Editor: Richard Doyle
Jet Propulsion Lab
rdoyle@jpl.nasa.gov

Machine Learning Tools for Automatic Mapping of Martian Landforms

Tomasz Stepinski, *Lunar and Planetary Institute*
Ricardo Vilalta, *University of Houston*
Soumya Ghosh, *University of Colorado*

Mars is at the center of current solar system exploration. Three spacecraft are presently orbiting the planet: NASA's Mars Odyssey and Mars Reconnaissance orbiters as well as the European Space Agency's Mars Express. These orbiters are surveying the entire Martian surface to understand the planet's past and the geological, climatic, and other processes responsible for its present state. They gather ever-increasing amounts of spatially extended data. Science data archived from all planetary missions prior to the 2001 Mars Odyssey mission totals approximately 5 terabytes. NASA expects that number to double over the Odyssey's duration and to increase by a factor of 10 with the newest Mars Reconnaissance mission.

This data deluge presents difficult transmission, storage, and distribution challenges. Less publicized is the scientific community's challenge to process, analyze, and ultimately turn a significant portion of the collected data into knowledge. Only a small fraction of the data is analyzed at this time because analysis still involves traditional methods that rely on visual inspection and descriptive characterization.

Automated or semiautomated tools for Martian data analysis can substantially broaden the scope of scientific inquiry. Recognizing this opportunity, we've undertaken research to apply pattern-recognition and machine-learning tools to automatic analysis and characterization of the Mars surface. This research includes machine surveys of specific landforms, such as impact craters and valley networks, and automatic generation of geomorphic maps.

Geomorphic mapping

A geomorphic map is a thematic map of topographical expressions or landforms. Machine learning can play a vital role in automating this mapping process. A learning system can employ *clustering* techniques to fully automate the discovery of meaningful landform classes. Alternatively, it can use *classification* techniques to predict unlabeled landform classes after an expert has manually labeled a representative landform sample. These techniques are often referred to as *unsupervised* and *supervised* learning, respectively (see the "Unsupervised versus Supervised Machine Learning" sidebar on p. 102). Given the size of the Martian surface that remains unmapped, they can be immensely valuable to topographical analysis.

We faced many design choices in developing appropriate tools. These included selecting an appropriate data set and machine-learning technique as well as topographic objects and their features. We based our tools on topographic data (see the "Measuring Martian Topography" sidebar on p. 104), which offers a more fundamental surface description than image data and is well suited for automated mapping. Topographic data is available as a grid-based digital elevation model. A DEM stores each pixel's elevation in a grid node. We developed both clustering and classification mapping techniques, with each approach calling for a different choice of objects and topographic features.

Clustering-based mapping

A clustering tool based on unsupervised learning offers maximum automation for geomorphic mapping. This approach can find clusters corresponding to novel topographical expressions that haven't been previously recognized. Our tool is somewhat analogous to multispectral image mapping of terrestrial land covers,[1] but it applies to topographic rather than image data. The basic topographic object is a DEM pixel that carries a vector of topographic features. The tool derives all the features from an elevation field, including variables such as slope and topographic curvature. It maps a given site by applying a clustering algorithm over all the site's feature vectors. The clustering algorithm outputs a set of mutually exclusive and exhaustive clusters. Each cluster comprises pixels carrying similar feature vectors, which are thought to be instances of a single landform.

We've applied our clustering-based mapping tool to several Martian sites.[2–5] The generated maps are qualitatively different from traditional geomorphic maps. Figure

1541-1672/07/$25.00 © 2007 IEEE
Published by the IEEE Computer Society

1a shows a standard, manually constructed Martian geomorphic map,[6] and figure 1b shows a map of the same region generated by our mapping tool. The manual map emphasizes specific landforms purposefully selected by an expert in geologic mapping. Each landform has a clearly defined semantic meaning. In contrast, the automatically generated map is more generic. Although it clearly partitions the site, some landforms lack a semantic meaning that a geologist could readily identify. This is because clusters derived under a proximity measure might not constitute a landform that an expert perceives as interesting.

For example, figure 1b divides the intercrater plateau into several landforms because they have different elevations. However, from the perspective of the particular domain expert who mapped this site, such differences were uninteresting and the intercrater plateau wasn't mapped at all. This doesn't mean that the generated map is wrong—merely that its content differs from traditional maps. In fact, such maps might be very useful for studying various statistical aspects of the Martian surface, but the context of such investigation would need to be developed outside the framework of traditional geomorphology.

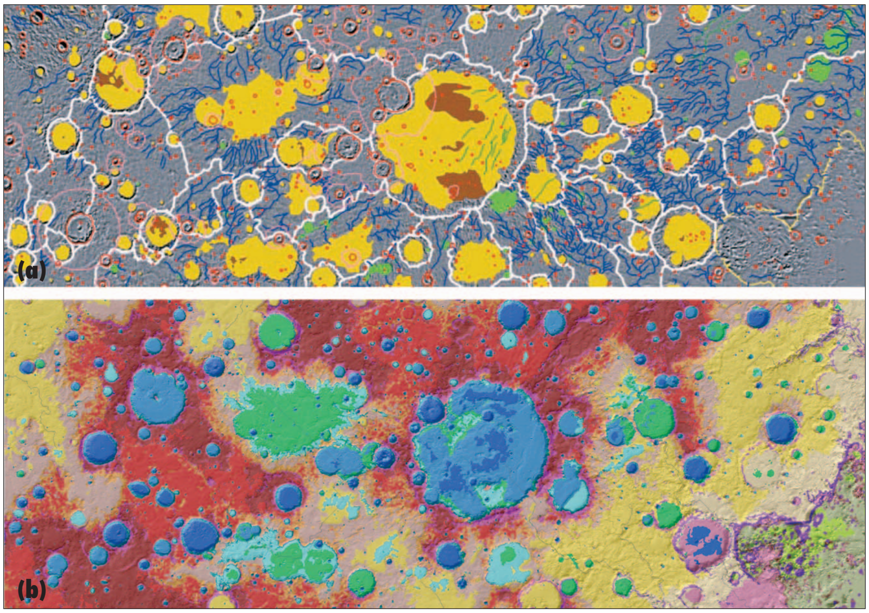


Figure 1. Geomorphic maps of Terra Cimmeria on Mars. (a) A manually constructed map reflects an expert's mapping of nine different landforms: fresh impact craters, fresh ejecta, wrinkle ridges, major valleys, dichotomy escarpment, drainage divides, basin interior mountains, intravalley basin floors, and terminal basin floors. The large gray portion of the site, consisting mostly of intercrater plateau, was assigned no landform designation. (b) The clustering-based mapping algorithm shows 19 "landforms" that represent clusters of feature vectors. An expert must interpret their meanings.

Classification-based mapping

Clustering-based mapping shows that manual geomorphic mapping relies on observation, comparison, and interpretation—processes that unsupervised learning has difficulty emulating. A mapping tool that can approximate the intangible qualities of manually constructed maps must rely on supervised learning. Therefore, we developed a classification-based mapping tool designed to map landforms as chosen and defined beforehand by an expert. Supervised learning requires a training set of terrain objects to which an expert has already assigned landform labels. The mapping tool constructs a labeling function that it then applies to label all other objects. The labeling function is an extensive, computer-derived rule set that reflects a connection between an object's numerical features and the landform labels assigned by an expert.

In supervised learning, we must carefully choose the objects to be classified. Classifying individual pixels has questionable value because pixels are too small for a human interpreter to assign them labels with a high degree of confidence. Researchers recognize a pixel's limitation as a classifiable object, and current research is focusing on a segmentation-based technique that first subdivides a site into meaningful segments for subsequent classification.[7]

We designed our classification-based mapping tool on a segmentation-based principle. It has two major components: the segmentation module and the classification module.

Segmentation module. This module divides the site into small segments that have approximately uniform pixel-based feature vectors. Computer vision researchers have studied raster segmentation intensely, but segmentation requirements in the computer-vision domain differ from those in the classification context. In computer vision, large segments are desirable as long as they contain uniform feature vectors. In our classification context, we prefer relatively small, approximately equal-sized segments, even if they cut through larger uniform fields of feature vectors. The smaller segments eliminate the danger of misclassifying a large segment and thus generating a grossly incorrect map.

On the other hand, a misclassification of a small segment results in a map that, although slightly less accurate, maintains its interpretability. In addition to terrain features, our segmentation module uses pixel spatial coordinates as additional features. This controls the segment sizes and produces segments with shapes that an algorithm can use as additional cues during classification. For example, in areas where terrain features change over a length scale larger than a segment size, the segments exhibit round shapes because the uniformity of spatial coordinates controls their extents. On the other hand, in areas where terrain features change over a length scale smaller than a segment size, the segments are elongated in a direction perpendicular to the change gradient.

Classification module. The classification module assigns a label (landform designation) to unlabeled segments (landscape elements) on the basis of patterns learned from examples. The classification subjects are segments that the segmentation module

Unsupervised versus Supervised Machine Learning

Unsupervised learning is instrumental in grouping similar data types without assuming any external classifications prior to analysis. It relies on clustering techniques to automatically discover natural similarities in data. This kind of learning is most useful in exploratory data analysis, such as pattern discovery. Post-processing of resultant clusters characterizes them by their number, size, and feature profiles. In addition, it can calculate a hierarchical tree structure of similarity between different clusters.

In our geomorphic mapping application, an unsupervised-learning algorithm groups similar pixels from a geomorphic feature space. Figure A illustrates how this process works. We employed a probabilistic algorithm working under a Bayesian framework to obtain the clusters. For sites too large to be clustered by a probabilistic algorithm, we switched to an algorithm implementing the self-organizing map and Ward hierarchical clustering method.[1]

Supervised learning is used to predict the class value of new objects. A teacher or domain expert determines the class of known objects prior to the data analysis. *Classification* is one type of supervised learning. The target classes correspond to nominal values, and the classification algorithm labels objects on the basis of their feature vectors.

In our application, supervised-learning algorithms assign landform labels to terrain segments. Figure B shows how it works. Terrain segments are more natural classification objects than pixels. To calculate segments, we employ a K-means clustering algorithm that groups pixels according to their similarity between their geomorphic features as well as their spatial proximity. A larger K value obtains smaller segments. To classify the cluster segments, we use three algorithms that represent conceptually different classification approaches and evaluate their performance using cross validation.

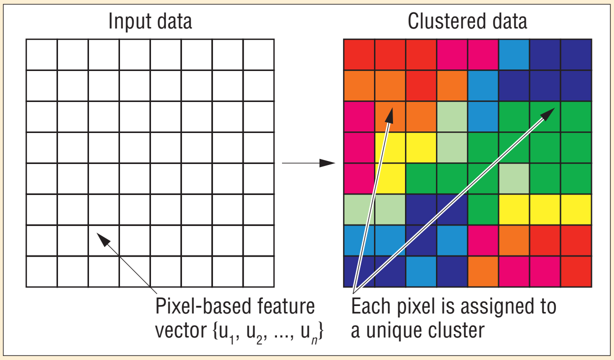

Input data
Clustered data
Pixel-based feature vector $\{u_1, u_2, ..., u_n\}$
Each pixel is assigned to a unique cluster

Figure A. Unsupervised learning. The learning algorithm groups 64 input-data pixels into nine clusters according to the similarity between pixel-based geomorphic feature vectors.

Reference

1. B.D. Bue and T.F. Stepinski, "Automated Classification of Landforms on Mars," *Computers & Geoscience*, vol. 32, no. 5, 2006, pp. 604–614.

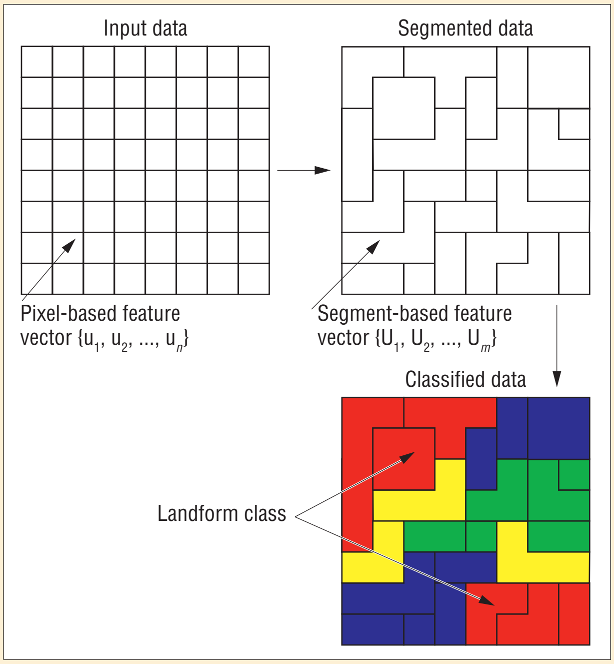

Input data
Segmented data
Pixel-based feature vector $\{u_1, u_2, ..., u_n\}$
Segment-based feature vector $\{U_1, U_2, ..., U_m\}$
Classified data
Landform class

Figure B. Supervised learning. The initial 64 pixels are first segmented into 23 segments on the basis of similarity between pixel-based feature vectors. These segments are then classified into four landform classes on the basis of similarity between segment-based feature vectors.

establishes, and a value of a segment-based feature vector determines the landform designation. A segment-based feature vector consists of physical and spatial features. The physical features are average values of pixel-based physical attributes calculated over an ensemble of pixels constituting a segment. The spatial features describe the segment itself—that is, its geometrical and neighboring properties.

To demonstrate how the classification-based mapping algorithm works, we applied it to a site on Mars located around Tisia Valles. This site represents a typical old Martian surface dominated by craters with various sizes, depths, and degrees of preservation. Figure 2a shows its topography. Planetary scientists are interested in the spatial variability of crater parameters, so the landforms of interest are crater floors, convex crater walls, concave crater walls, and intercrater plateaus. In addition, this site contains escarpments that aren't parts of craters, so we selected convex and concave ridges as additional landform classes. The algorithm segmented the site into 6,593 segments, as shown in figure 2b. An expert labeled 500 segments representative of all six landform classes to form a training set. We used this set to construct a classifier using a Support Vector Machines (SVMs) algorithm. Figure 2c shows the final map, resulting from applying the classifier to all 6,593 segments. Using the classification-based mapping tool with small labeling overhead, we obtained a map that geologists can readily interpret. By comparison, figure 2d shows the clustering-based mapping of the site. This map consists of 12 landforms, corresponding to the optimal number of clusters in the feature space, and it's not readily interpretable. Expert examination of these clusters regrouped them into four superclusters that geologists can more readily interpret.

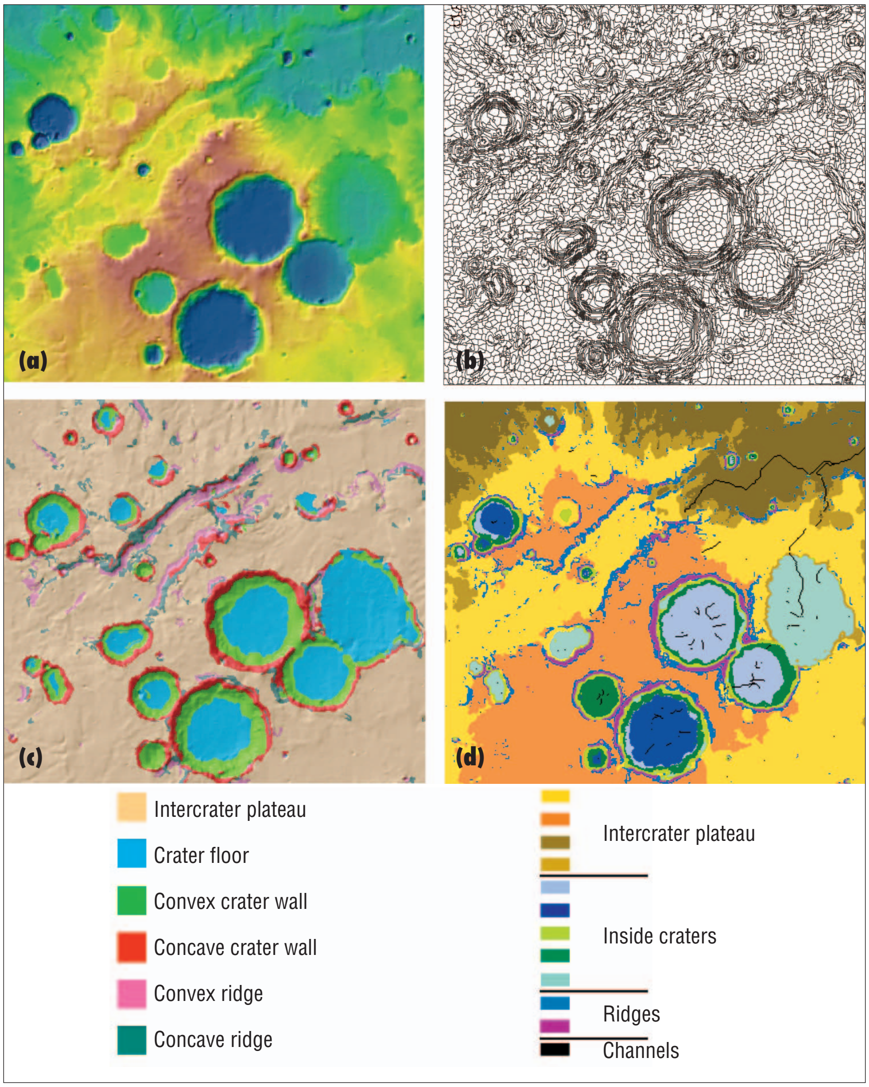


Figure 2. The Martian Tisia Valles site: (a) topographical site map approximately 215 km west to east and 192 km south to north (red-to-blue gradient indicates high-to-low elevation); (b) site segmentation into 6,593 segments; (c) geomorphic site map generated by the classification-based mapping algorithm using a Support Vector Machines classifier; and (d) site map generated by the clustering-based tool using a probabilistic algorithm working under the Bayesian framework.

Different classification algorithms

The map quality generated by the classification-based tool depends on many factors. These include available features, segmentation quality, training-set quality, and the classifier type used to assign segment labels. We used all available features, generated segmentation with desirable properties, and carefully established a representative training set. We tested three different learning algorithms to classify segments and generate geomorphic maps: naïve Bayes, bagging (using decision trees as base learners), and SVMs.

Naïve Bayes uses Bayes' theorem to estimate the landform label's posterior probability given a feature vector representing the segment. The computation of the likelihood is simplified by assuming feature independence given the landform label. Because we derive all features from the same DEM, our features don't meet the independence assumption, and we don't expect the naïve Bayes classifier to produce good maps. Rather, we expect them to serve as a baseline for comparison with the maps generated by other classifiers.

Bagging is an ensemble-learning algorithm. It generates multiple models by running a single learning algorithm multiple times over a bootstrapped sample of the training set. The final landform label is the result of voting over the contributing models—one from each bootstrap sample. Bagging has proved effective for complex data sets. It's particularly attractive when the training set is noisy, as is the case in our application. We use a C4.5 decision tree as the base learner.

Finally, SVMs use a statistical learning algorithm that works by finding an optimal hyperplane in a transformed feature space. The optimal hyperplane maximizes the separation between landform classes. SVMs exploit local data patterns and have proved effective in spatial data mining applications.

Figure 3 shows another set of Tisia Valles maps. Figure 3a shows the site topography from the southeast perspective. Figure 3b shows a geomorphic map of the six landforms obtained from an expert's manual labeling of the segments; this map serves as ground truth for maps generated on the same segmentation but with the naïve Bayes, bagging, and SVMs classifier algorithms in figures 3c, 3d, and 3e, respectively. Comparing the generated maps to the ground-truth map in figure 3b shows the naïve Bayes map to be overall inaccurate and inferior to the bagging and SVMs maps. The bagging and SVMs maps differ in character: the SVMs-generated map shows more detail, whereas the bagging-generated map shows better discrimination between crater walls and ridges.

Generalizing from training sets

The classification-based algorithm can achieve fast, accurate generation of multiple geomorphic maps if the training set is representative for all sites to be mapped. If one or more sites contain landforms not represented in the training set, the risk of mapping them incorrectly is high. To check the potential of our mapping algorithm to generalize from a single site to multiple sites, we applied it to five different sites using the training set established for Tisia Valles. The five sites—labeled Vichada, Al-Qahira, Dawes, Evros, and Margaritifer—

Measuring Martian Topography

A precise map of Martian topography, accurate around the planet to within 13 meters of elevation, is available to the scientific community as a result of measurements taken by the Mars Orbiter Laser Altimeter. The MOLA instrument went into space onboard the Mars Global Surveyor, which launched in November 1996 and arrived at Mars in September 1997.

The Laser Remote Sensing Branch of NASA Goddard Space Flight Center designed MOLA. It's a laser operating at a wavelength of 1,064 nm and emitting 8-ns pulses with energy of 40 mJ. MOLA measured the round-trip time that individual laser pulses took to travel between the spacecraft and the Martian surface. Interpolating the spacecraft orbital trajectory to the laser measurement's time and correcting for the atmosphere's refraction index gave the one-way light time between the spacecraft and the surface. Finally, subtracting the measured range from the spacecraft orbit yielded the Martian radius in a center-of-mass reference frame.

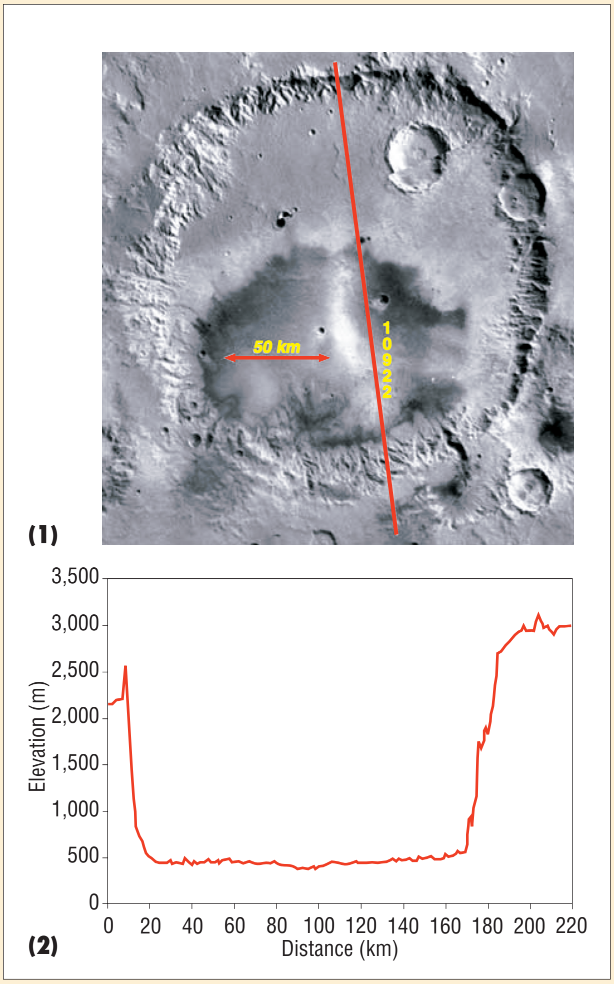


Figure C. Mars Orbiter Laser Altimeter elevation data over 180-km-diameter Dawes crater. (a) The red line shows the spacecraft's ground track against a background image on orbit 10922. (b) The MOLA measurements are combined to produce a topographic profile showing a flat-floored crater that's almost 2500 m deep.

The instrument was operational for 1,696 days and made about 640 million measurements of the Mars surface. The measurements' spatial resolution is 300 to 400 m along the track and about 1,000 m between tracks. The Martian topography is defined as the MOLA-measured planetary radius minus the radius of the geoid, which is a gravitational equipotential surface. Figure C shows an example of the MOLA measurements over a portion of a single track.

Figure D shows a global topographic map of Mars created by binning altimetry measurements from the mission's entire duration into grid-based, digital-elevation-model data structures called the MOLA Mission Experiment Gridded Data Record.[1] The MOLA team released the final MEGDR on 7 May 2003. The data is available from the Planetary Data Service geoscience node (http://pds-geosciences.wustl.edu) at resolutions of 4, 16, 32, 64, and 128 pixels per degree. Polar maps are available in resolutions of 128, 256, and 512 pixels per degree.

The algorithms for automatic geomorphic mapping described in this article use the MEGDR with resolution of 128 pixels per degree.

Reference

1. D. Smith et al., *Mars Global Surveyor Laser Altimeter Mission Experiment Gridded Data Record*, NASA Planetary Data System, MGS-M-MOLA-5-MEGDR-L3-V1.0, 2003; www.gps.caltech.edu/~marsdata/mars_MOLA_mgsl_300x_aareadme.txt.

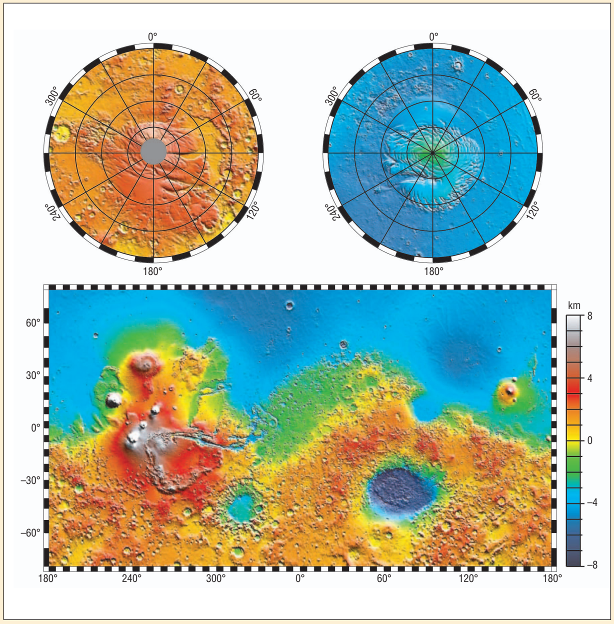


Figure D. Global topography maps of Mars. (Top left) South pole region in stereographic projection. (Top right) North pole region in stereographic projection. (Bottom) Regions between latitudes of 70S and 70N in Mercator projection showing the elevation difference between the northern and southern hemispheres.

have the same geologic character as the Tisia Valles site but differ in their dominant landforms, such as a big crater or a large valley. Figure 4 shows the generated maps for these sites. They depict the six predefined landforms quite well, except for Margaritifer, where significant confusion shows up between the “concave crater wall” landform and the “concave ridge” landform. These two landform classes are the most difficult to distinguish because they have similar physical features and differ only in a large-scale spatial context. The Margaritifer site contains segments that have feature values not well represented in the Tisia Valles site-based training set.

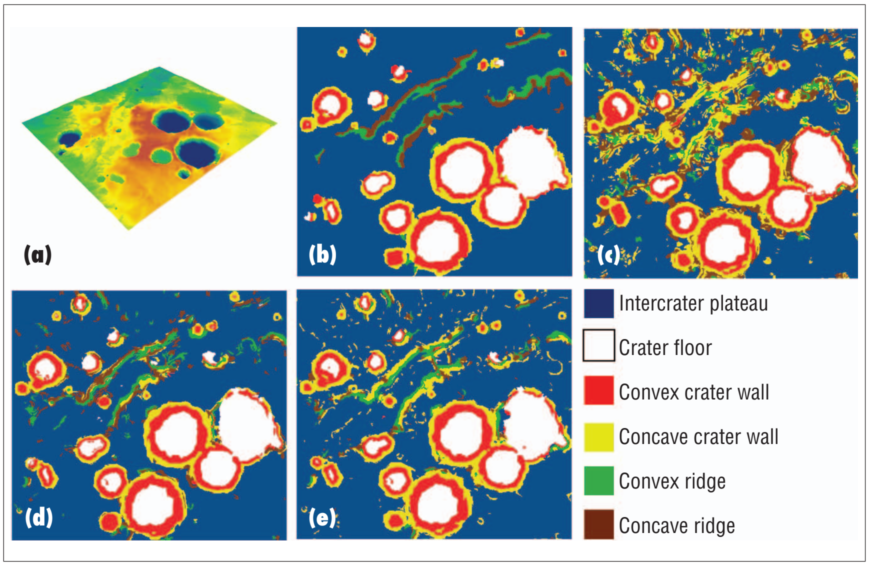


Figure 3. Geomorphic maps of the Tisia Valles site: (a) southeast perspective view of the site’s topography; (b) analyst-drawn maps of landforms; and maps generated by a classification-based tool using (c) naïve Bayes classifier, (d) bagging classifier, and (e) SVMs classifier.

Future work

The tools we’ve developed can map large regions on Mars quickly and consistently. We’re planning now to develop a method for automating the comparison and classification of geomorphic maps—an activity motivated by the large number of maps the automated mapping process could make available. Given a collection of maps, we would establish a similarity structure by calculating a “distance” between every pair of maps in the collection. The major intellectual challenge is to develop an appropriate measure of such a distance.

One approach is to use the concept of *mutual information*.[8] In information theory, mutual information of two variables quantifies the reduction in the degree of uncertainty about one variable when the other variable is known. We can measure the distance between two maps by calculating the degree of normalized mutual information (NMI) between the distribution of landform classes and the distribution of maps (a choice between the maps). A small NMI value indicates that knowing a distribution of landform classes doesn’t significantly decrease the uncertainty in distinguishing between two maps—the two landscapes are similar. On the other hand, a large NMI value indicates that knowing a distribution of landform classes helps to distinguish between the two landscapes because they’re dissimilar.

Another approach to measuring the distance between maps uses *normalized compression distance*. NCD is a practical implementation of information distance that uses file compressors or zippers. It calculates the proximity between objects that are represented as strings of characters from a finite alphabet. This “clustering by compression” approach has successfully classified objects such as text, music, and DNA

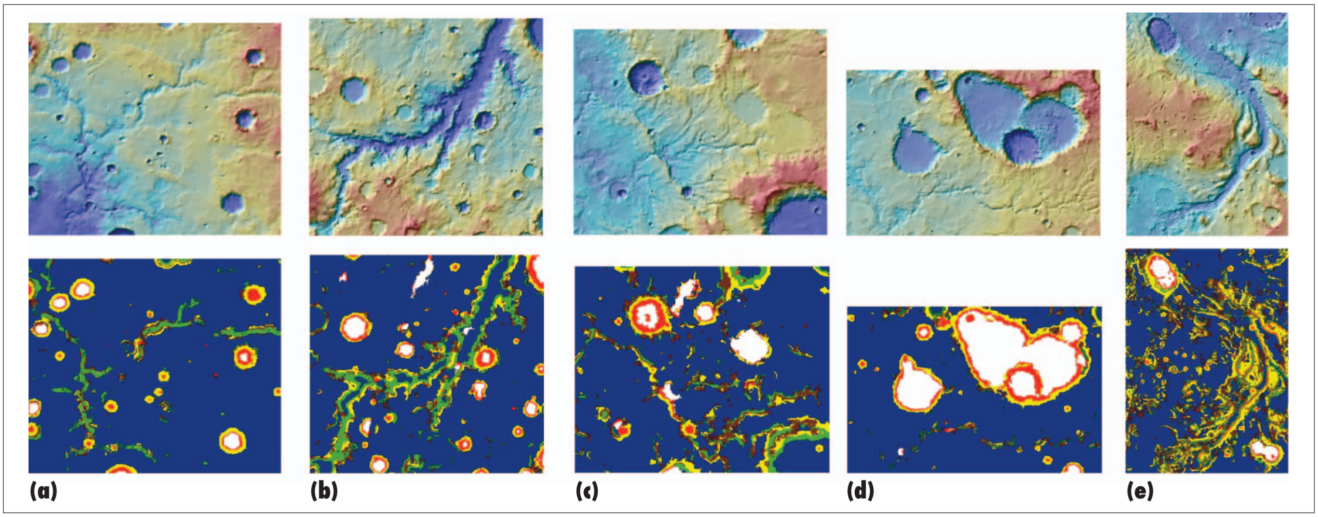


Figure 4. Geomorphic maps of (a) Vichada, (b) Al-Qahira, (c) Dawes, (d), Evros, and (e) Margaritifer Martian sites generated by the classification-based tool using SVMs. The top row shows the sites’ topography; the bottom row shows the actual maps.

sequences into meaningful categories.[9] The challenge to using it for classifying raster data sets, such as geomorphic maps, is finding a raster-string conversion method that's independent (or weakly dependent) on raster orientation.

Both these methods yield a *similarity matrix*, a data structure that must be visualized in terms of a dendrogram (binary tree) that agrees with the matrix but presents the similarity structure in a cognitively acceptable format.

Automatic map analysis will enable the study of spatial variability of surface expressions on a scale larger than individual landforms. The automated mapping we've described in this article identifies landforms from feature patterns associated with pixels. The map analysis we're planning will identify types of landscapes from the landform patterns, supporting rapid extraction of high-level information from topographic data—something only domain experts can do at this time. ■

Acknowledgments

The Lunar and Planetary Institute funded portions of this work under LPI contribution no. 1372. The Universities Space Research Association operates LPI under NASA contract CAN-NCC5-679. Other funding came from the US National Science Foundation under grants IIS-0431130, IIS-448542, and IIS-0430208 and from NASA under grant NNG06GE57G.

References

1. D. Landgrebe, "Information Extraction Principles and Methods for Multispectral and Hyperspectral Image Data," *Information Processing for Remote Sensing*, C.H. Chen, ed., World Scientific Publishing, 1999.

2. T.F. Stepinski and R. Vilalta, "Digital Topography Models for Martian Surfaces," *IEEE Geoscience and Remote Sensing Letters*, vol. 2, no. 3, 2005, pp. 260–264.

3. B.D. Bue and T.F. Stepinski, "Automated Classification of Landforms on Mars," *Computers & Geoscience*, vol. 32, no. 5, 2006, pp. 604–614.

4. T.F. Stepinski, S. Ghosh, and R. Vilalta, "Automatic Recognition of Landforms on Mars Using Terrain Segmentation and Classification," *Proc. Int'l Conf. Discovery Science*, LNAI 4265, Springer, 2006, pp. 255–266.

5. T.F. Stepinski, S. Ghosh, and R. Vilalta, "Machine Learning for Automatic Mapping of Planetary Surfaces," *Proc. 19th Innovative Applications of Artificial Intelligence Conf.*, AAAI Press, 2007, pp. 7–18.

6. R.P. Irwin and A.D. Howard, "Drainage Basin Evolution in Noachian Terra Cimmeria, Mars," *J. Geophysical Research*, vol. 107 E7, 2002, pp. 10-1–10-23.

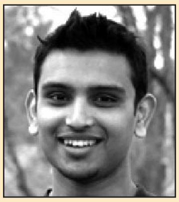
7. M. Baatz and A. Schäpe, "Multiresolution Segmentation: An Optimization Approach for High Quality Multi-Scale Image Segmentation," *Angewandte Geographische Informationsverarbeitung XII* [Applied Geographical Data Processing], Wichmann, 2000, pp. 12–23.

8. T.K. Remmel and F. Csillag, "Mutual Information Spectra for Comparing Categorical Maps," *Int'l J. Remote Sensing*, vol. 27, no. 7, 2006, pp. 1425–1452.

9. R. Cilibrasi and M.P. Vitanyi, "Clustering by Compression," *IEEE Trans. Information Theory*, vol. 51, no. 4, 2005, pp. 1523–1545.

Tomasz Stepinski is a staff scientist at the Lunar and Planetary Institute in Houston, Texas. His research interest is computational geomorphology with application to Mars and Earth. He received his PhD in applied mathematics from the University of Arizona. Contact him at tom@lpi.usra.edu.

Ricardo Vilalta is an assistant professor in the Department of Computer Science at the University of Houston. His research interests are in machine learning, statistical learning theory, data mining, and AI. He received his PhD in computer science from the University of Illinois at Urbana-Champaign. Contact him at vilalta@cs.uh.edu.



Soumya Ghosh is a graduate student in the Department of Computer Science at the University of Colorado, Boulder. His research interests lie at the intersection of machine learning and computer vision and in the applicability of this research to domains such as robotics. He received his MS in computer science from the University of Houston. Contact him at soumya.ghosh@colorado.edu.