Machine Learning for Automatic Mapping of Planetary Surfaces *

Tomasz F. Stepinski
Lunar and Planetary Institute
Houston TX 77058-1113, USA
{tom@lpi.usra.edu}

Soumya Ghosh
Dept. of Computer Science
University of Colorado at Boulder
Boulder, CO 80309-0430 USA
{soumya.ghosh@colorado.edu}

Ricardo Vilalta†
Dept. of Computer Science
University of Houston
Houston, TX 77204, USA
{vilalta@cs.uh.edu}

Abstract

We describe an application of machine learning to the problem of geomorphic mapping of planetary surfaces. Mapping landforms on planetary surfaces is an important task and the first step to deepen our understanding of many geologic processes. Until now such maps have been manually drawn by a domain expert. We describe a framework to automate the mapping process by means of segmentation and classification of landscape datasets. We propose and implement a number of extensions to the existing methodology with particular emphasis on the incorporation of machine learning techniques. These extensions result in a robust and practical mapping system that we apply on six sites on Mars. Support Vector Machines show the best mapping results with an accuracy rate of $\sim 91\%$. The resultant maps reflect the geomorphology of the sites and have appearance reminiscent of traditional, manually drawn maps. The system is capable of mapping numerous sites using a limited training set. Immediate and eventual applications of this automated mapping system are discussed in the context of planetary science and other domains.

Introduction

We are witnessing a rapid expansion of spatial datasets describing various properties of planetary surfaces. These datasets are gathered remotely by spacecrafts orbiting planets including Earth, Mars, Moon, Venus, Jupiter, and Saturn. A modern orbiter produces in the order of 10 terabytes of data from a single instrument onboard. This deluge of data challenges the ability of the scientific community to process, analyze, and ultimately turn the data into knowledge. The challenge is particularly acute in the case of the planet Mars. Due to an intense scientific and public interest in this planet, the scientific community operates currently four orbiters that send back to Earth terabytes of data expected to provide a clear understanding of past and present geological processes. Of salient interest is to obtain clues on the existence of liquid water, and ultimately on the possible existence of life. The major tool for studying Mars's surface is a geomorphic map manually drawn by a human expert on the basis of the photo-geologic interpretation of images (Wilhelms 1990; Tanaka 1994). Such manual mapping is slow and expensive; highly skilled labor is required to study and extract information. Traditional techniques are thus inappropriate to produce detailed and consistent maps on a regional scale, and to exploit high-resolution data. Innovative applications of AI technology are needed to automate the mapping process to guarantee that all data is turned into knowledge.

The task of geomorphic mapping is to divide the landscape (represented by an image, digital elevation model DEM, or other spatial datasets) into a set of mutually exclusive and exhaustive landscape elements –or regions– having specific surface patterns. These elements are later grouped into a set of clear landforms (e.g., craters, valleys, ridges, etc.); each landform representing a commonly recognizable abstraction of the elements it represents. This tasks presents significant challenges to automation including: (i) the need for a custom-designed segmentation algorithm, (ii) a classification procedure capable of assigning landform labels so that the resultant map mimics the traditional, manually derived map in appearance and content, and (iii) a substantial reduction in the size of the dataset.

In this paper we focus on the classification of landscape elements from topographic data alone as a first step towards an overall solution to the problem of automating the mapping process. Specifically, we have developed a mapping tool that implements a segmentation-based classification technique over a Martian topographic database. Our tool consists of a segmentation technique capable of procuring small segments that can guarantee accurate predictions, and a classification tool that is capable of approximating the intangible qualities of human expert mapping.

Related Work

Classification of multi-spectral images into different types of land covers (Landgrebe 1997; Landgrebe 1999) is an application of automatic mapping of spatially extended databases that has received much attention in the past. Because of this

*This work was supported by National Science Foundation under Grants IIS-0431130, IIS-448542, IIS-0430208, and by NASA under grant NNG06GE57G. A portion of this research was conducted at the Lunar and Planetary Institute, which is operated by the USRA under contract CAN-NCC5-679 with NASA. This is LPI Contribution No. 1344.

†Also affiliated to the Center for Research and Advanced Studies (CINVESTAV), Guadalajara, México.

focus on imagery data, most research represents the objects being classified in terms of individual pixels. Techniques for grouping pixels having similar spectral signatures into categories range from data clustering (e.g., k-means, self-organized maps), to hand-made rules, to supervised learning algorithms. An important problem with supervised learning is that classifying individual pixels has questionable value inasmuch as pixels are just too small to constitute units for which a label can be assigned by a human interpreter with high degree of confidence. As a result, most pixel-based classification of multi-spectral images is based on data clustering. The limitation of a pixel as a classifiable surface unit has been amply recognized lately, and the focus of research has switched to the technique of segmentation-based classification wherein a multi-spectral image is first subdivided into meaningful segments that are then subsequently classified (Baatz and Schäpa 2000). The segmentation-based approach is justified from the observation that human vision tends to separate images into regions of similar texture; meaning is easily ascribed to these regions rather than to smaller units of similar color (e.g., pixels).

Beyond classification of multi-spectral images, the topic of automatic mapping of spatial datasets (e.g., landscapes) has received little attention. This is because landscapes – more than images– are poorly suited for classification at the pixel level. Local values of topographic attributes do not uniquely determine topographic expressions. With the gain in popularity of segmentation-based classification methods, novel automated techniques have been proposed for mapping landforms of terrestrial landscapes (Gallant et al. 2005; Dragut and Blaschke 2006). These methods, however, rely on hand-made rules and fail to take advantage of AI technology.

Mars is the only planet besides the Earth for which topographic data is available in digital form (Smith et al. 2003). Automating the mapping of Martian landforms was first addressed by means of pixel-based classification (Stepinski and Vilalta 2005; Bue and Stepinski 2006). In these studies, landform categories are the result of clustering pixel-based topographic attributes using either a probabilistic clustering algorithm working under a Bayesian framework (Stepinski and Vilalta 2005), or a self-organizing map (Bue and Stepinski 2006). The major motivation behind such studies was a desire for maximum automation (minimum expert intervention) offered by the unsupervised approach. However, to obtain a map useful to planetary science, a significant manual post-clustering processing was necessary to interpret the output clusters. Moreover, the final maps had a qualitatively different character from manually drawn maps, with some landforms lacking customary geomorphic meaning. This is because a reasonable cluster derived under a proximity measure may not constitute a customary landform as perceived by a human expert.

Recognizing that automatically generated maps must conform to specifications and expectations of the particular domain of application, recent efforts explore the concatenation of a segmentation-based technique with supervised learning (Stepinski et al. 2006). This approach yields more "traditionally-looking" maps that are useful for domain experts. In this paper we describe a first solution to the automatic mapping problem by concatenating a segmentation module with classification; we also assess the feasibility of different classifiers to the mapping task.

Outline of Mapping Tool

We focus our study on spatially extended objects here referred to as *landscapes*. For each landscape, a number of datasets exists in the form of co-registered digital rasters (each describing a different attribute). Our tool accepts these attributes as input and automatically produces a categorical (thematic) map of desired landscape elements as the output. In the realm of planetary geology, a landscape is an entire region, whereas landscape elements are structures such as, for example, craters, valleys, and ridges. However, the framework presented here is domain independent and can be applied wherever rapid mapping of elements in spatially extended objects is required.

Input data from all datasets is organized into a 2-dimensional rectangular array of cells or pixels. Local landscape information is stored in each pixel in the form of a pixel-based, n-dimensional feature vector, $u(x, y) = \{u_1, u_2, \ldots, u_n\}(x, y)$. Every component of this vector corresponds to a specific landscape attribute. Fig. 1 shows an outline of our tool that consists of a segmentation and classification modules.

Segmentation

Landscape segmentation is a procedure that clusters pixels into spatially single-connected, mutually exclusive and exhaustive fragments. The procedure effectively subdivides the entire array into segments (patches) containing approximately uniform pixel-based feature vectors. Raster segmentation has been the subject of intense study in the domain of computer vision (Adams and Bischof 1994; Belongie et al. 1998; Deng and Manjunath 2001; Feng et al. 2001; Shi and Malik 2000; Wang 1998; Nock and Nielsen 2004). Although most segmentation techniques could be extended to landscape segmentation, there are clear difficulties to achieve that goal. One challenge is that a segmentation-based classification process should yield segments optimized for the subsequent classification module. Whereas in computer vision it is desirable to have large segments as long as they contain uniform feature vectors, in our context we prefer relatively small, approximately equal-sized segments, even if they cut through larger uniform fields of feature vectors. Such over-segmentation eliminates the danger of a particularly large segment being misclassified, leading to 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. Moreover, having approximately equal-sized segments assures that calculation of segment-based feature vectors is based on statistics calculated over similar-sized ensembles of pixels (explained next).

To achieve our particular segmentation procedure we integrate physical landscape attributes with spatial coordinates of pixels: $u(x, y) = \{u_1, u_2, \ldots, u_{n-2}, x, y\}(x, y)$. These

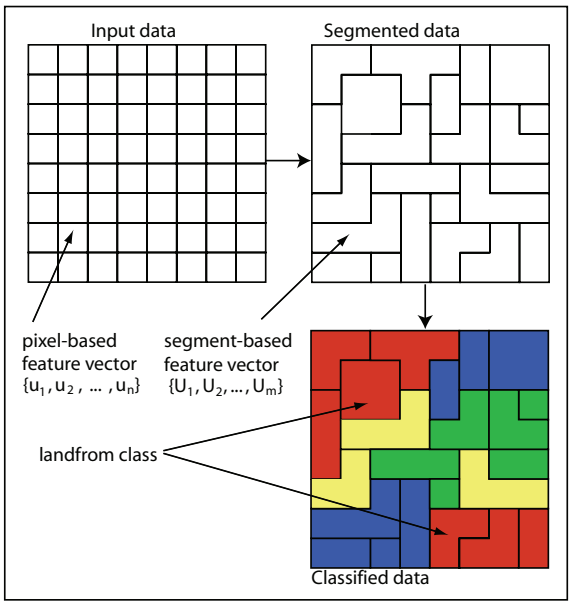

Figure 1: Diagram showing the two major components of our tool. In this illustration, 64 pixels are segmented into 23 segments on the basis of similarity between pixel-based feature vectors. These segments are classified into four landform classes on the basis of similarity between segment-based feature vectors.

additional "spatial" features enable us to control the size of the segments while providing the resultant segments with very desirable geometric properties. For example, in areas where the sub-vector of physical attributes is approximately uniform, the local gradient of u is dominated by changes in x and y leading to the formation of round-shaped segments. On the other hand, in areas where change of physical attributes dominates the local gradient of u, segments tend to exhibit an elongated shape in direction perpendicular to the gradient of the physical sub-vector. These properties constitute additional knowledge that could be exploited by the classification module.

In summary, we achieve our over-segmentation goal by integrating spatial coordinates to pixels. The actual segmentation invokes a simple k-means clustering technique applied to spatially-enriched feature vectors. The size of the segments is controlled by the value of k. The resulting k clusters do not correspond to k single-connected spatial segments; instead each cluster may contain a number of segments. To derive the final segmentation (with $K > k$ segments) we assign a unique segment identifier to each subset of a cluster corresponding to a single-connected region. The values of k and K are typically in the order of 10^3.

Classification

The classification module assigns a label (landform designation) to unlabeled segments (landscape elements) based on patterns learned from examples. Here, the subjects of classification are segments established by the segmentation module (described above). At this point we work with segment-based feature vectors. The m-dimensional segment-based feature vector, $U(i) = \{U_1, U_2, \ldots, U_m\}(i)$, $i = 1, \ldots, K$ describes landscape attributes at the level of an individual segment, as well as the spatial attributes of the segment itself. It is important to stress that these segment-based feature vectors, which are used for classification purposes, should not be mistaken with pixel-based feature vectors, which are used for segmentation purposes. 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 (i.e., its geometrical and neighboring properties).

A training set is established as a representative sample of segment-based feature vectors for which landform labels were assigned by a human interpreter from a limited collection of a number of (L) possible designations. This training set is used to build a classifier (data model) that is subsequently applied to all unlabeled segments to complete the mapping process.

We apply three different classification algorithms to obtain the final map: Naive Bayes, Bagging (using decision trees as base learners), and Support Vector Machines (SVMs). Naive Bayes is a simple probabilistic classifier that assumes feature independence given the class label. We include this algorithm as a baseline for comparison. Bagging is a meta-learning algorithm that generates an ensemble of training sets by randomly drawing, with replacement, samples from the original training set (Breiman 1996). The final class label is the result of voting over the contributing models (one from each bootstrap sample). We use a decision tree (C4.5) as the base learner (Quinlan 1993). Finally, SVMs operate by finding a hyper-plane in a transformed feature space that maximizes the margin (i.e., distance between the hyperplane and closest points from every class to that hyperplane) (Vapnik 1995). We employ these algorithms as implemented in the software package WEKA (Witten and Frank 2000) using default parameters.

Application to Martian Surfaces

Martian topographic data is contained in global digital elevation maps (DEMs) with a resolution of $\sim$ 500 meters/pixel (Smith et al. 2003). We demonstrate the feasibility of our mapping tool by generating a six-landform geomorphic map, geared toward rapid characterization of impact craters. We focus on six different sites on Mars referred to as Tisia, Al-Qahira, Dawes, Evros, Margaritifer, and Vichada. We aim at identifying six landforms (i.e., $L = 6$ possible class labels): crater floors, convex crater walls, concave crater walls, convex ridges, concave ridges, and inter-crater plateau. The choice of the first three landforms stems from our interest in the automatic characterization of impact craters. The next two landforms are justified because the sites contain escarpments that are not parts of craters. The inter-crater plateau is a dominant landform on Mars that must be included on any geomorphic map.

The selection of physical pixel-based features is dictated by the choice of landforms. We have selected slope (s), cur-

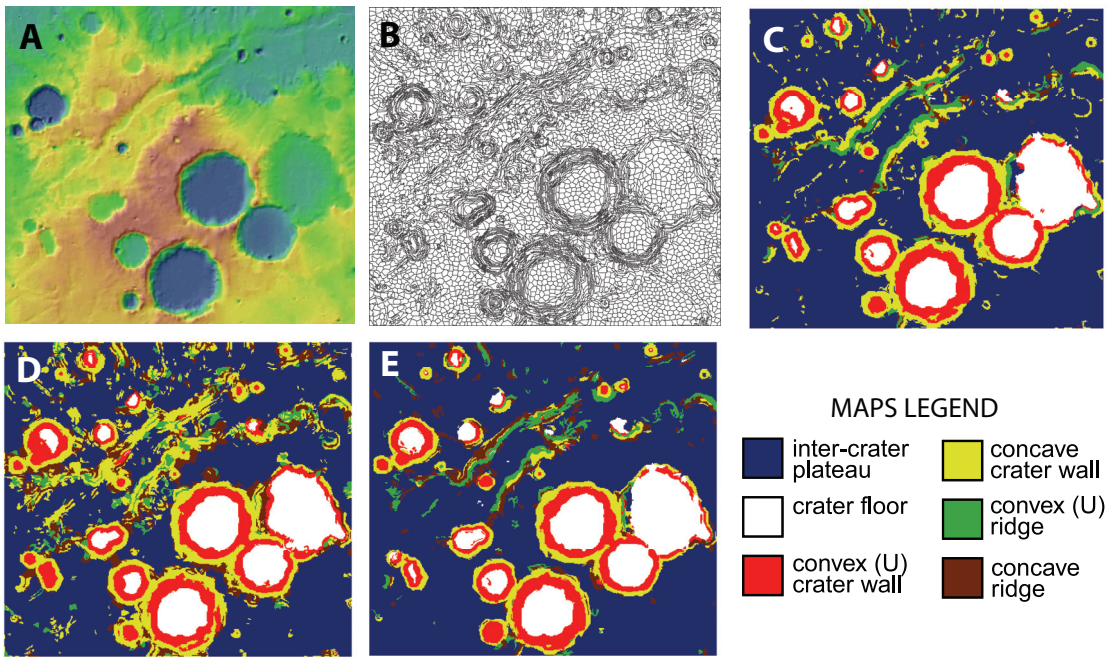

Figure 2: (A) Topography of Tisia site, red-to-blue gradient indicate high-to-low elevation. (B) Segmentation achieved by k-means clustering. (C) Automatically generated geomorphic map of Tisia site using SVMs. (D) Map generated using Naive Bayes. (E) Map generated using Bagging.

vature (κ), and flood (f) as landscape attributes; adding spatial coordinates makes $n = 5$ and our pixel-based feature vector is $u(x, y) = \{s, \kappa, f, x, y\}(x, y)$. The slope, $s(x, y)$, is the local rate of maximum elevation change. The (profile) curvature, $\kappa(x, y)$, measures the local change of slope angle ($\kappa > 0$ correspond to convex topography, whereas $\kappa < 0$ correspond to concave topography). The flood, $f(x, y)$, is a binary variable such that pixels located inside topographic basins have $f(x, y) = 1$, and all other pixels have $f(x, y) = 0$. These attributes are calculated directly from the DEM dataset using a 3×3 pixels moving window.

In this application we use $m = 13$ segment-based features. The dimensional feature vector characterizing each segment is described as follows:

$$U = \left\{ \bar{s},\ \bar{\kappa},\ \bar{f},\ SCI,\ a_1^s,\ a_2^s,\ a_3^s,\ a_1^\kappa,\ a_2^\kappa,\ a_3^\kappa,\ a_1^f,\ a_2^f,\ a_3^f \right\}$$

The first three coordinates of U are physical features – averages of s, κ, and f calculated over the extent of the segment. The remaining ten coordinates of U are spatial features. The fourth coordinates of U is the **S**hape **C**omplexity **I**ndex (SCI) (Hengl 2003), computed as follows:

$$SCI = \frac{P}{2\pi r};\quad r = \sqrt{\frac{A}{\pi}}$$

where P is the perimeter of the segment boundary, A is the area of the segment, and r is the radius of a circle with the same surface area as the segment. SCI is essentially a perimeter-to-circumference ratio, a crude measure of circularity (values ~ 1.0 signify a circular shape). The last nine coordinates of U encapsulate information about neighborhood properties. Ideally, we would like to know landform classes of segment neighbors, but such information is not available prior to classification. However, a preliminary categorization of segments into low, medium, and high slope is possible on the basis of statistics of the values of $\bar{s}(i), i = 1, 2, \ldots, K$. Such categorization is used to calculate neighborhood properties of a segment $\{a_1^s, a_2^s, a_3^s\}$, where a_j^s, $j = 1, 2, 3$, is the percentage of the object boundary adjacent to neighbors belonging to slope category j. Similar neighborhood properties, $\{a_1^\kappa, a_2^\kappa, a_3^\kappa\}$, $\{a_1^f, a_2^f, a_3^f\}$, are calculated on the basis of curvature and flood values.

Empirical Results

We first apply our tool to the Tisia site. The topography of this site is shown in Fig. 2A and serves as a visual ground truth for maps generated by our tool. The segmentation ($k = 5000$) of the site (composed of 163,240 pixels) results in 6593 segments that are shown in Fig. 2B. A total of 829 segments, representing all six landform classes, were labeled by a domain expert and used for training. The six-landform geomorphic maps generated by applying SVMs, Naive Bayes, and Bagging are shown in Figs. 2C, 2D, and 2E, respectively. Accuracy is measured using 10-fold cross-validation.

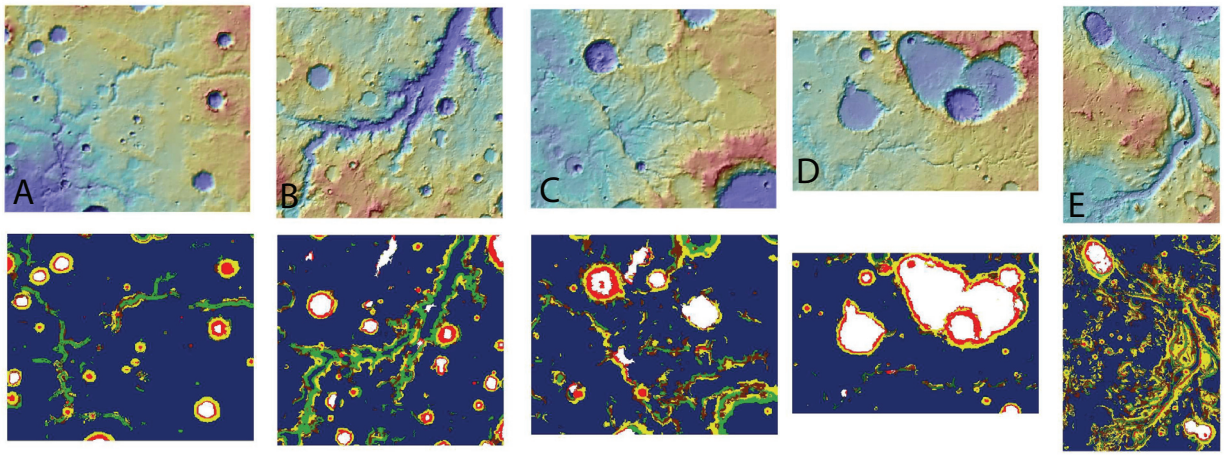

Figure 3: Automatically generated geomorphic maps of Vichada (A), Al-Qahira (B), Dawes (C), Evros (D), and Margaritifer (E) sites using SVMs. The top row shows the sites' topography; the bottom row shows the actual maps.

Accuracy rates are as follows[1]: SVMs: $91.06(2.5)$, Naive Bayes: $88.65(3.17)$, and Bagging: $90.95(2.57)$. According to expert input, visual agreement with topography (Fig 1A) is best achieved with SVMs and Bagging.

The benefit of automating the mapping process lies on the fast and accurate generation of multiple maps. We have used the training set established for the Tisia site to map all other five sites. Fig. 3 shows the resultant maps generated using SVMs. All maps reflect the actual topography well, except for the Margaritifer site where there is a significant confusion between the "concave crater wall" landform and the "concave ridge" landform. These two landform classes are most difficult to distinguish because they have very similar physical attributes and differ only in a large scale spatial context. The Margaritifer site contains segments that are characterized by segment-based feature vectors not well represented in the Tisia site-based training set.

Discussion and Future Work

Conventional ways of mapping the geomorphology of a planetary site take place by manually drawing a map on the basis of visual interpretation of features contained on images. This requires a significant commitment of skilled human resources –an impractical proposition for mapping large regions characterized by high resolution.

In this paper we present an automated mapping tool that employs AI technology (i.e., machine learning tools), aimed at fast generation of large number of maps that closely resemble traditional, manually drawn maps. Our approach incorporates spatial coordinates into the pixel-based feature vectors; this helps to achieve over-segmented images that facilitate the accurate classification of segments. Empirical results show our automatic mapping of planetary surfaces feasible and practical. Our tool produces maps that, in appearance and context, mimic manually derived maps. These maps represent a significant improvement over maps generated after clustering single pixels on the basis of similarity of pixel-based feature vectors (Stepinski and Vilalta 2005; Bue and Stepinski 2006). Fig. 4 shows the Tisia site divided into landforms using the pixel-clustering technique (Stepinski and Vilalta 2005). The optimal number of 12 clusters does not translate into 12 landforms of interest. For the domain user (i.e., the geologist) the map produced by our tool (see Fig. 2C) is clearly superior to the map shown in Fig. 4.

This initial success invites for further improvements. First, we are working on a hierarchical segmentation module with a parameter k rather small (e.g., ~ 10 instead of the current value of 5000). In a second stage each resultant segment is itself segmented (keeping a relatively small value of k). This process can be repeated until a fine segmentation is achieved. This form of hierarchical segmentation enables us to be more flexible on the types of regions where over-segmentation is necessary. Moreover, the hierarchical structure can be exploited by the classification algorithm by learning concepts at different levels of abstraction.

We also plan to incorporate meta-learning techniques into the classification module. The goal of such techniques would be to provide continuous adaptation of a classifier to new sites. In our current application the training set (a subset of Tisia site segments) was successfully applied to map other sites of similar character. However, application to Margaritifer, a site of somewhat different character, resulted in a map of degraded quality. In an adaptive scheme, segments not well represented in the training set would be labeled "unknown" and left for manually labeling. This improvement is particularly valuable for mapping large number of diverse sites.

On the application site, our tool will be first used for characterization of Martian craters. A large number of sites will be mapped to identify crater components (e.g., floor, walls, rim, etc.). Identification of components is necessary for cal-

[1]Numbers enclosed in parentheses represent standard deviations.

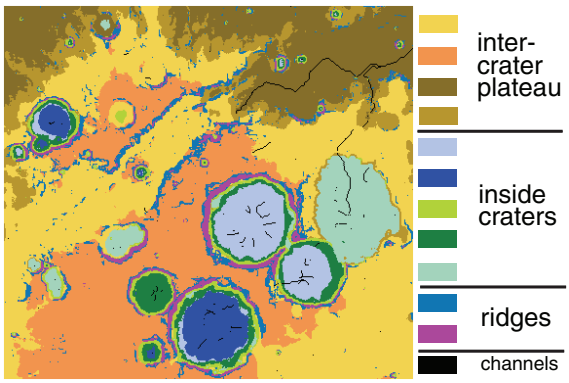

Figure 4: Landforms identified by clustering individual pixels (Stepinski and Vilalta 2005). Twelve clusters (landforms) are indicated by different colors, and are manually grouped into larger clusters with geological meaning.

culating crater characteristics (such as the size of the floor, crater depth, the curvature of the walls, etc.) of interest to the domain user. The end result will be a comprehensive catalog of Martian craters. Eventually, our tool will be used to map a large number of sites on Mars; these maps will be used for quantitative, automated comparative analysis of landscapes located at different regions of the planet using map comparison techniques (Remmel and Csillag 2006). Our tool can also be applied without major modification to terrestrial geospatial datasets: DEMs, multi-spectral images, census data, etc.; applications abound in ecology, urban development, forestry, and agriculture.

References

Adams, R.; and Bischof, L. 1994. Seeded Region Growing. *IEEE Trans. Pattern Analysis and Machine Intelligence*. 16(6):641–647.

Baatz, M.; and Schäpe, A. 2000. Multiresolution Segmentation: An Optimization Approach for High Quality Multi-Scale Image Segmentation. In Strobl, J. et al. (eds.): Angewandte Geographische Infor-mationsverarbeitung XII. Wichmann, Heidelberg. 12-23.

Belongie, S.; Carson, C.; Greenspan, H.; and Malik, J. 1998. Color and Texture-based Image Segmentation using EM and its Application to Content-Based Image Retrieval. In *Proc. of Sixth IEEE Int. Conf. Comp. Vision*, 675-682.

Breiman, L. 1996. Bagging predictors, *Machine learning*, 24(2):123-140.

Bue, B.D.; and Stepinski, T.F. 2006. Automated Classification of Landforms on Mars. *Computers & Geoscience* 32(5):604–614.

Deng, Y.; and Manjunath, B.S. 2001. Unsupervised Segmentation of Color-Texture Regions in Images and Video. *IEEE Trans. Pattern Analysis and Machine Intelligence* 23(8):800–810.

Dragut, L.; and Blaschke, T. 2006. Automated Classification of Landform Elements Using Object-Based Image Analysis, *Geomorphology* 81:330–344.

Feng, H.; Castanon, D.A.; and Karl, W.C. 2001. A Curve Evolution Approach for Image Segmentation Using Adaptive Flows. *Proc of Eight IEEE Int. Conf. Comp. Vision* 494–499.

Gallant, A.L.; Brown, D.D.; and Hoffer, R.M. 2005. Automated Mapping of Hammond's Landforms, *IEEE Geoscience and Remote Sensing Letters* 2(4):384–288.

Hengl, T. 2003. Pedometric Mapping: Bridging the Gaps Between Conventional and Pedometric Approaches. Ph.D. diss., Wageningen University.

Landgrebe, D. 1997. The Evolution of Landsat Data Analysis. *Photogrammetric Ingeenering and Remote Sensing*, LXIII(7):859–867.

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

Nock, R.; and Nielsen, F. 2004. Stochastic Region Merging. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(11):1452–1458.

Quinlan, J.R. 1993. *C4.5: Programs for Machine Learning*, San Francisco: Morgan Kaufmann.

Remmel, T.K.; and Csillag, F. 2006. Mutual Information Spectra for Comparing Categorical Maps. *Inter. Journal of Remote Sensing* 27(7):1425-1452.

Shi, J.; and Malik, J. 2000. Normalized Cuts and Image Segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence* 22(8):888–905.

Smith, D.; Neumann, G.; Arvidson, R.E.; Guinness, E.A.; and Slavney, S. 2003. Mars Global Surveyor Laser Altimeter Mission Experiment Gridded Data Record, *NASA Planetary Data System*, MGS-M-MOLA-5-MEGDR-L3-V1.0.

Stepinski, T.F.; and Vilalta, R. 2005. Digital Topography Models for Martian Surfaces. *IEEE Geoscience and Remote Sensing Letters* 2(3):260–264.

Stepinski, T.F.; Ghosh, S.; and Vilalta, R. 2006. Automatic Recognition of Landforms on Mars Using Terrain Segmentation and Classification, In Proceedings of the International Conference on Discovery Science:255–266.

Tanaka, K.L. 1994. The Venus Geologic Mappers' Handbook, *U.S. Geol. Surv. Open File Rep.*:99–438.

Vapnik, V.N. 1995. *The Nature of Statistical Learning Theory*, Springer.

Wang, J.P. 1998. Stochastic Relaxation on Partitions with Connected Components and its Application to Image Segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence* 20(6):619–636.

Wilhelms, D.E. 1990. Geologic Mapping. *Planetary Mapping*, Greeley. R.; and Batson, R. Eds. Cambridge, UK: Cambridge Univ. Press, 209–260.

Witten, I. H.; and Frank, E. 2000. *DataMining: Practical Machine Learning Tools and Techniques with Java Implementations*. Academic Press, London U.K.