. III: Volcanic angular gravel and blocks in poorly-sorted clay matrix. IV: Oxidized clay. V: Massive volcanic blocks.
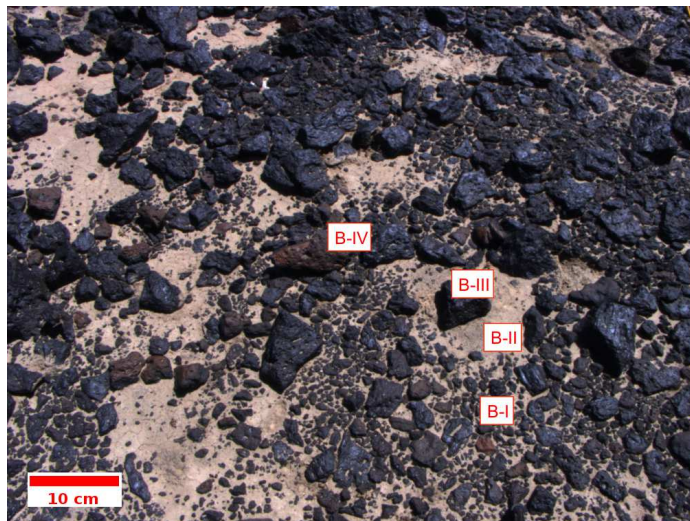

FIG. 8. Surfaces at Site B. I: Gravel and small clasts (< 64 mm in size). II: Clay sediment. III: Basalt cobbles. IV: Basalt cobbles with iron oxide coating.
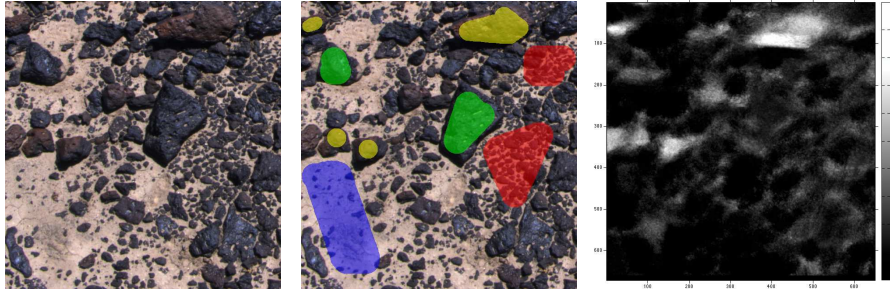
FIG. 9. (Left, Center) Training labels for Site B. Here we trained the classifier using a single image. Red: B-I gravel (< 64 mm). Blue: B-II sediment. Green: B-III basalt cobbles. Yellow: B-IV basalt cobbles with iron oxide coating. (Right) the inferred probability map for surface class B-IV, basalt cobbles with iron oxide coatings, produced automatically by the model after training. Higher intensity corresponds to a higher probability that the surface belongs to the iron oxide class.

least five distinct layers (Figure 7): (A-I) the channel bed, consisting primarily of bright sand; (A-II) a layer of volcanic angular gravel, well sorted and punctuated by locally embedded lava bombs; (A-III) poorly sorted, angular volcanic gravel, including boulder-sized clasts, suspended in a clay matrix; (A-IV) a thin and somewhat intermittent layer of red oxidized clay; and (A-V) a layer of massive basalt material. We also delineated several visually distinctive surface classes at Site B: (B-I) regions with predominant gravel-sized particles (< 64 mm in size), including local patches of desert pavement; (B-II) surface covered by clay sediment; (B-III) basalt cobbles (> 64 mm in size); and (B-IV) basalt cobbles with a red oxidized surface indicating differential weathering or exposure to moisture.

3.2 Data Acquisition

We began each session by acquiring a single three-channel image and manually labeling examples of different classes in an image editing application (Figure 9). This process produced a map of integer values delineating classified and unclassified areas. We then trained the random forest on a field laptop computer, building the decision tree from these training examples. This process typically required 10 to 15 minutes, with the majority of the time used for manual image labeling. At the end of the training process we uploaded the resulting random forest definition to a flash memory card and inserted it into the instrument; all subsequent classifications used this trained model. Contrary to some previous studies [Peralta et al., 2008] we did not experience significant problems from changing light conditions, and a single training session typically provided consistent performance for the remainder of the trial. That we experienced fewer illumination issues is understandable since our study differed from the Calderón work in several significant ways. First, the Calderón study was conducted near the winter solstice, with very low and fast-changing light angles. In contrast we operated at favorable times of the day when the light remained relatively constant for the tar-

gets during acquisition of both training and test images. Second, that previous work focused on the problem of tracking individual features using Scale Invariant Feature Transform (SIFT) descriptors, a challenging problem that relies on perfect matches from one scene to the next. Tracking a specific feature is an all-or-nothing problem, and a changing shadow can cause a lost track that cannot be recovered thereafter. SIFT matching is known to be somewhat sensitive to lighting changes [Bauer et al., 2007]. Finally, we provided some additional illumination invariance by preprocessing each image and converting it to a Hue Saturation Value (HSV) color representation.

After training, we acquired a sequence of images to cover the area of interest. For Site A, we began approximately 12 meters from the exposed wall, with the camera at 0 degrees elevation (i.e., pointed at the horizon) We collected a single training example from the extreme right side of the outcrop. We then panned the camera leftward, collecting a sequence of four images of the entire visible face of the wall and forming a mock rover panorama spanning approximately 120 degrees of azimuth. At Site B we began with a training image with the tripod-mounted camera pointed down at approximately -45 degree elevation, observing a patch of terrain approximately 1 m wide. We then moved the camera to the transect location, where we acquired a sequence of 16 images at 1 meter intervals with the camera facing the direction of motion. All image processing occurred onboard the instrument; images were also saved to the instrument computer for later analysis.

We evaluated the system's performance by comparing the automatic classifications to ground truth classifications made by a human. We generated these ground truth labels in the same fashion as the training data, by manually labeling all regions of the test images according to their surface type. Ambiguous areas that could be reasonably classified as more than one class were left unlabeled and excluded from the analysis.

The classification maps would reflect any spatial correlations in the different terrain types. We quantified the degree of spatial correlation by calculating the fractional coverage of each class over the foreground of the transect images. Looking at just the foreground ensures that the areas analyzed do not overlap from one frame to the next. If the terrain types were not spatially correlated the resulting coverage fractions would be Independent and Identically Distributed (IID) for each class. We test this for each terrain type by treating the coverage fraction as a time series and applying a Durbin-Watson test [Durbin and Watson, 1971]. This is a test for no autocorrelation based on the residuals from a regression analysis. We favor the test for its simplicity, though it does not explicitly account for the constraint that the mixing ratios must sum to one. There exist Bayesian alternatives, such as hidden state time series representations, but inference in such models can be challenging so we will favor to the simpler Durbin Watson test here. Gravel and clast coverage has been found to be similar across larger spatially-extended units [Wells et al., 1985] so we expect these classes to fail the no autocorrelation test. The spatial distribution of iron oxide weathering is less clear in visual inspection, so we do not have any prior expectation for that result.
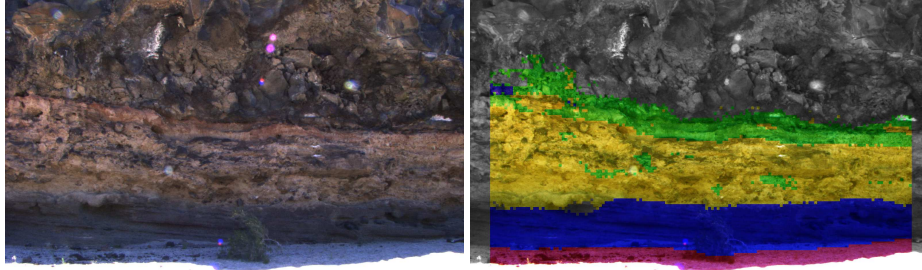
FIG. 10. Example of automatic image classification (Site A). Magenta: A-I Channel Bed. Blue: A-II Sorted volcanic gravel. Yellow: B-III Volcanic angular gravel and blocks in clay matrix. Green: A-IV: Oxidized clay. Class A-V (massive volcanic blocks) is excluded.

4 Field Experiment Results

We first consider results at Site A. Figure 10 shows a typical classification result assigning each surface class a different color. The system correctly interprets this scene as having contiguous horizontal layers. There is some slight mixing between iron oxide and clay classes, but the border of the sorted gravel layer is well localized. For clarity we have removed classification labels above the iron oxide layer and below the channel bed. The massive basalt class (A-V) was never classified accurately (for reasons that we will discuss below); after initial attempts to retrieve it we finally excluded it from our analysis.

Figure 11 quantifies performance at Site A using a Receiver Operating Characteristic (ROC) curve. The ROC curve treats each class as a one-vs-many classification problem, showing true positive and false positive rates that can be achieved for different probability thresholds. The operator can modify acceptance thresholds to move along this curve; for example, one can reduce the number of false positives by requiring classifications to be very certain. However, this will find fewer examples of each class. Conversely, lower probability thresholds result in more false positives. The ROC curve characterizes this tradeoff. Curves in the upper left of the plot suggest powerful classifiers with desirable performance profiles. These curves were produced by averaging the ROC performance from all images in the dataset.

The ROC curve for site A shows near-perfect performance on three of five classes; the system recovers 90% of classes A-I, A-III, and A-IV with negligible false positives. We are mainly interested in the leftmost portion of the ROC curve, since this is appropriate for a rover that collects just a few followup observations of high probability targets. The performance demonstrated at site A would allow a rover to perform representative sampling — a single followup measurement of each surface — with very high confidence. Class A-II is the most difficult, and the rate of false positives becomes non-negligible if one wishes to capture the entire surface, as the dark gravel area grades
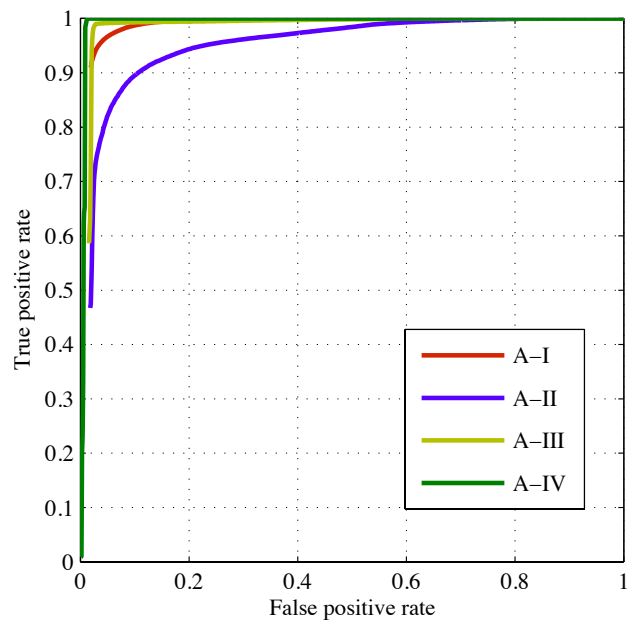
FIG. 11. ROC curve for Site A. The upper left corner represents ideal performance, detecting 100% of true positives with no false alarms.

	Predicted			
		A-I	A-II	A-III	A-IV
	A-I Sediment	**0.796**	0.204	0.000	0.000
Actual	A-II Gravel	0.031	**0.965**	0.004	0.000
	A-III Clay matrix	0.000	0.000	**0.982**	0.018
	A-IV Iron oxide	0.000	0.000	0.003	**0.997**

Table 1. Confusion matrix for classes at Site A. Bold entries show the precision of each automatic class label, e.g. the fraction that are correctly assigned.
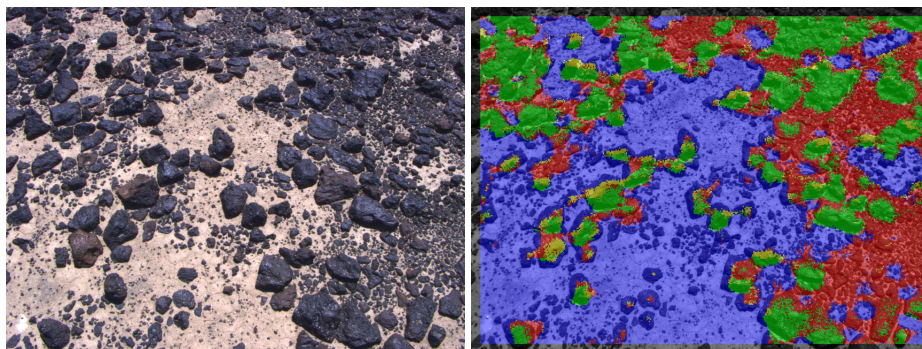

FIG. 12. Example of Cima image classification (Site B). Red: B-I gravel (< 64 mm). Blue: B-II sediment. Green: B-III basalt cobbles. Yellow: B-IV basalt cobbles with iron oxide coating.

into the shadowed sediment directly below it. Table 1 shows confusion matrices calculated over all pixels in the dataset. This also shows some shadow-induced confusion between sediment and gravel classes. We have normalized each entry, dividing by the number of true examples of that class. Thus the rows of the confusion matrices sum to one, but the columns in general do not.

Figure 12 provides a typical classification result for Site B, which captures both the size and distribution of rocks as well as the presence or absence of iron oxide coatings. Figure 13 shows the ROC curve calculated over all images at Site B. The system is highly effective at mapping the fundamental basalt/sediment distinction, and less accurate for the iron oxide and gravel classes. Again, this level of accuracy is sufficient for representative sampling because it is possible to select high-probability pixels that guarantee a negligible false positive rate. Table 2 shows a confusion matrix. Overall performance is good but there are some classes that are occasionally confused. For instance, the distinction between gravel and cobbles is less obvious in the upper portion of the images since the camera sees little of the sediment substrate at extreme aspect ratios.

However, most mislabelings represent the intrinsic subjectivity of the classes at Site B rather than difficulty in recognizing visual appearance. The distinction between sediment with small, sparsely scattered rocks and the gravel is somewhat blurred. Often
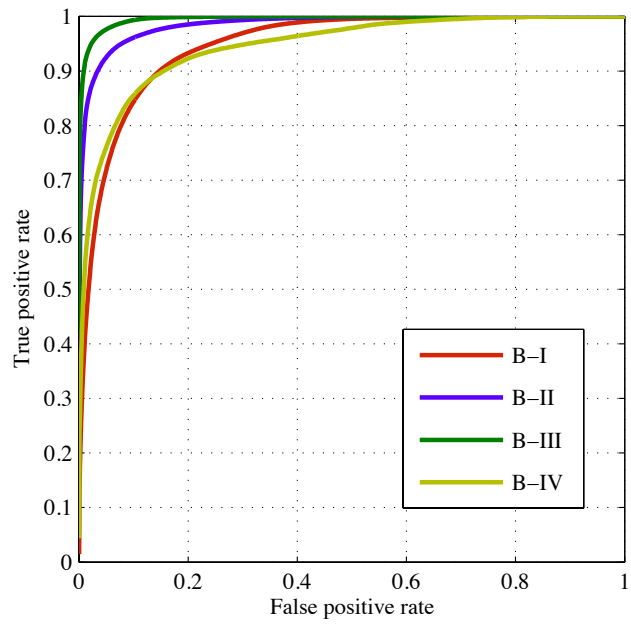
FIG. 13. ROC curve Site B. The best performance is in the upper left corner.

		Predicted			
		B-I	B-II	B-III	B-IV
	B-I Basalt gravel	**0.807**	0.182	0.010	0.001
Actual	B-II Sediment	0.005	**0.993**	0.001	0.001
	B-III Basalt cobbles	0.023	0.013	**0.961**	0.003
	B-IV Iron oxide	0.000	0.516	0.005	**0.479**

Table 2. Confusion matrix for classes at Site B. Bold entries show the precision of each automatic class label.

isolated gravel particles are justifiably counted as belonging to the surrounding sediment unit. The iron oxide class performs worst, with approximately half of the area labeled as basalt cobbles. This also reflects some inherent subjectivity since the two are simply endmembers of a distribution of coatings with different degrees of weathering and iron oxide coatings. The fact that approximately half of these pixels are mislabeled is compatible with the good precision evidenced in the ROC curve: there are many fewer iron oxide pixels than basalt, so one can ignore a large proportion of them without significantly increasing the overall false positive rate. That places some rare iron oxide within the more mundane background class, which is not a problem for rovers attempting to use this information for instrument targeting - operators can use a high probability threshold to ensure any iron oxide pixels they do select are likely to be real.

Table 3 shows the fractional coverage of each terrain type for the 16 images in the Cima sequence, and the p-value of the Durbin-Watson test for no autocorrelation of the time series. Low p values for the sediment, basalt, and gravel class suggest that these three units are not uncorrelated. This is appropriate because these three classes represent spatially extended phenomena on meter scales, while the iron oxide class represents an anomalous weathering condition that occurs intermittently on the surfaces of individual rocks. Figure 14 displays the same information as a bar chart. Figure 15 shows two extreme examples from the sequence: Image 2 (Left) which is estimated to have the smallest fraction of sediment, and Image 12 (Right) which has the largest. We conclude that spatial trends in the mapped sequence are consistent with the mesoscale spatial correlations expected for these terrain types.

5   Discussion

The experiments described above represent a challenging test of automated geologic image analysis. They evaluated the instrument in true field conditions with just a single training episode and no effort made to re-tune instrument parameters on the fly. These tests involved large contiguous mosaics of both vertical and horizontal surfaces, both with and without obvious stratigraphy. To our knowledge, it is the first time automated geologic image analysis of this kind has been demonstrated in real time instrument hardware. We will continue to refine the instrument, methods, and training sets, but this work provides an initial test of system performance and a bound on the errors that one might expect in an operational setting.
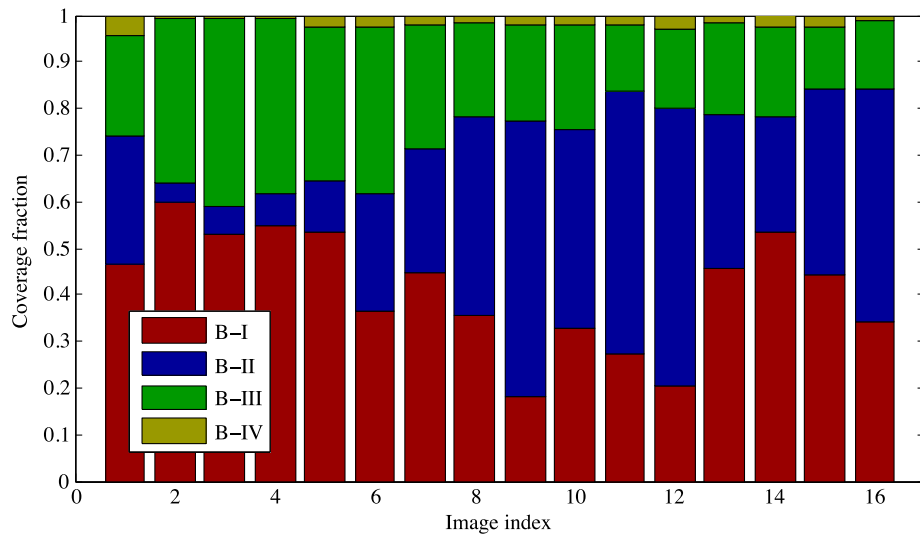
FIG. 14. Coverage fractions for each terrain type, for each image in the site B traverse sequence.
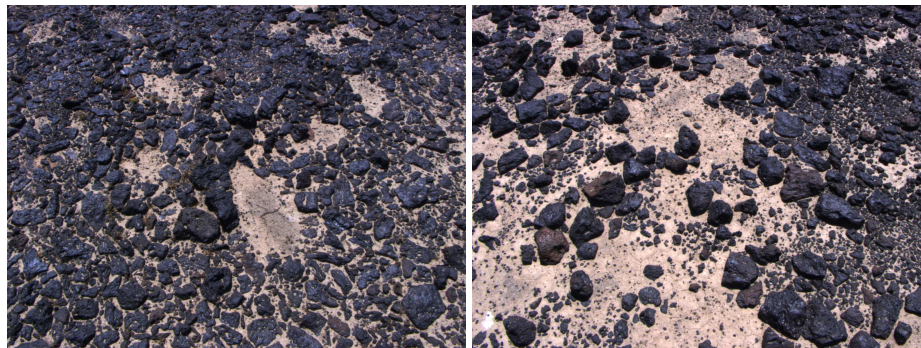

FIG. 15. Transect image 2, classified as having the smallest sediment fraction (Left) and transect image 12 having the largest fractional area of estimated sediment (Right)

19

Image	B-I	B-II	B-III	B-IV
1	0.47	0.27	0.21	0.05
2	0.60	0.04	0.35	0.01
3	0.53	0.06	0.41	0.01
4	0.55	0.07	0.38	0.01
5	0.54	0.11	0.33	0.02
6	0.37	0.25	0.36	0.02
7	0.45	0.26	0.27	0.02
8	0.35	0.43	0.20	0.01
9	0.18	0.59	0.21	0.02
10	0.33	0.43	0.22	0.02
11	0.27	0.56	0.14	0.02
12	0.20	0.60	0.17	0.03
13	0.46	0.33	0.20	0.02
14	0.53	0.25	0.19	0.03
15	0.44	0.40	0.13	0.03
16	0.34	0.50	0.15	0.01
p-value	0.011	0.002	0.008	0.240

Table 3. Fractional coverage of each surface type in each transect image, and p values of a Durbin-Watson test for autocorrelation in each class.

The overall result is encouraging, though we found a wide variation in mapping accuracy across different surface types. Certain distinctions were quite simple to teach the classifier, such as the difference between differently-colored volcanic strata (Site A) or the obvious visual distinction between sediment and rock (Site B). Accuracy rates for pure, unambiguous classes are typically above 95% representing a high likelihood of acquiring good targeted data regardless of the class abundance. This level of performance is already sufficient to improve rover exploration efficiency over status quo methods: rock detection is important for instrument targeting [Estlin et al., 2012] while layer detection is useful for a range of instrument placement and mapping operations [Gulick et al., 2001, Wagstaff et al., 2013]. In some cases, the automatic classification outperformed the initial human labeling, as in the iron oxide class at Site A. Here small image regions labeled as iron oxide led to the discovery of isolated patches of this class outside the main layer.

These tests also hint at more challenging classes, which may benefit from special pre-processing or post-processing. For example, the massive block class (A-V) is visually distinctive when looking at the entire rock face, but the instrument must make a classification decision based only on the local information near each pixel. Local surface textures provided insufficient information to reliably delineate the class. The iron oxide coatings on rocks at Site B are another challenging scenario. These targets were small and scattered, so that a labeling which did not correctly identify the perimeter of each iron oxide surface would incur a high penalty in terms of total pixel area. Even with this caveat, instrument targeting with these classifications would significantly outperform blind targeting when the features at a new locale are not known in advance.

But the performance variation by class means that operators seeking maximum reliability should define the desired targeting behavior in terms of simple and localized visual patterns.

The rare classes highlight the related and important issue of class imbalance: our baseline classifier does not typically take the relative abundance of each class into account, but the true probability may shift as the predominance of classes change. This is easy to remedy; in the case of class imbalance one could apply an additional informed prior $P(y)$ using Bayes' rule. This would amount to scaling each output probability by the relative abundance of each class. A similar scaling can capture different misclassification costs, leading to a Bayes' optimal classification with maximum total utility. By applying these simple postprocessing operations, a single generic random forest could be applied for many different scenarios.

6 Future Work

There are two principal avenues for future work: algorithm improvements and hardware integration. We plan to improve classification performance using more sophisticated pre-processing and a richer set of pixel attributes. Local descriptors [Dalal and Triggs, 2005] or oriented bar filters [Wagstaff et al., 2013] could improve performance, though their computational expense would probably demand specialized hardware support. We will also consider ways to address the challenge of spatial context, evidenced in class A-V where non-local information is required. Candidate solutions include input attributes that capture a larger spatial context, or class prior probabilities based on image location [Crandall et al., 2005].

We are also investigating further hardware upgrades in preparation for our next field expedition scheduled for summer 2014. We are currently adding the second Hitachi camera and porting the algorithm into the FPGA fabric. The addition of the second camera will enable stereo vision processing on the computer. The resulting range estimates will be used to select appropriate classifiers for each part of the scene. In porting the algorithm from the embedded PowerPC processor inside the FPGA to the surrounding fabric, we hope to achieve a 1 Hz processing rate by performing multiple per-pixel decision tree evaluations in parallel. We will also make the instrument more convenient to use by extending the WiFi range and reducing Ethernet data drops.

7 Conclusions

This work provides several unique new contributions to the field. First, we have taken random forest algorithms recently demonstrated for Mars surface geology [Wagstaff et al., 2013] and applied them to real-time geologic surface analysis in a field environment. This demonstrates that the same basic algorithm applies to both scenarios and suggests a potential for a wide range of different rover science autonomy tasks. Unlike previous tests, which have generally relied on traditional computing platforms, we have implemented our classifier system directly in embedded instrument hardware. The resulting system was evaluated at two distinct regions in the Cima Volcanic fields, where

it successfully classified visually-distinctive surface patterns using color and texture cues. Further work will continue to refine pre-processing, the algorithm itself, and the instrument platform.

Acknowledgements

References

[Bakambu et al., 2012] Bakambu, J. N., Langley, C., Pushpanathan, G., MacLean, W. J., Mukherji, R., and Dupuis, E. (2012). Field trial results of planetary rover visual motion estimation in Mars analogue terrain. *Journal of Field Robotics*, 29(3):413–425.

[Bauer et al., 2007] Bauer, J., Sunderhauf, N., and Protzel, P. (2007). Comparing several implementations of two recently published feature detectors. In *Proc. of the International Conference on Intelligent and Autonomous Systems*, volume 6.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32. 10.1023/A:1010933404324.

[Crandall et al., 2005] Crandall, D., Felzenszwalb, P., and Huttenlocher, D. (2005). Spatial priors for part-based recognition using statistical models. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005.*, 1:10–17.

[Dalal and Triggs, 2005] Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005.*, 1:886–893.

[Deans et al., 2012] Deans, M., Lees, D., Smith, T., Cohen, T., Morse, T., and Fong, T. (2012). Field testing next-generation ground data systems for future missions. In *Lunar and Planetary Institute Science Conference Abstracts*, volume 43, page 2518.

[Dohrenwend et al., 1984] Dohrenwend, J. C., McFadden, L. D., Turrin, B. D., and Wells, S. G. (1984). K-Ar dating of the Cima volcanic field, eastern Mojave Desert, California: Late Cenozoic volcanic history and landscape evolution. *Geology*, 12(3):163–167.

[Durbin and Watson, 1971] Durbin, J. and Watson, G. S. (1971). Testing for serial correlation in least squares regression. *Biometrika*, 58 Number 1:1–19.

[Estlin et al., 2012] Estlin, T. A., Bornstein, B. J., Gaines, D. M., Anderson, R. C., Thompson, D. R., Burl, M., Castaño, R., and Judd, M. (2012). AEGIS automated science targeting for the MER Opportunity rover. *ACM Trans. Intell. Syst. Technol.*, 3(3):50:1–50:19.

[Farr, 2004] Farr, T. G. (2004). Terrestrial analogs to mars: The NRC community decadal report. *Planetary and Space Science*, 52(1–3):3–10.

[Foil et al., 2013] Foil, G., Thompson, D. R., Abbey, W., and Wettergreen, D. (2013). Probabilistic surface classification for rover instrument targeting. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE.

[Francis et al., 2014] Francis, R., McIsaac, K., Thompson, D. R., and Osinski, G. R. (2014). Multiclass linear discriminant analysis as a tool for image segmentation in geology. *IEEE International Conference on Robotics and Automation (submitted)*.

[Furgale and Barfoot, 2010] Furgale, P. and Barfoot, T. D. (2010). Visual teach and repeat for long-range rover autonomy. *Journal of Field Robotics*, 27(5):534–560.

[Grotzinger et al., 2012] Grotzinger, J., Crisp, J., Vasavada, A., Anderson, R., Baker, C., Barry, R., Blake, D., Conrad, P., Edgett, K., Ferdowski, B., Gellert, R., Gilbert, J., Golombek, M., Gmez-Elvira, J., Hassler, D., Jandura, L., Litvak, M., Mahaffy, P., Maki, J., Meyer, M., Malin, M., Mitrofanov, I., Simmonds, J., Vaniman, D., Welch, R., and Wiens, R. (2012). Mars Science Laboratory mission and science investigation. *Space Science Reviews*, 170(1-4):5–56.

[Gulick et al., 2001] Gulick, V. C., Morris, R. L., Ruzon, M. A., and Roush, T. L. (2001). Autonomous image analyses during the 1999 Marsokhod rover field test. *Journal of Geophysical Research*, 106(E4):77457763.

[Hitachi, 2013] Hitachi (2013). HV-F202SCL/GV, KP-FD202PCL/SCL Specifications. `http://www.hitachikokusai.com`.

[McFadden et al., 1987] McFadden, L., Wells, S., and Jercinovich, M. (1987). Influences of eolian and pedogenic processes on the origin and evolution of dsert pavements. *Geology*, 15:504-508.

[McGuire et al., 2010] McGuire, P. C., Gross, C., Wendt, L., Bonnici, A., Souza-Egipsy, V., Ormoe, J., Diaz-Martinez, E., Foing, B., Bose, R., Walter, S., et al. (2010). The cyborg astrobiologist: testing a novelty detection algorithm on two mobile exploration systems at rivas vaciamadrid in spain and at the mars desert research station in utah. *International Journal of Astrobiology*, 9(1):11.

[Mlynarczuk et al., 2013] Mlynarczuk, M., Górszczyk, A., and Slipek, B. (2013). The application of pattern recognition in the automatic classification of microscopic rock images. *Computers & Geosciences*, 60:126–133.

[Mustard et al., 2013] Mustard, J. F., Adler, M., Allwood, A., Bass, D. S., Beaty, D. W., Bell III, J. F., Brinckerhoff, W. B., Carr, M., Des Marais, D. J., Drake, B., Edgett, K. S., Eigenbrode, J., Elkins-Tanton, L. T., Grant, J. A., Milkovich, S. M., Ming, D., Moore, C., Murchie, S., Onstott, T. C., Ruff, S. W., Sephton, M. A., Steele, A., and Treiman, A. (2013). Report of the Mars 2020 Science Definition Team, 154 pp., posted july, 2013, by the mars exploration program analysis group (mepag). http://mepag.jpl.nasa.gov/reports/MEP/Mars_2020_SDT_Report_Final.pdf.

[Peralta et al., 2008] Peralta, F. C., Thompson, D. R., and Wettergreen, D. (2008). Autonomous rover reflectance spectroscopy with dozens of targets. *International Symposium on Artificial Intelligence, Robotics and Automation in Space (ISAIRAS)*.

[Shotton et al., 2008] Shotton, J., Johnson, M., and Cipolla, R. (2008). Semantic texton forests for image categorization and segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (2008)*, 259(4 Pt 2):1–8.

[Squyres et al., 2003] Squyres, S. W., Arvidson, R. E., Baumgartner, E. T., Bell, J. F., Christensen, P. R., Gorevan, S., Herkenhoff, K. E., Klingelhfer, G., Madsen, M. B., Morris, R. V., Rieder, R., and Romero, R. A. (2003). Athena Mars rover science investigation. *Journal of Geophysical Research: Planets*, 108(E12).

[Thompson et al., 2008] Thompson, D. R., Smith, T., and Wettergreen, D. (2008). Information-optimal selective data return for autonomous rover traverse science and survey. In *IEEE Conference on Robotics and Automation*.

[Thompson et al., 2011] Thompson, D. R., Wettergreen, D., and P., F. J. C. (2011). Autonomous science for large-scale robotic survey. 28(4).

[Wagstaff et al., 2004] Wagstaff, K., Castano, R., Dolinar, S., Klimesh, M., and Mukai, R. (2004). Science-based region-of-interest image compression. *Lunar and Planetary Institute Science Conference Abstracts*, 35:1934.

[Wagstaff et al., 2013] Wagstaff, K. L., Thompson, D. R., Abbey, W., Allwood, A., Bekker, D. L., Cabrol, N. A., Fuchs, T., and Ortega, K. (2013). Smart, texture-sensitive instrument classification for in situ rock and layer analysis. *Geophysical Research Letters*, 40(16):4188–4193.

[Wells et al., 1985] Wells, S., Dohrenwend, J., McFadden, L., Turrin, B., and Mahrer, K. (1985). Late cenozoic landscape evolution on lava flow surfaces of the cima volcanic field, mojave desert, california. *GSA Bulletin*, 96: 1518-1529.

[Wells et al., 1995] Wells, S. G., McFadden, L. D., Poths, J., and Olinger, C. T. (1995). Cosmogenic 3He surface-exposure dating of stone pavements: Implications for landscape evolution in deserts. *Geology*, 23(7):613–616.

[Wilshire, 1992] Wilshire, H. G. (1992). Geologic map of the Indian Spring Quadrangle, San Bernardino County, California. In *Open-File Report*, volume 92-181.

[Woods et al., 2009] Woods, M., Shaw, A., Barnes, D., Price, D., Long, D., and Pullan, D. (2009). Autonomous science for an exomars roverlike mission. *Journal of Field Robotics*, 26(4):358–390.

[Woods et al., 2011] Woods, M., Shaw, A., and Rendell, P. (2011). High-level autonomy and image prioritisation for long distance Mars rovers. In *Proc. 11th Symposium on Advanced Space Technologies in Robotics and Automation, ESA/ESTEC, The Netherlands*, pages 12–14.