2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

Probabilistic Surface Classification for Rover Instrument Targeting

Greydon Foil[1], David R. Thompson[2], William Abbey[2], and David S. Wettergreen[1]

Abstract— **Communication blackouts and latency are significant bottlenecks for planetary surface exploration; rovers cannot typically communicate during long traverses, so human operators cannot respond to unanticipated science targets discovered along the route. Targeted data collection by point spectrometers or high-resolution imagery requires precise aim, so it typically happens under human supervision during the start of each command cycle, directed at known targets in the local field of view. Spacecraft can overcome this limitation using onboard science data analysis to perform autonomous instrument targeting. Two critical target selection capabilities are the ability to target priority features of a known geologic class, and the ability to target anomalous surfaces that are unlike anything seen before.**

This work addresses both challenges using probabilistic surface classification in traverse images. We first describe a method for targeting known classes in the presence of high measurement cost that is typical for power- and time-constrained rover operations. We demonstrate a Bayesian approach that abstains from uncertain classifications to significantly improve the precision of geologic surface classifications. Our results show a significant increase in classification performance, including a seven-fold decrease in misclassification rate for our random forest classifier. We then take advantage of these classifications and learned scene context in order to train a semi-supervised novelty detector. Operators can train the novelty detection to ignore known content from previous scenes, a critical requirement for multi-day rover operations. By making use of prior scene knowledge we find nearly double the number of abnormal features detected over comparable algorithms. We evaluate both of these techniques on a set of images acquired during field expeditions in the Mojave Desert.

I. Introduction

Exploration spacecraft are increasing in both mobility and their capacity to collect large volumes of science data, making communications latency and bandwidth a critical bottleneck for mission science return [1], [2]. Recent tests have demonstrated the ability of rovers to traverse multiple kilometers a day [3], and the recently-launched Mars Science Laboratory rover carries an order of magnitude more instrument mass than any previous rover mission [4]. Missions are exploiting these payloads to actively and passively analyze habitats, detect biosignatures, and characterize chemical abundances, yet spacecraft mobility and potential to collect scientific data greatly outpace their communications ability [2]. Limited communication windows with Earth, communications blackouts, and low-bandwidth transfer methods hinder the amount of information returned and the rate of discovery of spacecraft [1], [2], greatly reducing the lifetime scientific return of a mission and necessitating the development of reliable onboard analysis methods.

This is particularly important for the next generation of astrobiology-inspired rover missions. Even ignoring communication restraints, the search for peleohabitat indicators is particularly challenging because any evidence is likely to be sparse, isolated, and difficult to detect from a distance [5]. Onboard data analysis can play an important role to maximize scientific return and reduce the number of missed observation opportunities. In recent years a number of algorithms have been developed that aid in scene understanding, automated targeting, and data summary. Advances include rock detection and classification of rock characteristics [2], [6], and detection and tracking of dust devils [7]. The AEGIS system [8], currently operating on the Mars Exploration Rovers, autonomously discovers scientifically interesting features and targets them for followup observations by high-resolution imagery on the same command cycle.

Onboard data analysis can be particularly transformative for missions to the farthest and harshest regions of the Solar System. Hostile environments such as the intense heat on the surface of Venus or the radiation of Europa mean that landers have an extremely limited amount of time to target instruments, take readings, and transmit this data to Earth. When the communications delay between the probe and Earth is nearly as long or longer than the expected lifetime of the spacecraft, onboard data analysis can be used to prioritize key measurements, detect and target anomalous regions, and return low bandwidth maps and compressed representations of scenes.

Instrument targeting typically involves finding specific features of scientific interest like rock outcrop, layered strata, or specific geologic facies [9], and then taking aimed measurements with instruments like high-resolution cameras or point spectrometers. Finding these target surfaces is tantamount to a traditional classification problem. However, planetary science has special requirements that differ from those of generic classification tasks. First, automatic instrument targeting requires excellent classifier precision. Time and power resources for followup data collection are highly constrained so the system must be very confident in the classification before diverting from its plan for opportunistic data collection. It is likely that an explorer robot will visit environments viewed in a variety of lighting or environmental conditions. It will be critical to for the analysis system to be able to adjust to these conditions, even if that means forgoing classification of portions of the scene. Second, it is also important that the

[1] G. Foil and D. S. Wettergreen are with The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213 gfoil@cs.cmu.edu, dsw@ri.cmu.edu

[2] D. R. Thompson and William Abbey are with the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA David.R.Thompson@jpl.nasa.gov, William.J.Abbey@jpl.nasa.gov

978-1-4673-6357-0/13/$31.00 ©2013 European Union

system find anomalous surfaces that appear different from anything the mission has seen previously. Such anomalies are critically valuable for exploration.

This work demonstrates a probabilistic surface classification approach for target selection during rover traverse. We first introduce a general algorithm for increasing the precision of surface classification. The proposed system prioritizes low-risk classifications, removing the most ambiguous pixels from the final output. We describe the cost of abstention as a fixed error, making it straightforward to compute Bayes-optimal decisions about which pixels to classify. This exploits the fact that full scene classification is not needed for most instrument targeting applications. By only classifying regions of high confidence, a rover may operate in a large variety of environments or new imaging conditions over long periods while maintaining a high level of targeting precision. The cost of abstention is directly related to classifier confidence and misclassification costs, and can easily be adjusted for varying applications.

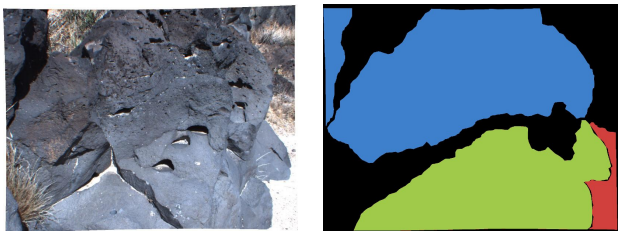
Building on these results, we use these classifications to direct novelty detection in a semi-supervised manner. We use the classification results from a new image to seed a distance-based anomaly detector with known background regions. The system can thus be trained to ignore specific features or surfaces that are known to be non-anomalies. The novelty score not only provides an in-image ranking of each window's novelty, but also provides a standardized metric with which we can compare windows across an entire traverse, allowing the rover to make more intelligent decisions about instrument targeting or optimize the return of the most novel regions when a data uplink becomes available.

We evaluate our algorithms on a random forest classifier trained on geologic textures. The classifier identifies patterns in image channels, such as an object's color, range from the camera, vertical height, or state of illumination, in order to provide scene analysis and autonomous instrument targeting [10]. These patterns are used to ascribe classification labels to physical surfaces in the spacecraft environment. Classified images can be used to guide instrument placement to features that exemplify local terrain, as well as score dissimilar regions for further analysis.

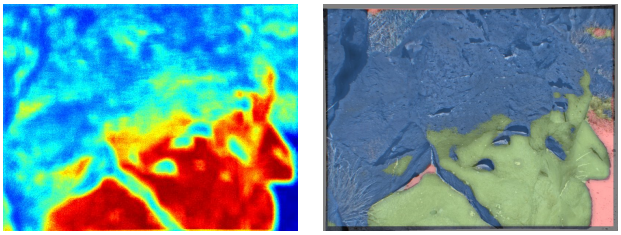
Section II discusses related work, the classification algorithm, and our use of these classifications for instrument targeting. Section III outlines our experimental dataset as well as our training and testing procedures. Section IV reports experimental results.

II. Approach

The initial scene analysis classifies each pixel of an image according to the physical texture. We favor a random forest classifier [11], an ensemble of decision tree classifiers T, each trained on a random subset of the training data. Trees are made up of nodes n and leaves $L = (l_1, ...l_m)$, where each leaf stores a learned class distribution $P(c|n)$ [Here and elsewhere P designates probability, while pixels are designated p_i], or simply the probability of being a member



(a) Example test image (b) Ground truth comparison



(c) Smooth basalt class probability (d) Overlayed classification

Fig. 1: Sample classification process. We learn texture from images such as (a) using ground truth shown in (b). The blue label indicates vesicular basalt, green indicates smooth basalt, and red indicates sand. (c) shows a heatmap of the smooth basalt class, warmer colors indicating higher confidence in that classification. (d) is the final output of the classifier overlayed on the original image, showing excellent agreement with the geologist classification.

of class c given the path of nodes traversed before reaching the given leaf.

Classifications are made by starting at the root node of each tree and branching left or right at each node according to the response of a node-specific operation on the vector of inputs. The class of the input vector is then assigned as the Maximum A Posteriori (MAP) class after averaging across all trees:

$$P(c|L) = \frac{1}{T}\sum_{t=1}^{T} P(c|l_t) \quad (1)$$

A. Random Forest Generation

To generate a random forest, we train each tree on a small random subset $I' \subseteq I$ of available training images I. We select training points from I' using a random sampling method. At each node n of the tree, we split the data at that node, I_n, recursively into left and right subsets, I_l and I_r, respectively, according to a threshold τ of a split function f on the feature vector $\mathbf{v}$:

$$I_l = i \in I_n | f(\mathbf{v}_i) < \tau \quad (2)$$

$$I_r = I_n - I_l \quad (3)$$

At each node, we generate a number of candidate functions for f. Our algorithm uses five main binary comparisons: the value at pixel p_i, or the difference, sum, absolute difference, or ratio between two pixel values p_i and p_j within a window. Pixel values can be from any color or pre-processing channel, and the channels of p_i and p_j do not have to be the same.

Candidates for f are a random combination of these binary comparisons and relative pixel locations for p_i and p_j.

We select the candidate function and threshold that maximizes the expected information gain over the classes in I_n:

$$E[\Delta H] \quad = \quad -\frac{|I_l|}{|I_n|}H(I_l) - \frac{|I_r|}{|I_n|}H(I_r) \qquad (4)$$

where $H(I)$ is the Shannon entropy of the classes in I. Once the best operation from the random set has been chosen, we split the training points and create two leaf nodes. We then compute $P(c|n)$ for each class using its population of training data. This splitting continues recursively until a maximum number of splitting operations has occurred. We disallow splits that would create a child population with fewer than 32 pixels.

We select thresholds with an exhaustive search. It should also be noted that the training data may be biased towards more frequently occurring classes. To account for this, we normalize the leaf distributions by weighting each sample by the inverse of its class frequency.

We use a random forest classifier inspired by Shotton et al. [12]. The classifier used here resembles their work with two main differences. First, the Shotton et al. work uses splitting functions based on single pixel values and fast comparisons between them: pixel differences, absolute differences, and sums. We augment these features with pixel ratios, increasing performance slightly for our dataset. Second, their work includes a number of post-processing steps based on *bag of semantic textons* to improve overall scene consistency. We forgo these steps, and instead only perform a single pixel-wise classification.

B. Classification with Abstention

Instrument targeting requires very precise classification since time and power are constrained and any opportunistic data collection displaces other science activities. However, there is typically a surplus of target surface so that precision only matters where the rover chooses to collect data, i.e. in small regions of an image. Consequently we can improve task performance by abstaining from uncertain classifications. In our scenario, the random forest classifier provides an average estimate of posterior class probabilities. We propose a Bayesian approach that uses these classification confidences to choose when to abstain. Our cost function gives appropriate penalties to both misclassifications and abstentions.

During a single pixel's classification, each tree in the forest provides a posterior probability $P(C = c|x)$ of a pixel of class C being classified as class c given inputs x. The agent takes an action a, either ascribing a specific class or abstaining. The agent then incurs a cost based on the true class c and the action taken. Correct classifications are considered to have no cost, while misclassifications have a per-class cost of β_c and abstentions have a cost α ($0 \leq \alpha \leq \min_c \beta_c$). In an instrument targeting scenario, the flexibility of per-class misclassification costs allows us to weight the risk of misclassification against the cost of sampling. Common classes may have a high misclassification cost to reduce mistargeting risk, while infrequent classes may warrant a lower one.

Given n classes and the pixel's features, the expected loss of choosing an action $a \in \{abstain,\ classify\ as\ c\}$ is:

$$E[L|x,a] \quad = \quad \frac{1}{n}\sum_c L(c,a)P(C=c|x) \qquad (5)$$

$$\text{where } \mathrm{L(c,a)} \quad = \quad \begin{cases} 0 \ \text{ if } \ \mathrm{a = c} \\ \beta_c \ \text{ if } \ \mathrm{a} \neq \{\mathrm{c, abstain}\} \\ \alpha \ \text{ if } \ \mathrm{a = abstain} \end{cases} \qquad (6)$$

It is optimal to choose the action that is associated with the smallest expected loss. Extrapolating to m trees $(t_1, ..., t_m)$, the optimal action is found by averaging the expected loss across all trees:

$$argmin_a\ E[L|x,a] = \frac{1}{n}\sum_c L(c,a)\frac{1}{m}\sum_t p_t(C=c|x) \qquad (7)$$

By varying α, we increase or decrease the number of unclassified pixels. A high value for α will lead to fewer abstentions, while a lower value will lead to many.

There is considerable prior research in abstaining and reliable classifiers, especially in medical disciplines demanding highly accurate diagnoses. Elazmeh et al. [13] propose using Tango's test to find regions of receiver operating characteristic (ROC) curves that contain reliable classifications. Vanderlooy et al. [14] propose ROC isometrics, or methods with which to find unreliable regions of the ROC graph. Eliminating these regions increases the overall precision of the classifier. Unfortunately, many of these works deal with binary classification problems, so they do not translate as well to the multiclass scenario proposed here.

The abstention method that most closely resembles this work is that of Chow [15]. Chow proposes a Bayes-optimal method for selecting a 'rejection threshold' (a confidence score below which a classifier should abstain). Chow shows that when the loss is defined by rejection and misclassification rates, the rejection threshold is sufficient to exactly specify a particular classifier in the Pareto-optimal set. Here we expand this formulation to an arbitrary multiclass loss function, in which both the threshold and cost for incorrect classifications can be altered to best suit the task at hand. This retains the theoretical elegance of the Chow approach, while providing the additional flexibility of asymmetric misclassification loss to distinguish high-value targets.

C. Novelty Detection

When scenes are well-represented by training data, classifiers can provide a complete interpretation. However, when anomalous scene characteristics are encountered, novelty detection can leverage a probabilistic classification to mark regions of image for further analysis, instrument targeting, or prioritized uplink. Taken across an entire traverse, novelty detection can provide summary statistics that alert scientists to unusual events. Such a system must also be trainable so that it is not fooled by local anomalies that are pervasive in historical data.

We propose a semi-supervised system that utilizes the results of the abstaining classifier as a background model leading to a per-pixel novelty detection score. We begin by classifying the image and build a probability density model using high-confidence classifications.

The feature space consists of the HSV color values and m Gabor wavelets of various scales and orientations to provide local texture statistics. Thus each pixel p_i in the scene becomes a vector of length $m+3$:

$$p_i = (x_h,\ x_s,\ x_v,\ x_{g1},\ ...,\ x_{gm})^T \tag{8}$$

Using the classification labels as pixel priors, the mean μ_c and covariance S are calculated over the pixels of each class c:

$$\mu_c = (\mu_{h,c},\ \mu_{s,c},\ \mu_{v,c},\ \mu_{g1,c},\ ...,\ \mu_{gm,c})^T \tag{9}$$

Using these, a per-pixel distance to each class is then calculated using the Mahalanobis distance metric:

$$D_c(p_i) \quad = \quad \sqrt{(p_i - \mu_c)^T S^{-1} (p_i - \mu_c)} \tag{10}$$

The novelty score for a given pixel $S(p_i)$ is thus the minimum Mahalanobis distance to any class:

$$S(p_i) \quad = \quad \min_c D_c(p_i) \tag{11}$$

As the novelty scores for nearby pixels can have a high variance, we average the pixel-level scores within a window w of a fixed size N:

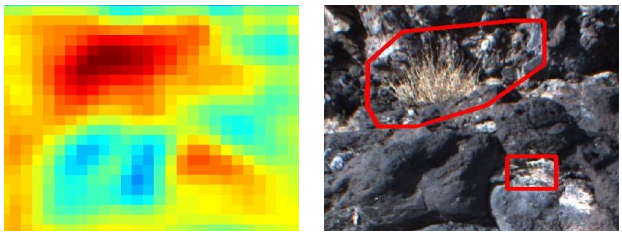
$$S_w = \frac{1}{N} \sum_{i=1}^{N} S(p_i) \tag{12}$$

Window novelty scores can then be compared directly to determine the regions within the scene of the highest novelty. See Figure 2 for example scores and resulting novel regions.

It is worth noting that a Mahalanobis distance over $m+3$ image channels is directly related to a p-value from a χ^2 distribution with $m+3$ degrees of freedom. Thus, using Mahalanobis as a distance metric provides us with a normalized score that is consistent across an entire traverse, allowing our system to also rank novel regions between multiple images.

There has been much prior work on calculating image novelty, as novel regions in an image often contain features of high information content. Multiple approaches have used saliency as a method for automated video indexing [16], [17], greatly reducing the required amount of manual review. Other works use machine learning techniques to find abnormalities in an image by training against sample data, making the assumption that such training sets exist and have consistent features [18], [19].

Much prior work presents anomaly detection under the more general theme of image saliency. Given a sample image or series of images, their detectors find a ranked list of salient regions. Itti et al. propose a method that simulates the visual search process of humans [20]. They calculate responses to center-surround operations, color contrast, and contrast between local orientation responses, then linearly combine them to calculate saliency scores. Hou et al. consider the same scenario, yet approach it using a spectral residual model [21]. The spectral residual approach shows improvement over Itti's method, and we compare against it in the evaluations below. For a more comprehensive survey on saliency techniques, we refer the reader to [22].



(a) Novelty heatmap (b) Original image with novel regions marked

Fig. 2: Example novelty detection. We use classifications of the true color image shown in (b) as class priors. Using the average minimum Mahalanobis distance over image windows we calculate the heatmap shown in (a). Windows are ranked with a novelty score and are shown overlayed in (b).

In an exploration context, finding novel regions or images has many benefits. Primarily, novelty algorithms are able to address data management issues, such as automating data analysis of a traverse or improving the relevancy of information returned through limited data uplinks. Johnson-Robertson et al. use an entropy-based saliency method to analyze large amounts of underwater data, both to find salient images and salient pixel regions, greatly reducing the amount of manual examination required [23]. Wagstaff et al. do not explicitly calculate novel pixels, yet propose a method in which regions of images are selectively compressed if they have low information content [24]. Thompson et al. use a semi-supervised eigenbasis approach to novelty detection in radio astronomy time series data to distinguish abnormal radio signals from background noise [25].

III. Experimental Details

This section evaluates image interpretation for instrument targeting in a challenging field environment: the Cima volcanic fields in the Mojave National Preserve.

A. Dataset

The dataset used in this work is a series of images of vesicular and smooth basalt formations in the Cima volcanic field in the Mojave Desert. These fields are a part of the Mojave Cima Volcanic Range, which includes 40 volcanic cinder cones and associated basaltic lava flows. The flows formed when the cinder cones erupted as relatively benign liquid fountains. Gases escaping from the cooling lava created bubbles or "vesicles" directly visible in the rock surface. The number and size of vesicles indicates distinctive processes and compositions associated with different parts of the flow. We identified a portion of the flow containing three main classes: smooth basalt, vesicular basalt, and sand. While sand is typically easily identifiable, smooth and vesicular

basalt formations are often of similar color and illumination, mainly differing in texture. Furthermore, without contextual cues it is often unclear where one class ends and another begins, making classification difficult.

We simulated multiple rover imaging sequences, approaching the rock face from a distance and collecting images at 0.5m intervals. This is similar to a navigation sequence collected by a rover approaching a sampling target during single-command instrument placement. We acquired images from a color stereo rig with a 12cm baseline, providing color as well as accurate range information. The midday sun provided bright overhead illumination and clear distinctive cast shadows.

We create two versions of our ground truth images. One version contains labels for smooth and vesicular basalt, as well as sand. These are used as labels in the classification training and evaluation processes, and were labeled by a geologist. Vegetation and ambiguous regions in which a trained geologist cannot identify a class are left unlabeled, excluded from both the training and test process. Evaluation of the abstaining classifier is performed using this version.

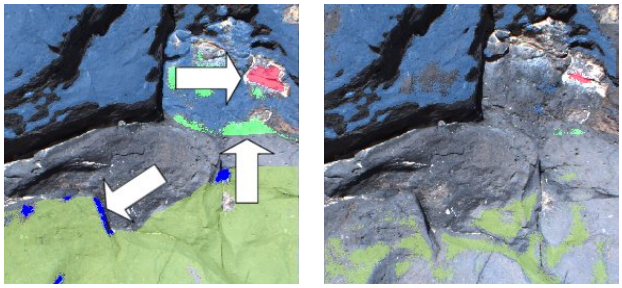
The second version of ground truth images contains the same labels as the first, but also includes the labeling of vegetation and interesting rock features. Interesting rock features in this context are discolorations in a rock face, rocks of unusual color or texture, or veins within a rock face. All test images contain between five and ten of these features. Evaluation of all novelty detection methods are performed using this set.

B. Classifier Training and Test Procedure

We evaluate performance on 23 images from our dataset. We convert all images to the HSV color space which we have found gives a slight increase in classification performance over an RGB representation. We train our classifier using 100 trees, 50,000 sample points per tree, 64 expansions within each tree, and a window size of 41 pixels in which the algorithm searches for pixel comparisons that maximize information gain. We find our test results are insensitive to the precise training parameters used.

We evaluate performance using leave-one-out cross validation, comparing posterior classification probabilities against ground truth. We ignore classification of abstained pixels and assume a constant misclassification cost β_c for all classes.

We experimented with using stereo range values as an additional input channel as in [26]. For this dataset, incorporating the range data as a feature decreased generalization performance and provided no other noticeable benefits. However, we did find an increase in classification performance when multiple classifiers were used, each trained on a subset of range values. We partition the training data into groups of pixels having similar distances from the camera, creating three subsets: a close-range subset from $0-2$m, a mid-range subset from $2-6$m, and a and long-range subset beyond 6m. These distinctions are significant for rover operations; the first interval is roughly the workspace for arm-mounted contact sensors, the second interval may have valid remote sensing targets, and the long-range interval would probably require further driving to investigate. Range data from a new scene determines the appropriate classifier for each new pixel, and pixels without range data are ignored in both training and classification.



(a) $\alpha = 0.5$ (b) $\alpha = 0.3$

Fig. 3: Decreasing the cost of abstention causes the classifier to become more conservative, leaving a larger fraction of pixels unlabeled. Brighter colors are misclassifications of the correspondingly-colored class and are indicated in (a) by arrows.

We note that classification is greatly improved when performed on well-lit regions. In this case we smooth the image with a boxcar filter and then apply an intensity threshold to find lit and shadowed surfaces, operating under the assumption that the image has first been contrast-normalized in prior processing or through automatic exposure and gain settings on the physical camera.

C. Novelty Test Procedure

We evaluate novelty performance using the same set of 23 images. We perform classification with abstention as described above using a conservative abstaining threshold of $\alpha = 0.4$, chosen because it provides a good tradeoff between the number of pixels classified and classifier performance, as shown in Figure 4. The classified pixels are then used to seed a background model in the density estimation. As outlined above, we combine HSV color channels with ten Gabor wavelets of differing orientation and scale, then calculate the mean of each image channel and covariance of each class and use these to calculate the minimum Mahalanobis distance from each unclassified pixel to the class means.

We calculate the average novelty score for image regions using windows of size 60 by 60 pixels, evaluated every 20 pixels. The highest-scoring windows are then combined together until there are five distinct regions, and the convex hull of these regions are checked for vegetation or rock feature classes. A limit of five regions is chosen because, on average, scenes in the dataset are labeled with approximately eight novel regions, but sometimes as few as five.

We compare our method to three others: Hou's spectral residual method, a Gaussian mixture model (GMM), and a similar semi-supervised approach which treats image channels as independent.

The GMM approach initializes three Gaussian clusters using K-means clustering, then performs Expectation Maxi-

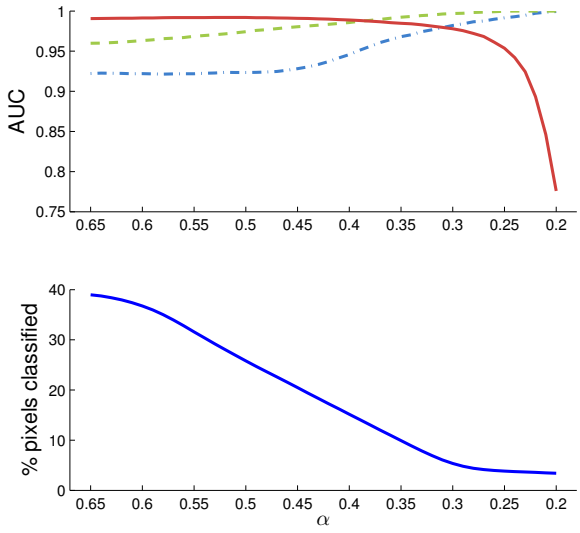


Fig. 4: Top: AUC performance as the confidence threshold α changes. The green line represents smooth basalt, blue represents vesicular basalt, and red represents sand. An AUC of 1 indicates perfect classification. Bottom: Percent of pixels classified as α changes.

mization (EM) to shift the distributions to better fit the data. If EM fails due to numeric instability or other causes, the test uses the initial K-means solution. The means and covariances of each class are calculated from the final mixture model and the Mahalanobis distance is used to score novelty in a similar manner to our proposed approach. Note that this method simulates an unsupervised version of our approach in which class labels are not known but are instead estimated.

The alternate semi-supervised method treats each channel of the image as an independent Gaussian. The abstaining classifications are used to calculate channel means and variances for each class, and pixels are given the novelty score of the minimum number of standard deviations from class means across all channels.

For all test methods we designate a novel region as correctly labeled if that region contains at least 40% vegetation or interesting rock feature classes. 40% is chosen here due to the size of the vegetation and rock feature labels as compared to the region size, as the convex hull of regions in all methods often includes an area larger than the ground truth label.

IV. Results

A. Abstention Results

We first investigate improving precision by abstaining from classification. We evaluate this method on a classifier trained using range segmentation on illuminated surfaces, the highest performing strategy. We vary α, the cost of abstention, increasing or decreasing the number of pixels within the scene that are classified. At lower values of α the classifier is very strict, ascribing only its most confident classifications, while higher values of α lead to a higher number of classified

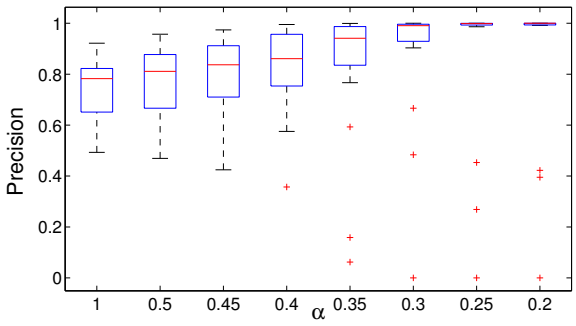


Fig. 5: Percent of pixels correctly classified across the entire dataset while varying α.

pixels. Since the system incurs a misclassification cost of one, it will ascribe a class whenever its posterior probability exceeds $1 - \alpha$. There are three possible classes, so reducing α may not always reduce the number of classified pixels.

We evaluate the abstaining classifier by observing the performance of each class as α is changed. We use the Area Under the Curve, or AUC, metric, in which ROC curves with a higher area are said to have better performance. An AUC value of 1 indicates perfect performance. Figure 4 shows that as α decreases, the AUC increases for both the smooth and vesicular basalt classes. The sand class has high initial performance, yet tapers off as α decreases, possibly indicating mislabeled ground truth regions or confident misclassifications of high saturation regions. However, for the main basalt classification task, the AUC does increase, suggesting the posterior probabilities are meaningful estimates of classification confidence. As shown in the lower part of Figure 4, there is a trade off between classification performance and the amount of scene classified, but even a moderate abstention cost improves performance, and can be further adjusted to meet mission requirements.

Figures 3 and 6 show visually the change in classified regions as α is decreased. As expected, increasing α causes the system to omit ambiguous regions which are also the most likely to be misclassified. Furthermore, some classes tend to have higher confidence than others in the same scenes, suggesting that some class distinctions are intrinsically more subtle and challenging.

Figure 5 shows quantitatively the precision of our abstention policy on a pixel-wise level. As expected, increasing α greatly increases the precision of the classifier. A handful of images, approximately five of our 23 test images, include high-confidence misclassifications and are shown as outliers. Over half of the images have a precision of greater than 99% when $\alpha = 0.3$ is used, and all but the five outliers are classified with a precision greater than 99% when $\alpha = 0.2$. This has large implications for scenarios such as instrument placement, providing approximately a 22% increase in precision over non-abstaining classifiers (shown in Figure 5 when $\alpha = 1$) and a 66% increase over random classification.

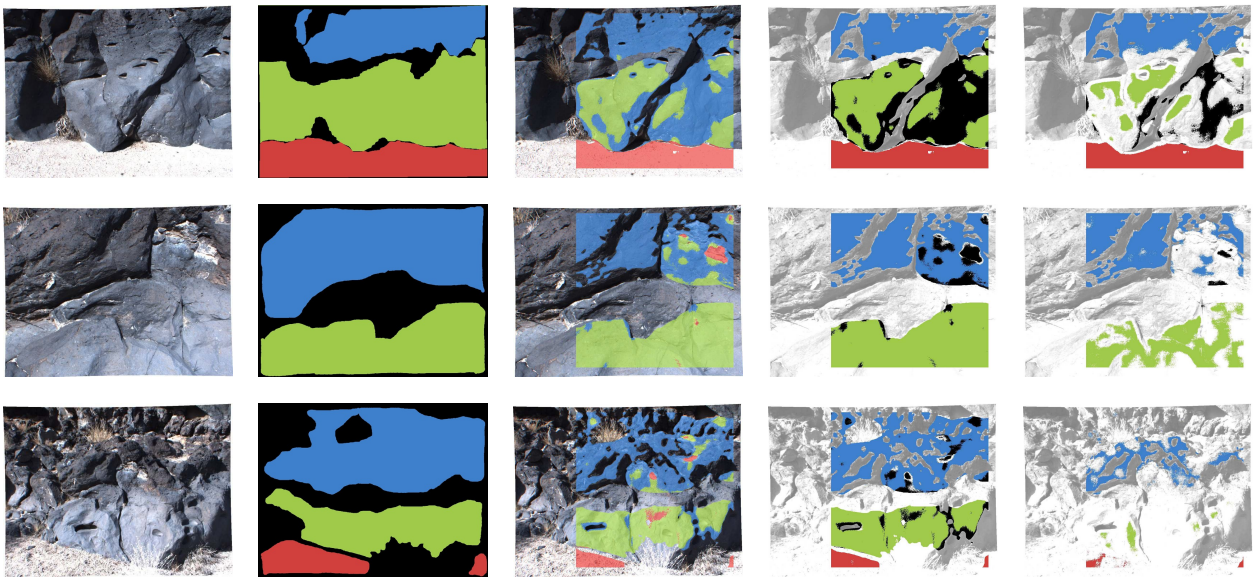


Fig. 6: Column 1 shows original images. Column 2 shows ground truth labels. Column 3 shows a scene labeling. Columns 4 and 5 show the classifications when α equals 0.6 and 0.45, respectively. Misclassified regions are colored black. Regions without ground truth or range information are excluded, and multiple range segmentations are combined As α increases the number of classified pixels is reduced, but the accuracy of the overall classification increases.

B. *Novelty Results*

We compute the top five most novel regions for every training image using each of the described algorithms, giving a total of 115 regions. Results are show in Table I.

TABLE I: Overall novelty results

Algorithm	Correctly Novel	% Correct
Mahalanobis Distance	**74**	**64.3%**
Independent Channels	61	53.0%
GMM	44	38.3%
Hou	37	32.2%

We see that the Mahalanobis distance metric outperforms all other novelty metrics, correctly finding novel regions approximately two thirds of the time. Table II shows the class breakdown of the novel regions for each method. We note that our proposed method tends to focus on vegetation, potentially due to a covariance relationship between strongly-reacting Gabor wavelets. The independent channel method is able to find more overall rock features, yet also labels a number of smooth and vesicular regions as novel.

V. Conclusion

We have demonstrated methods for enabling autonomous rover instrument targeting using probabilistic geologic surface classifications. We describe a Bayes-optimal approach to an abstaining classifier in which the overall precision of a scene classification can be increased by only evaluating pixels which have high confidence. Our results show increased precision in classification, a characteristic paramount to remote exploration spacecraft applications, while retaining the flexibility to be tailored to mission requirements. Furthermore, the generic nature of our approach lends itself to any classification method with a confidence metric or posterior class probability.

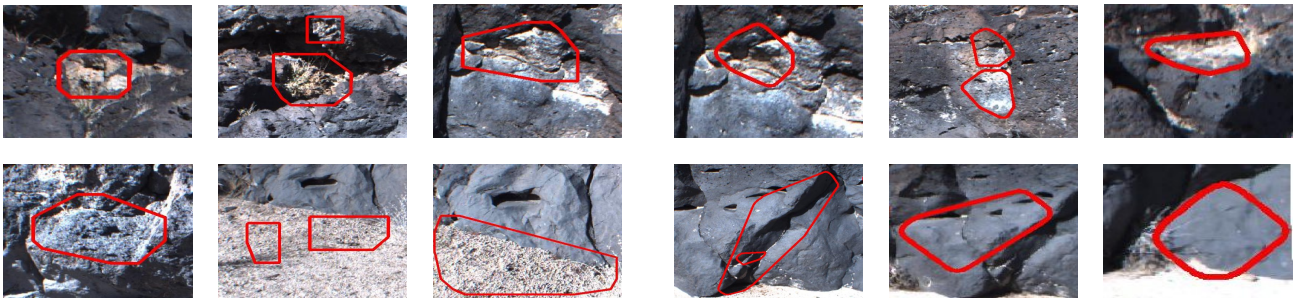
We also describe a semi-supervised method for using these precise classifications to improve in-image novelty detection. Given a scene classification, the method is straightforward and efficient to calculate and greatly improves novelty detection rates by taking advantage of contextual scene information.

We have shown the class posterior probabilities produced by random forest classification are meaningful, providing a method for greatly increasing the precision of an instrument without costly pre- or post-processing steps. We are able to decrease the overall failure rate of the pixelwise classification from approximately 14% to 2%, a seven-fold decrease over a non-abstaining classifier. This is a crucial decrease when dealing with space applications, allowing for high-precision autonomous instrument placement or scene interpretation.

By eliminating ambiguous classifications, it is possible and indeed beneficial to use the remaining classifications as class priors for other operations, such as novelty detection. When compared against state of the art saliency methods we find that utilizing this information can nearly double the efficiency of a novelty detection system. Furthermore, this method has the advantage that novelty scores are inherently normalized, allowing for the valid comparison or ranking of novel regions not just within a single window, but across an entire traverse.

VI. Acknowledgments

This research was partially carried out at the Jet Propulsion Laboratory, California Institute of Technology, with



Our method Hou's method

Fig. 7: Example novel regions. The top row shows success cases. The bottom row shows failure cases. The left three columns are calculated using our semi-supervised method. The right three columns are calculated using Hou's spectral residual method.

TABLE II: Class breakdown of novel regions

Algorithm	Unlabeled	Vesicular	Smooth	Sand	Vegetation	Rock Feature
Mahalanobis Distance	8	25	3	9	**52**	18
Independent Channels	11	18	21	8	33	**24**
GMM	18	24	16	15	28	14
Hou	21	9	46	2	15	22

support from the JPL Graduate Fellowship program. Copyright 2013 California Institute of Technology. All Rights Reserved; U.S. Government Support Acknowledged. The TextureCam project is supported by the NASA Astrobiology Science and Technology Instrument Development program (NNH10ZDA001N-ASTID). This work was supported by a NASA Office of the Chief Technologists Space Technology Research Fellowship, as well as ASTEP grant NNX11AJ87G and STTR grant NNX11CC51C.

References

[1] V. C. Gulick *et al.*, "Autonomous image analyses during the 1999 marsokhod rover field test," *Journal of Geophysical Research*, p. 77457763, 2001.

[2] R. Castaño *et al.*, "Rover traverse science for increased mission science return," *2003 IEEE Aerospace Conference Proceedings*, 2003.

[3] D. Wettergreen *et al.*, "Long-distance autonomous survey and mapping in the robotic investigation of life in the atacama desert," in *International Symposium on Artificial Intelligence, Robotics and Automation in Space (iSAIRAS)*, February 2008.

[4] J. P. L. M. Relations, "Mars science laboratory fact sheet," Jet Propulsion Laboratory, Tech. Rep., 2012. [Online]. Available: $http://mars.jpl.nasa.gov/msl/news/pdfs/MSL\_Fact\_Sheet.pdf$

[5] K. Warren-Rhodes *et al.*, "Robotic ecological mapping: Habitats and the search for life in the atacama desert," *Journal of Geophysical Research*, 2007.

[6] R. Castaño *et al.*, "Onboard autonomous rock shape analysis for mars rovers," *IEEE Aerospace Conference*, 2002. [Online]. Available: ml.jpl.nasa.gov/papers/castano/castano-IEEEAC02.pdf

[7] ——, "Opportunistic rover science: finding and reacting to rocks, clouds and dust devils," in *Aerospace Conference, 2006 IEEE*, 2006.

[8] T. A. Estlin *et al.*, "Aegis automated science targeting for the mer opportunity rover," *ACM Trans. Intell. Syst. Technol.*, 2012.

[9] L. A. Edgar *et al.*, "Sedimentary facies and bedform analysis observed from the rocknest outcrop (sols 59-100), gale crater, mars," in *Lunar and Planetary Science*, 2013.

[10] D. R. Thompson *et al.*, "Smart cameras for remote science survey," *International Symposium on Artificial Intelligence, Robotics and Automation in Space (iSAIRAS)*, 2012 (in press).

[11] L. Breiman, "Random forests," *Machine Learning*, 2001.

[12] J. Shotton *et al.*, "Semantic texton forests for image categorization and segmentation," *IEEE Conference on Computer Vision and Pattern Recognition (2008)*, 2008.

[13] W. Elazmeh *et al.*, "A framework for comparative evaluation of classifiers in the presence of class imbalance," 2006.

[14] S. Vanderlooy, I. G. Sprinkhuizen-Kuyper, E. N. Smirnov, and H. J. van den Herik, "The roc isometrics approach to construct reliable classifiers," *Intelligent Data Analysis*, 2009.

[15] C. Chow, "On optimum recognition error and reject tradeoff," *Information Theory, IEEE Transactions on*, 1970.

[16] M. Datcu *et al.*, "Introduction to the special section on image information mining for earth observation data," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 795–798, 2007.

[17] K. Lebart *et al.*, "Automatic indexing of underwater survey video: algorithm and benchmarking method," *IEEE Journal of Oceanic Engineering*, pp. 673–686, 2003.

[18] L. Tarassenko *et al.*, "Novelty detection for the identification of masses in mammograms," in *Fourth International Conference on Artificial Neural Networks*, 1995, pp. 442–447.

[19] A. Nairac *et al.*, "Choosing an appropriate model for novelty detection," in *Fifth International Conference on Artificial Neural Networks*, 1997, pp. 117–122.

[20] L. Itti *et al.*, "A model of saliency-based visual attention for rapid scene analysis," in *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 1998.

[21] X. Hou *et al.*, "Saliency detection: A spectral residual approach," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[22] K. Duncan *et al.*, "Saliency in images and video: a brief survey," *IET Computer Vision*, 2012.

[23] M. Johnson-Roberson *et al.*, "Saliency ranking for benthic survey using underwater images," in *International Conference on Control, Automation, Robotics, and Vision (ICARCV)*, 2010.

[24] K. Wagstaff *et al.*, "Science-based region of interest image compression," in *Lunar and Planetary Science*, 2004.

[25] D. R. Thompson *et al.*, "Semi-supervised eigenbasis novelty detection," *Statistical Analysis and Data Mining*, 2012.

[26] G. J. Brostow *et al.*, "Segmentation and recognition using structure from motion point clouds," *ECCV*, 2008.