Machine Learning, 30, 165–194 (1998)

Learning to Recognize Volcanoes on Venus

MICHAEL C. BURL burl@aig.jpl.nasa.gov
Jet Propulsion Laboratory, MS 525-3660, 4800 Oak Grove Drive, Pasadena, CA 91109, USA and California Institute of Technology

LARS ASKER
Jet Propulsion Laboratory, MS 525-3660, 4800 Oak Grove Drive, Pasadena, CA 91109, USA and Stockholm University

PADHRAIC SMYTH
Jet Propulsion Laboratory, MS 525-3660, 4800 Oak Grove Drive, Pasadena, CA 91109, USA and University of California, Irvine

USAMA FAYYAD
Jet Propulsion Laboratory, MS 525-3660, 4800 Oak Grove Drive, Pasadena, CA 91109, USA and Microsoft Research

PIETRO PERONA
California Institute of Technology and Università di Padova

LARRY CRUMPLER
Brown University and New Mexico Museum of Natural History & Science

JAYNE AUBELE
Brown University and New Mexico Museum of Natural History & Science

Editors: Ron Kohavi and Foster Provost

Abstract. Dramatic improvements in sensor and image acquisition technology have created a demand for automated tools that can aid in the analysis of large image databases. We describe the development of JARtool, a trainable software system that learns to recognize volcanoes in a large data set of Venusian imagery. A machine learning approach is used because it is much easier for geologists to identify examples of volcanoes in the imagery than it is to specify domain knowledge as a set of pixel-level constraints. This approach can also provide portability to other domains without the need for explicit reprogramming; the user simply supplies the system with a new set of training examples. We show how the development of such a system requires a completely different set of skills than are required for applying machine learning to "toy world" domains. This paper discusses important aspects of the application process not commonly encountered in the "toy world," including obtaining labeled training data, the difficulties of working with pixel data, and the automatic extraction of higher-level features.

Keywords: machine learning, pattern recognition, learning from examples, large image databases, data mining, automatic cataloging, detection of natural objects, Magellan SAR, JARtool, volcanoes, Venus, principal components analysis, trainable

1. Introduction

Detecting all occurrences of an object of interest in a set of images is a problem that arises in many domains, including industrial product inspection, military surveillance, medical diagnosis, astronomy, and planetary geology. Given the prevalence of this problem and the

fact that continued improvements in image acquisition and storage technology will produce ever-larger collections of images, there is a clear need for algorithms and tools that can be used to locate automatically objects of interest within such data sets.

The application discussed in this paper focuses on data from NASA/JPL's highly successful Magellan mission to Venus. The Magellan spacecraft was launched from Earth in May of 1989 with the objective of providing global synthetic aperture radar (SAR) mapping of the entire surface of Venus. In August of 1990 the spacecraft entered a polar elliptical orbit around Venus. Over the next four years Magellan returned more data than all previous planetary missions combined (Saunders et al., 1992), specifically, over 30,000 1024×1024 pixel images covering 98% of the planet's surface. Although the scientific possibilities offered by this data set are numerous, the sheer volume of data is overwhelming the planetary geology research community. Automated or semi-automated tools are necessary if even a fraction of the data is to be analyzed (Burl et al., 1994a).

1.1. Scientific Importance

Volcanism is the most widespread and important geologic phenomenon on Venus (Saunders et al., 1992), and thus is of particular interest to planetary geologists studying the planet. From previous low-resolution data, it has been estimated that there are on the order of one million small volcanoes (defined as less than 20 km in diameter) that will be visible in the Magellan imagery (Aubele and Slytua, 1990). Understanding the global distribution and clustering characteristics of the volcanoes is central to understanding the geologic evolution of the planet (Guest et al., 1992; Crumpler et al., 1997). Even a partial catalog including the size, location, and other relevant information about each volcano would enable more advanced scientific studies. Such a catalog could potentially provide the data necessary to answer basic questions about the geophysics of Venus, questions such as the relationship between volcanoes and local tectonic structure, the pattern of heat flow within the planet, and the mechanics of volcanic eruption.

1.2. Impact of an Automated System

A catalog of large Venusian volcanoes (greater than 20 km in diameter) has been completed manually (Crumpler et al., 1997; Stofan et al., 1992). However, by optimistic estimates the time for a geologist to generate a comprehensive catalog of small volcanoes on Venus would be ten to twenty man-years. In fact, in our experiments we have found that humans typically become quite fatigued after labeling only 50–100 images over a few days. Thus, large-scale sustained cataloging by geologists is not realistic even if they had the time to devote this task. An automated system would provide many benefits, including the ability to maintain a uniform, objective standard throughout the catalog, thereby avoiding the subjectivity and drift common to human labelers (Cooke 1991; Poulton 1994). Even a partially automated system that functions as an "intelligent assistant" would have considerable impact.

1.3. Motivation for a Learning Approach

There are two approaches one could follow for building an automated volcano cataloging system. The first would be to develop hand-coded, volcano-specific detectors based on a high-level description of the problem domain provided by human experts. There are, however, a number of drawbacks to this method. Geologists are quite good at identifying examples of the objects of interest, but it is often difficult for them to identify precisely which characteristics of each volcano in the image led to its detection. High-level features, such as circularity or the presence of a summit pit, are difficult to translate into pixel-level constraints. Finally, visual recognition of localized objects is a problem that arises in many domains; using a hand-coded approach would require a significant amount of reprogramming for each new domain.

The second approach is to use learning from examples. Since the geologists can identify examples of volcanoes with relative ease, their domain knowledge can be captured implicitly through the set of labeled examples. Using a learning algorithm, we can extract an appearance model from the examples and apply the model to find new (previously unseen) volcanoes. This approach can also provide portability since the user must merely supply a new set of training examples for each new problem domain—in principle no explicit reprogramming is required.

1.4. Related Work

Most prior work on automated analysis of remotely sensed imagery has focused on two problems: (1) classification of homogeneous regions into vegetation or land-use types, e.g., (Richards, 1986) and (2) detection of man-made objects such as airports, roads, etc. The first technique is not applicable to the volcano detection problem, and the second is inappropriate because naturally occurring objects (such as volcanoes) possess much greater variability in appearance than rigid man-made objects. Several prototype systems (Flickner et al. 1995; Pentland, Picard, and Sclaroff, 1996; Picard and Pentland, 1996) that permit querying by content have been developed in the computer vision community. In general, these systems rely on color histograms, regular textures, and boundary contours or they assume that objects are segmented and well-framed within the image. Since the small volcanoes in the Magellan imagery cannot be characterized by regular textures or boundaries, none of these approaches is directly applicable to the volcano cataloging problem. (For example, we found that the edge contrast and noise level in the SAR images did not permit reliable edge-detection.)

In general, there has been relatively little work on the problem of finding natural objects in a cluttered background when the objects do not have well-defined edge or spectral characteristics. Hough transform methods were used for the detection of circular geologic features in SAR data (Cross, 1988; Skingley and Rye, 1987) but without great success. For the small volcano problem, Wiles and Forshaw (Wiles and Forshaw, 1993) proposed using matched filtering. However, as we will see in Section 3, this approach does not perform as well as the learning system described in this paper. Fayyad and colleagues (Fayyad et al., 1996) developed a system to catalog sky objects (stars and galaxies) using decision tree classification methods. For this domain, segmentation of objects from the background and

conversion to a vector of feature measurements was straightforward. A good set of features had already been hand-defined by the astronomy community so most of the effort focused on optimizing classification performance. In contrast, for the Magellan images, separating the volcanoes from the background is quite difficult and there is not an established set of pixel-level features for volcanoes.

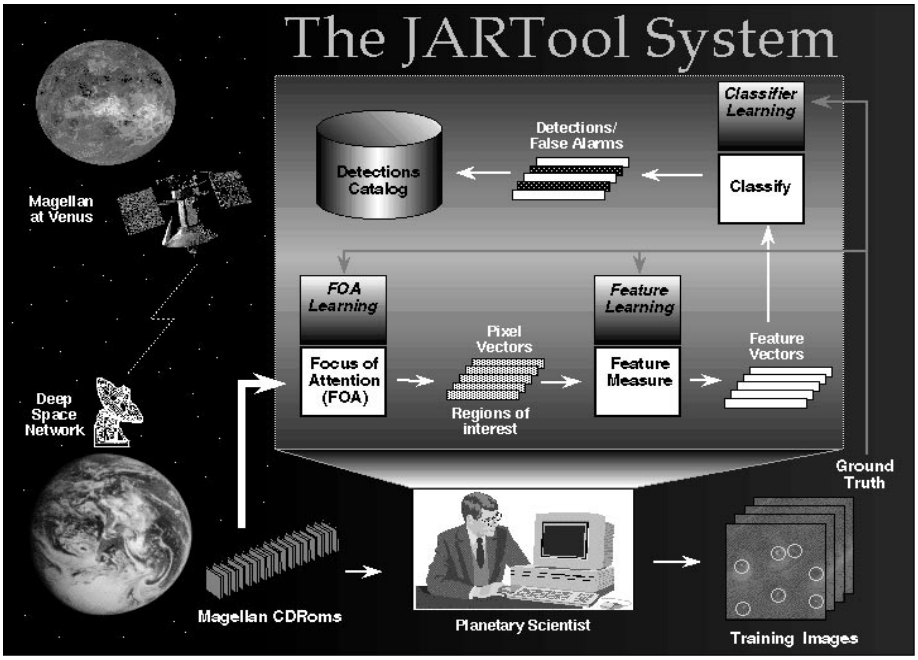
1.5. The JARtool System

JARtool (JPL Adaptive Recognition Tool) is a trainable visual recognition system that we have developed in the context of the Magellan volcano problem. The basic system is illustrated in Figure 1. Through a graphical user interface (GUI), which is shown in Figure 2, a planetary scientist can examine images from the Magellan CD-ROMs and label examples in the images. The automated portion of the system consists of three components: focus of attention (FOA), feature extraction, and classification. Each of these components is trained for the specific problem of volcano detection through the examples provided by the scientist.

The specific approach taken in JARtool is to use a matched filter derived from training examples to focus attention on regions of interest within the image. Principal components analysis (PCA) of the training examples provides a set of domain-specific features that map high-dimensional pixel data to a low-dimensional feature space. Supervised machine learning techniques are then applied to derive a mapping from the PCA features to classification labels (volcano or non-volcano). The PCA technique, which is also known as the discrete Karhunen-Loeve transform (Fukunaga, 1990), has been used extensively in statistics, signal processing, and computer vision (Sirovich and Kirby, 1987; Turk and Pentland 1991; Moghaddam and Pentland, 1995; Pentland et al., 1996) to provide dimensionality reduction. PCA seeks a lower-dimensional subspace that best *represents* the data. An alternate approach is linear discriminant analysis (LDA) (Duda and Hart, 1973; Swets and Weng, 1996), which seeks a subspace that maximizes the separability between classes. However, in the volcano context, the non-volcano class is so complex that LDA methods at the pixel-level do not work well.

1.6. Outline

In Section 2 the JARtool system design process is described with an emphasis on the real-world issues that had to be addressed before standard "off-the-shelf" classification learning algorithms could be applied. In Section 3 we provide an empirical evaluation of our learning-based system and compare the performance with that of human experts. Section 4 indicates the current status of the project. In Section 5 we discuss the lessons learned from the project and how these application lessons could provide useful directions for future machine learning research. In Section 6 we conclude with a summary of the main points of the article and indicate directions for future work.



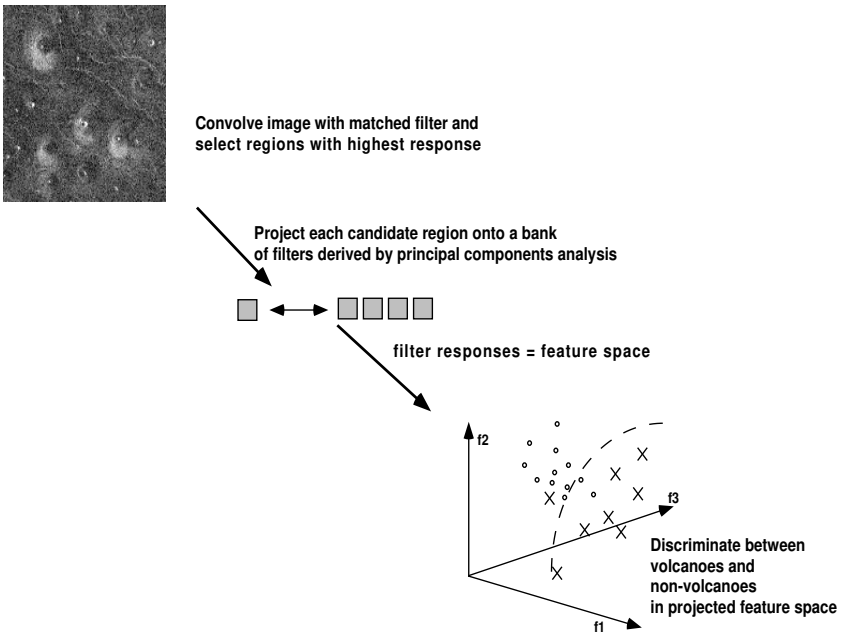


Figure 1. Overview of the JARtool system.

Figure 2. In addition to the standard image browsing and labeling capabilities, the JARtool graphical user interface enables the user to learn models of an object and then look for novel instances of the object. The image displayed here is a 30km $\times$ 30km region on Venus containing a number of small volcanoes. (See Figure 5 to find out where the volcanoes are located.)

2. System Design

2.1. Magellan Imagery

Pettengill and colleagues (Pettengill et al., 1991) give a complete description of the Magellan synthetic aperture radar system and associated parameters. Here we focus only on how the imaging process affects the appearance of the volcanoes in the dataset.

Figure 2 shows a 30km $\times$ 30km area of Venus as imaged by Magellan. This area is located near lat 30°N, lon 332°. Illumination is from the lower left and the pixel spacing[1] is 60m. Observe that the larger volcanoes in the image have the classic radar signature

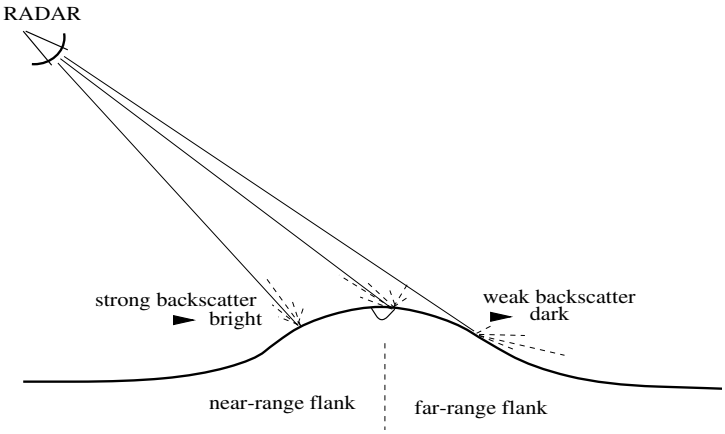


Figure 3. Because of the topography, the near-range volcano flanks scatter more energy back to the radar and appear bright. In contrast, the far-range flanks scatter energy away and appear dark.

one would expect based on the topography; that is, the side of the volcano closest to the radar (near-range) appears bright and the side away from the radar (far-range) appears dark. The reason is that the near-range side scatters more energy back to the sensor than the surrounding flat plains, while the far-range side scatters most of the energy off into space. The brightness of each pixel is proportional to the log of the returned energy, so volcanoes typically appear as a *bright-dark* pair within a circular planimetric outline. Near the center, there is often a visible summit pit that appears as a *dark-bright* pair since the radar energy backscatters strongly from the far-range rim. However, if the pit is too small relative to the image resolution, it may not appear at all or may appear just as a bright spot. A high-level illustration of the imaging process is given in Figure 3.

These topography-induced features are the primary visual cues that geologists report using to locate volcanoes. However, there are a number of other, more subtle cues. The apparent brightness of an area in a radar image depends not only on the macroscopic topography but also on the surface roughness relative to the radar wavelength. Thus, if the flanks of a volcano have different roughness properties than the surrounding plains, the volcano may appear as a bright or dark circular area instead of as a bright-dark pair. Volcanoes may also appear as radial flow patterns, texture differences, or as disruptions of graben. (Graben are ridges or grooves in the planet surface, which appear as bright lines in the radar imagery–see Figure 2.)

2.2. Obtaining a Labeled Training Database

Although the Magellan imagery of Venus is the highest resolution available, expert geologists are unable to determine with 100% certainty whether a particular image feature is indeed a volcano. This ambiguity is due to a variety of factors such as image resolution,

signal-to-noise level, and difficulties associated with interpreting SAR data. For the same image, different geologists will produce different labelings, and even the same geologist may produce different labelings at different times.

To help quantify this uncertainty, the geologists are asked to assign the training examples to subjective probability "categories." Based on extensive discussions with the geologists, five categories are used.

Category 1 almost certainly a volcano ($p \approx 0.98$); the image clearly shows a summit pit, a bright-dark pair, and a circular planimetric outline.

Category 2 probably a volcano ($p \approx 0.80$); the image shows only two of the three category 1 characteristics.

Category 3 possibly a volcano ($p \approx 0.60$); the image shows evidence of bright-dark flanks or a circular outline; summit pit may or may not be visible.

Categroy 4 a pit ($p \approx 0.50$); the image shows a visible pit but does not provide conclusive evidence for flanks or a circular outline.

Category 5 not a volcano ($p \approx 0.0$).

The probability p attached to category i is interpreted as follows. Given that a geologist has assigned an image feature to category i, the probability that the feature is truly a volcano is approximately p_i. Figure 4 shows some typical volcanoes from each category. The use of quantized probability bins to attach levels of certainty to subjective image labels is not new. The same approach is used routinely in the evaluation of radiographic image displays to generate subjective ROC (receiver operating characteristic) curves (Bunch, 1978; Chesters, 1992).

A simple experiment was conducted to assess the variability in labeling between two planetary geologists, who will be referred to as A and B. Both of these geologists were members of the Volcanism Working Group of the Magellan science team and have extensive experience in studying Earth-based and planetary volcanism. They have published some of the standard reference works on Venus volcanism (Guest et al., 1992; Aubele and Slyuta, 1990; Crumpler et al., 1997). Each geologist separately labeled a set of four images known as `HOM4`. The labels were then compared using a simple spatial thresholding step to determine the correspondence between label events from the two geologists. (A "label event" simply refers to a labeler circling an image feature and assigning a subjective confidence label.) The resulting confusion matrix is given in Table 1.

The (i, j)th element of the confusion matrix counts the number of times that labeler A assigned a visual feature to category i while labeler B assigned the same feature to category j. Two label events are considered to belong to the same visual feature, if they are within a few pixels of each other. The $(i, 5)$ entries count the instances where labeler A provided label i, but labeler B did not provide any label (and vice versa for the $(5, j)$ entries). Entry (5,5) is not well-defined.

If both labelers agreed completely (same location and label for all events), the confusion matrix would have only diagonal entries. In the case shown in Table 1, there is clearly substantial disagreement, as evidenced by the off-diagonal elements in the matrix. For example, label 3 is particularly noisy in both "directions." Label 3 is actually noisier than

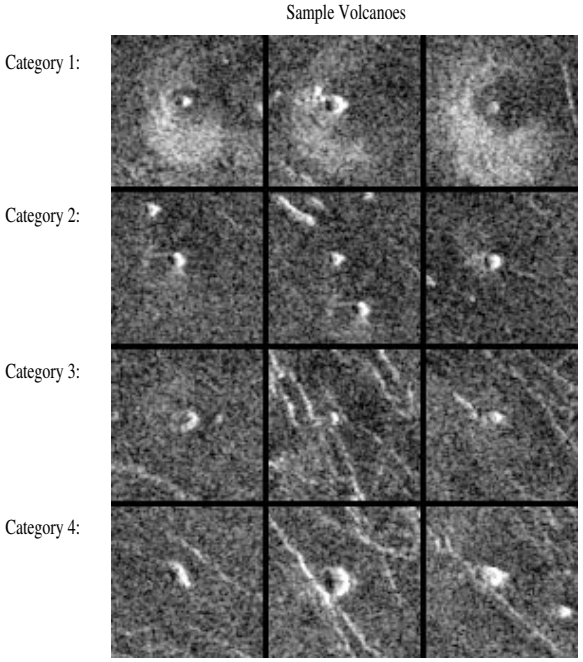


Figure 4. A selection of volcanoes from each of the confidence categories.

Table 1. Confusion matrix of geologist A vs. geologist B on HOM4.

	geologist B				
	Label 1	Label 2	Label 3	Label 4	Label 5
geologist A					
Label 1	19	9	13	1	4
Label 2	8	8	12	4	8
Label 3	4	6	18	5	29
Label 4	1	5	1	24	16
Label 5	3	5	37	15	X

label 4 because there is greater variability in the appearance of category 3 compared to category 4 (4's are simple pits, while 3's are less well-defined). About 50% of the objects assigned label 3 by either labeler are not detected at all by the other labeler. On the other hand, only about 30% of the objects assigned label 4 and 10% of the objects assigned label 1 by one labeler are missed by the other.

The confusion matrix clearly illustrates that there is considerable ambiguity in small volcano identification, even among experts. Success for the task can only be measured in a relative manner. To evaluate performance, we treat one set of labels as ground truth and measure how well an algorithmic detector agrees with this set of reference labels. In this paper, reference labels 1–4 are all considered to be true volcanoes for the purpose of performance evaluation. An alternative "weighted" performance metric is discussed in (Burl et al., 1994b). We also measure how well human labelers agree with the reference labels. Ideally, an algorithm should provide the same consistency with the reference labels as the human experts. A consensus labeling generated by a group of geologists working together and discussing the merits of each image feature is often used as the reference, since in general this labeling will be a more faithful representation of the actual ground truth. A typical consensus labeling is shown in Figure 5. From the geologists' point of view, it is a useful achievement to detect most of the category 1's and 2's, as the category 3's and 4's would not be used as part of a conservative scientific analysis.

2.3. Focus of Attention (FOA)

The first component in the JARtool system is a focus of attention (FOA) algorithm that is designed to take as input an image and produce as output a discrete list of *candidate* volcano locations. In principle, every pixel in the image could be treated as a candidate location; however, this is too expensive computationally. A better approach is to use the FOA to quickly exclude areas that are void of any volcanoes. Only local regions passing the FOA are given to subsequent (computationally more expensive) processes. Hence, the FOA should operate in an aggressive, low-miss-rate regime because any missed volcanoes at this stage will be lost for good. The rate of false positives (false alarms) from the FOA is not critical; these may still be rejected by later stages (classification).

Given the constraints of the FOA (low miss rate and low computational cost), a reasonable approach is to use a matched filter, i.e., a linear filter that matches the signal one is trying to find. The matched filter is optimal for detecting a known signal in white (uncorrelated) Gaussian noise (Duda and Hart, 1973). Of course, the volcano problem does not quite satisfy these underlying assumptions. Specifically, the set of observed volcanoes shows structured variations due to size, type of volcano, etc., rather than "isotropic" variations implicit with a signal plus white noise model. Likewise, the clutter background cannot be properly modeled as white noise. Despite these caveats, we have found empirically that the following modified matched filtering procedure provides a reasonable focus of attention mechanism.

Let $\mathbf{v}_i$ denote a $k \times k$ pixel region around the i-th training volcano. There is some loss of information due to this windowing process (especially for larger volcanoes). However, in our experiments, the results have not been particularly sensitive to the value of k (default = 15 spoiled pixels) (Burl et al., 1996). This may indicate that most of the information is concentrated at the center of the volcano (for example, the transition in shading and presence of a summit pit) or that the matched filter is not able to exploit the information from the periphery–both explanations probably have merit.

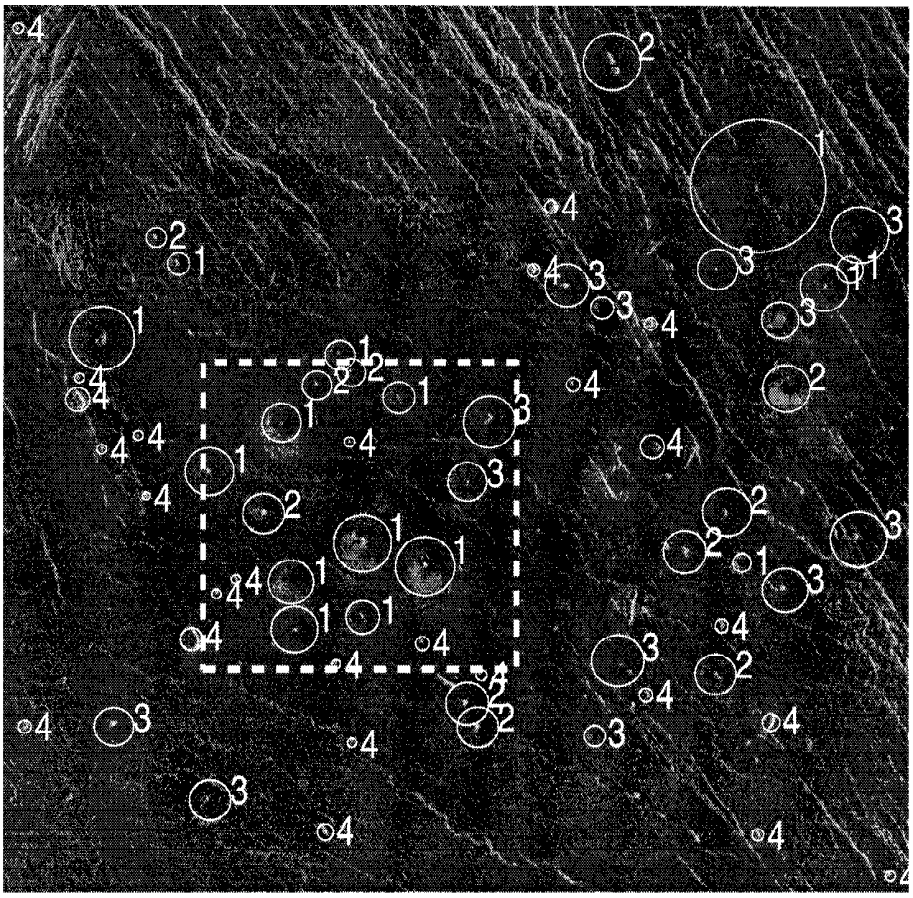


Figure 5. Consensus labeling of a Magellan SAR image of Venus. The labeling shows the size, location, and subjective uncertainty of each image feature. The dashed box corresponds to the subimage shown in Figure 2.

Each $k \times k$ region can be normalized with respect to the local average image brightness (DC level) and contrast as follows:

$$\tilde{\mathbf{v}}_i = \frac{\mathbf{v}_i - \mu_i \cdot \mathbf{1}}{\sigma_i} \tag{1}$$

where μ_i is the mean of the pixels in $\mathbf{v}_i$, σ_i is their standard deviation, and $\mathbf{1}$ is a $k \times k$ matrix of ones. This normalization is essential since there are large fluctuations in the DC and contrast between images and even between local areas of the same image. A modified

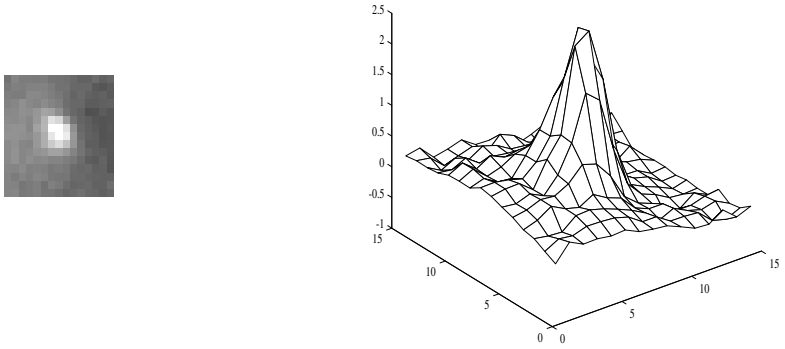


Figure 6. The matched filter displayed as a template (left) and as a surface plot (right). The matched filter captures many of the characteristics that planetary geologists report using when manually locating volcanoes. In particular, there is a bright central spot corresponding to the volcanic summit pit and left-to-right bright-dark shading.

matched filter $\mathbf{f}$ is constructed by averaging the normalized volcano examples from the training set. Figure 6 shows the resulting filter.

Applying the matched filter to an image involves computing the normalized cross-correlation of $\mathbf{f}$ with each $k \times k$ image patch. The cross-correlation can be computed efficiently using separable kernel methods to approximate the 2-D kernel $\mathbf{f}$ as a sum of 1-D outer products (Treitel and Shanks, 1971). High response values indicate that there is strong correlation between the filter and the image patch. Candidate volcano locations are determined by thresholding the response values and spatially aggregating any threshold crossings that are within a prescribed distance from each other (default distance = 4 pixels).

Obviously one concern with such a simple FOA is that if the population of volcanoes contains significant subclasses then a single filter would not be expected to perform well. However, experiments with an alternative mechanism that uses clustering to find several matched filters has provided only limited improvement (Stough and Brodley, 1997).

2.4. Feature Extraction

A region of interest (ROI) identified by the focus of attention algorithm can be viewed as a point in a k^2-dimensional space by stringing the $k \times k$ pixel values out into a long vector. Note, however, that there is a loss of spatial neighborhood information. Algorithms that treat the data in this form will not *explicitly* know that certain pixels were adjacent in the original image data. Also, given the small number of training examples relative to the dimensionality of the data, there is little hope of learning anything useful without additional constraints.

Experimental results with a variety of feedforward neural network classification models verified this hypothesis (Baldi, 1994). The training data were often linearly separable in pixel space resulting in an underconstrained training procedure that allowed the model to essentially memorize the training data perfectly, but with poor generalization to unseen data. Thus, direct use of the pixels as input to a classification algorithm is not practical.

To work around the small number of training examples, we make use of the fact that for visual data, there is additional prior information that helps constrain the problem. Specifically, there is reason to believe that the volcanoes "live" on a low-dimensional manifold embedded in k^2-dimensional space. Although the manifold is certainly nonlinear, we make use of the principal components analysis (PCA) paradigm to approximate the manifold with a low-dimensional hyperplane. This approximation can be viewed as a mapping from the high-dimensional pixel space to a lower dimensional feature space in which the features consist of linear combinations of the pixel values. We have also experimented with clustering the training data in pixel space and applying PCA separately to each cluster. This extension yields an approximation to the manifold by a union of hyperplanes. (See Section 2.7 for additional discussion.)

Before presenting a more detailed view of the PCA approach, we remark that PCA is not the only method available for linear feature extraction. The assumption behind PCA is that it is important to find features that represent the data. Other approaches, such as linear discriminant analysis (LDA), seek to find *discriminative* features that separate the classes. In the context of finding volcanoes, however, the "other" class is quite complex consisting of all patterns that are not volcanoes. Direct application of LDA in pixel space leads to poor results.

Recently, a method was proposed that combines PCA and LDA to find "most discriminative features" (Swets and Weng, 1996). In this approach, PCA is used on the pooled set of examples (volcanoes and non-volcanoes) to project the pixel data to a lower dimensional feature space. LDA methods are then applied in the projected space. Effectively this amounts to using a "linear machine" classifier (Duda and Hart, 1973) in the space of principal components features. In Section 3 we demonstrate that by performing PCA on only the positive examples and allowing more complex classifiers in PCA space, the JARtool algorithm is able to outperform the method of Swets and Weng by a significant margin.

PCA can be summarized as follows. The goal is to find a q-dimensional subspace such that the projected data is closest in L_2 norm (mean square error) to the original data. This subspace is spanned by the eigenvectors of the data covariance matrix having the highest corresponding eigenvalues. Often the full covariance matrix cannot be reliably estimated from the number of examples available, but the approximate highest eigenvalue basis vectors can be be computed using singular value decomposition (SVD).

Each normalized training volcano is reshaped into a vector and placed as a column in an $n \times m$ matrix X, where n is the number of pixels in an ROI ($n = k^2$) and m is the number of volcano examples. The SVD produces a factorization of X as follows:

$$X = USV^T \tag{2}$$

For notational convenience, we will assume m is less than n. Then in Equation 2, U is an $n \times m$ matrix such that $U^TU = I_{m\times m}$, S is $m \times m$ and diagonal with the elements on the diagonal (the singular values) in descending order, and V is $m \times m$ with $V^TV = VV^T =$

$I_{m\times m}$. Notice that any column of X (equivalently, any ROI) can be written exactly as a linear combination of the columns of U. Furthermore, if the singular values decay quickly enough, then the columns of X can be closely approximated using linear combinations of only the first few columns of U. That is, the first few columns of U serve as an approximate basis for the entire set of examples in X.

Thus, the best q-dimensional subspace on which to project is spanned by the first q columns of U. An ROI is projected into this q-dimensional feature space as follows:

$$\mathbf{y} = \left[\begin{array}{cccc} \mathbf{u}_1 & \mathbf{u}_2 & \ldots & \mathbf{u}_q \end{array} \right]^T \mathbf{x} \tag{3}$$

where $\mathbf{x}$ is the ROI reshaped as an n-dimensional vector of pixels, $\mathbf{u}_i$ is the i-th column of U, and $\mathbf{y}$ is the q-dimensional vector of measured features.

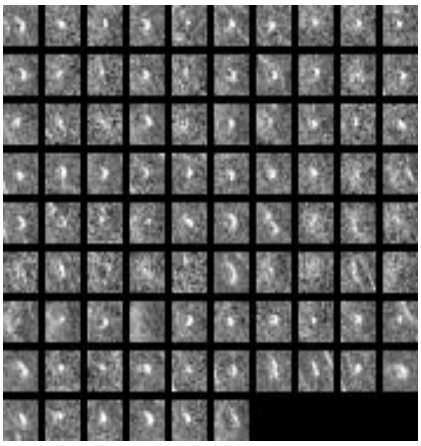
Figure 7b shows the columns of U reshaped as ROIs. The templates are ordered according to singular value so that the upper left template corresponds to the maximum singular value. Notice that the first ten templates (top row) exhibit structure while the remainder appear very random. This suggests using a subspace of dimension ≤ 10. The singular value decay shown in Figure 7c also indicates that 6 to 10 features will be adequate to encode most of the information in the examples. Indeed, parameter sensitivity experiments, which are reported in (Burl et al., 1996) show that values of q in the range 4–15 yield similar overall performance.

2.5. Classification

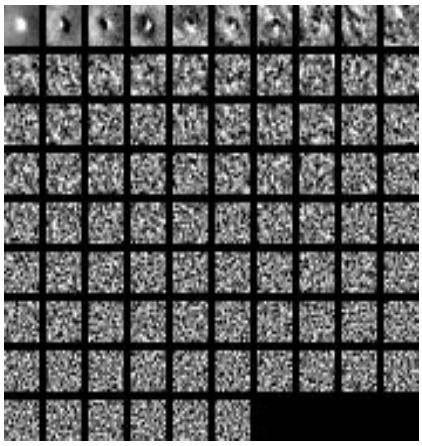
The FOA and feature extraction steps transform the original Magellan images into a discrete set of feature vectors that can be classified with "off-the-shelf" learning algorithms. The remaining step is to classify ROIs into volcano or non-volcano. FOA and feature learning are based exclusively on positive examples (volcanoes). The classifier could also be trained in this manner. However, there are arguments (Fukunaga, 1990) showing that single-class classifiers are subject to considerable error even in relatively low dimensions because the location of the "other" distribution is unknown. Experiments based on non-parametric density estimation of the volcano class verified this hypothesis: the method gave poorer performance than the two-class methods described below.

The negative examples were not used in the FOA and feature learning steps due to the complexity of the non-volcano class. Nonetheless these steps provide substantial conditioning of the data. For example, the FOA centers objects within a $k \times k$ window. The feature extraction step uses prior knowledge about visual data (i.e., the fact that certain object classes can be modeled by linear combinations of basis functions) to map the data to a lower-dimensional space in which there is an improved opportunity for learning a model that generalizes to unseen data. Hence, in the PCA space it is reasonable to use supervised two-class learning techniques. We have experimented with a variety of algorithms including quadratic (or Gaussian) classifiers, decision trees, linear discriminant analysis, nearest neighbors using Euclidean and spatially weighted distance measures (Turmon, 1996), tangent distance (Simard, le Cun, and Denker, 1993), kernel density estimation, Gaussian mixture models, and feedforward neural networks (Asker and Maclin, 1997b; Cherkauer, 1996).

(a) Training Volcanoes



(b) Principal Components



(c) Singular Values

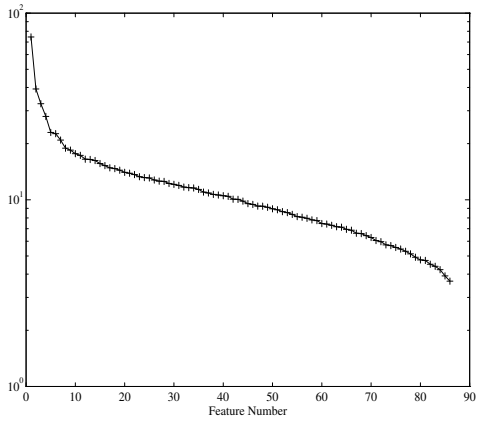


Figure 7. (a) The collection of volcanoes used for feature synthesis. (b) The principal components derived from the examples. (c) The singular values indicate the importance of each of the features for representing the examples.

All of these methods (with the exception of linear discriminant analysis) yielded similar performance on an initial test set of images. We interpret this to mean that the critical system design choices were already made, specifically in the feature learning stage; the choice of classifier is of secondary importance. In the experiments reported in Section 3, the quadratic classifier is used as the default since it is optimal for Gaussian data and provides posterior

probability estimates, which can be thresholded to vary the trade-off between detection and false alarm rate. Letting ω_1 designate the volcano class and ω_2 the non-volcano class, we have the following from Bayes' rule:

$$p(\omega_i|\mathbf{y}) = \frac{p(\mathbf{y}|\omega_i) \cdot p(\omega_i)}{p(\mathbf{y}|\omega_1) \cdot p(\omega_1) + p(\mathbf{y}|\omega_2) \cdot p(\omega_2)} \tag{4}$$

where $\mathbf{y}$ is the observed feature vector. For the quadratic classifier, the posterior probabilities are estimated by assuming the class-conditional densities are Gaussian

$$p(\mathbf{y}|\omega_i) = \mathcal{N}(\mathbf{y};\ \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{5}$$

where the statistics of each class ($\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$) are estimated from labeled training data.

2.6. Summary of the Training Procedure

In summary, training consists of a three-step process based on the geologist-labeled images:

1. Construct the FOA detection filter from the volcanoes labeled in the training images. Apply the FOA to the training images and then use the "ground truth" labels to mark each candidate ROI as a volcano or non-volcano.
2. Determine principal component directions from the ROIs that were detected in step 1 and marked as volcanoes.
3. Estimate the parameters of a classifier from the labeled feature vectors obtained by projecting all of the training data detected in step one onto the PCA templates of step two. ROIs marked as true volcanoes in step one serve as the positive examples, while ROIs marked as non-volcanoes serve as the negative examples.

Comment: This training procedure contains some non-idealities. For example, the positive examples supplied to the classifier are the same examples used to derive the features (principal component directions). It would clearly be better if the classifier were to receive a disjoint set of positive training examples, but given the limited number of examples, we compromised on the procedure described above.

2.7. Extension to the Basic Algorithm

One objection to the baseline approach presented thus far is that there are various subtypes of volcanoes, each with unique visual characteristics. One would not expect the (approximate) hyperplane assumption implicit in the PCA approach to hold across different volcano subtypes. This limitation could affect the algorithm's ability to generalize across different regions of the planet, and in fact in the experiments reported later (Section 3), we have observed that the baseline system performs significantly worse on heterogeneous sets of images selected from various areas of the planet.

One solution we investigated involves using a combination of classifiers in which each classifier is trained to detect a different volcano subclass. The outputs from all the classifiers

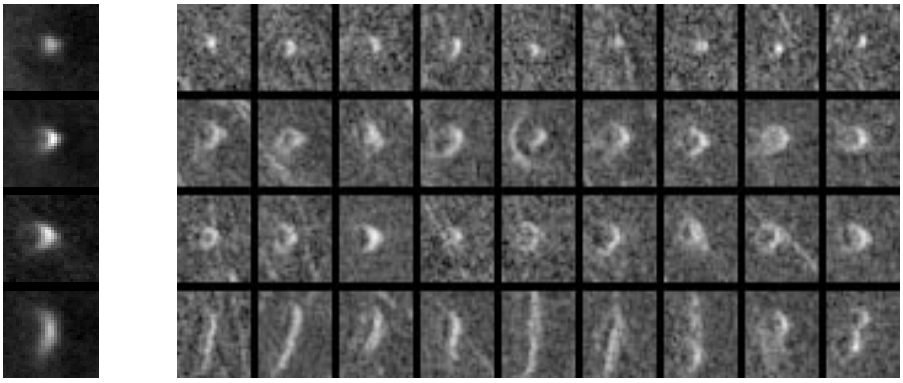


Figure 8. Example volcanoes from four different clusters and their respective cluster centers. Each row represents a sample of volcanoes that have been clustered together using K-means.

are then combined to produce a final classification. Subclasses of volcanoes are found automatically by clustering the raw pixel representation of the volcanoes in the training set using k-means (Duda and Hart, 1973). In Figure 8 we show the results of clustering the volcanoes into four classes. Each row corresponds to a cluster; the first column shows the cluster center, while the other columns show selected instances. For each cluster, principal components analysis is performed separately yielding a set of features (basis functions) specific to a subclass of volcanoes. A classifier is then trained for each subclass, and in the final step the predictions of all the classifiers are combined into one. Details of the method for combining classifiers are given in (Asker and Maclin, 1997a). Experimental results comparing the combined classifier approach with the baseline are given in Section 3.

3. Performance Evaluation

Initial experiments were conducted using a small set of images called `HOM4` (denoting a set of *four* images which were relatively *homogeneous* in appearance). The results from these experiments were used to provide feedback in the algorithm development process and also served to fix the values of *miscellaneous* parameters such as the ROI window size, FOA threshold, number of principal components, and so forth. Because of this feedback, however, performance on `HOM4` cannot be considered as a fair test of the system (since in effect one is training on the test data). In addition, `HOM4` did not include enough images to provide a thorough characterization of the system performance and generalization ability.

After these initial experiments, the algorithm and all the *miscellaneous* parameters were frozen at specific values (listed in the Appendix). Based on empirical sensitivity studies (Burl et al., 1996), we believe the system is relatively insensitive to the exact values of these parameters. Note that "freezing" does not apply to parameters normally derived during

Table 2. Experiments and image sets used to evaluate system performance.

Experiment	Image Set	#Volcanoes	Description
Initial Testing	HOM4	160	4 images from lat 30°N, 332°
Extended Testing	HOM38	480	38 images from lat 30°N, 332°
	HOM56	230	56 images from lat 30°N, 123°
	HET36	670	36 images from various locations
Follow-up	HET5	131	5 images from various locations

learning such as the matched filter, principal components, or statistics used by the classifier. These are recalculated for each experiment from the stated set of training examples.

Extensive tests were conducted on three large image sets (HOM38, HOM56, HET36). The naming convention for the image sets is to use HOM if the set is considered homogeneous (images from the same local region) and HET if the set is heterogeneous (images selected from various locations). The numerical suffix indicates the number of images in the data set. Note that the smallest of these datasets covers an area of 450km $\times$ 450km.

A summary of the experiments and image sets is given in Table 2. The number of volcanoes listed corresponds to the number of label events in the "ground-truth" reference list, i.e., each label event is counted as a volcano regardless of the assigned confidence. The main conclusion from these tests was that the baseline system performed well on homogeneous sets in which all images were taken from the same region of the planet, but performed poorly on heterogeneous sets in which images were selected randomly from various locations on the planet.

To better understand this difference in performance, we conducted a follow-up experiment using a small set of heterogeneous images HET5. Our initial hypothesis was that the discrepancy occurred because the volcanoes from different regions looked different. However, what we found was that "knowing the local volcanoes" was not nearly as important as knowing the local non-volcanoes. The argument used to arrive at this conclusion is somewhat subtle, but is explained in detail in Section 3.4.

3.1. ROC and FROC

As explained in Section 2.2, we evaluate performance by measuring how well a detector (algorithmic or human) agrees with a set of reference labels. A "detection" occurs if the algorithm/human indicates the presence of an object at a location where a volcano exists according to the reference list. Similarly, a "false alarm" occurs if the algorithm/human indicates the presence of an object at a location where a volcano does not exist according to the reference list. Consider a system which produces a scalar quantity indicating detection confidence (e.g., the estimated posterior class probability). By comparing this scalar to a fixed threshold, one can estimate the number of detections and false alarms for that particular threshold. By *varying* the threshold one can estimate a sequence of detection/false-alarm points. The resulting curve is known as the receiver operating characteristic (ROC) curve (Green and Swets, 1966; MacMillan and Creelman, 1991; Spackman, 1989; Provost and Fawcett, 1997).

The usual ROC curve plots the probability of detection versus the probability of false alarm. The probability of detection can be estimated by dividing the number of detections by the number of objects in the reference list. Estimating the probability of false alarm, however, is problematic since the number of possible false alarms in an image is not well-defined. A practical alternative is to use a "free-response" ROC (FROC) (Chakraborty and Winter, 1990), which shows the probability of detection versus the *number* of false alarms (often normalized per image or per unit area). The FROC methodology is used in all of experiments reported in this paper; in particular, the x-axis corresponds to the number of false alarms per square kilometer.

The FROC shares many of the properties of the standard ROC[2]. For example, the best possible performance is in the upper left corner of the plot so an FROC curve that is everywhere above and to the left of another has better performance. The FROC curve is implicitly parameterized by the decision threshold, but in practice the geologist would fix this threshold thereby choosing a particular operating point on the curve.

3.2. Initial Experiments

Experiments on HOM4 were performed using a generalized form of cross-validation in which three images were used for training and the remaining image was reserved for testing; the process was repeated four times so that each image served once as the test image. This type of training-testing procedure is common in image analysis problems (Kubat, Holte, and Matwin, 1998).

The system output was scored relative to the consensus labeling with all subjective confidence categories treated as true volcanoes. The FROC performance curve is shown in Figure 9a. The horizontal dashed line across the top of the figure (labeled FOA=0.35) shows the best possible system performance using an FOA threshold of 0.35. (The line is not at 100% because the FOA misses some of the true volcanoes.)

The performance points of two individual geologists are also shown in the figure. Geologist A is shown with the '*' symbol, while geologist B is shown with the '+'. Note that for these images the system performance (at an appropriately chosen operating point) is quite close to that of the individual geologists. The effect of using different operating points is shown in table form in Figure 10a.

3.3. Extended Performance Evaluation

3.3.1. Homogeneous Images Given the encouraging results on HOM4, we proceeded to test the system on larger images sets. The HOM4 images were part of a 7×8 block of images comprising a full-resolution Magellan "data product." Within this block 14 images were blank due to a gap in the Magellan data acquisition process. The remaining 38 (56 minus 4 minus 14) images were designated as image set HOM38. Training and testing were performed using generalized cross-validation in which the set of images was partitioned into six groups or "folds." Two of the images did not contain any positive examples, so these were used only for training. The other 36 images were partitioned randomly into six groups of six with the constraint that each group should have approximately the same number of positive examples. Five folds were used for training and the remaining fold was

used for testing; the process was repeated so that each fold served once as the test set. This *leave-out-fold* method was used rather than *leave-out-image* to reduce the run time.

The FROC performance is shown in Figure 9b (solid line). Since we did not have consensus labeling available for the entire image set, the labels of geologist A were used as the reference. The '+' symbol shows the performance of geologist B (relative to A), while the 'o' symbol shows the performance of one of the non-geologist authors (Burl). The performance of the algorithm is similar to the HOM4 case except at higher false alarm rates where the HOM38 performance is lower by approximately 12%. The discrepancy is probably due to differences in the FOA performance. Note that for HOM4 the FOA asymptote is around 94%, while it is only at 83% for HOM38.

For comparison FROC curves are plotted for two other methods. The dashed curve labeled "FOA" shows the performance that could be achieved by using only a matched filter but with a lower threshold. The combination of matched filter and classification yields better performance than the matched filter alone. (Matched filtering was proposed as a possible solution to the volcano-detection problem in (Wiles and Forshaw, 1993)). Also shown is the FROC for the discriminative Karhunen-Loeve (DKL) approach (Swets and Weng, 1996), which combines principal components and linear discriminant analysis. Observe that the JARtool approach provides significantly better performance (an increase in detection rate by 10 percentage points or more for a given false alarm rate).

For the HOM38 experiments, the training images and test images were geographically close together. To test the system's generalization ability, another experiment was performed in which training was carried out on HOM4+HOM38 and testing was carried out on a geographically distinct set of homogeneous images HOM56. The HOM56 images were from the same latitude as the training images and visually appeared to have similar terrain and volcano types. For this data set, reference labels were provided by one of the non-geologist authors (Burl).

The baseline performance is shown as a solid curve in Figure 9c. The clustering extension to the baseline algorithm was also applied to the data. The corresponding FROC is shown with the dashed curve. The clustering approach appears to provide a slight improvement over the baseline algorithm, consistent with other results reported in (Asker and Maclin, 1997a). However, the baseline algorithm is still used in the fielded system because of its simplicity and shorter training time. (These factors are believed to outweigh the marginal loss in performance.)

3.3.2. Heterogeneous Images Finally, the system was evaluated under the most difficult conditions. A set of 36 images (HET36) was selected from random locations on the planet. These images contained significantly greater variety in appearance, noisiness, and scale than the previous image sets. Training was done on HOM4+HOM38. The system FROC performance (relative to consensus) is shown in Figure 9d, and the performance at selected operating points is shown in Figure 10d. Here the classifier performs much worse than on the more homogeneous data sets. For example at 0.001 false alarms/km^2 the detection performance is in the 35-40% range whereas for all homogeneous image sets, the detection rates were consistently in the 50-65% range. For the few images where we also have individual labels, the geologists' detection performance is roughly the same as it was on

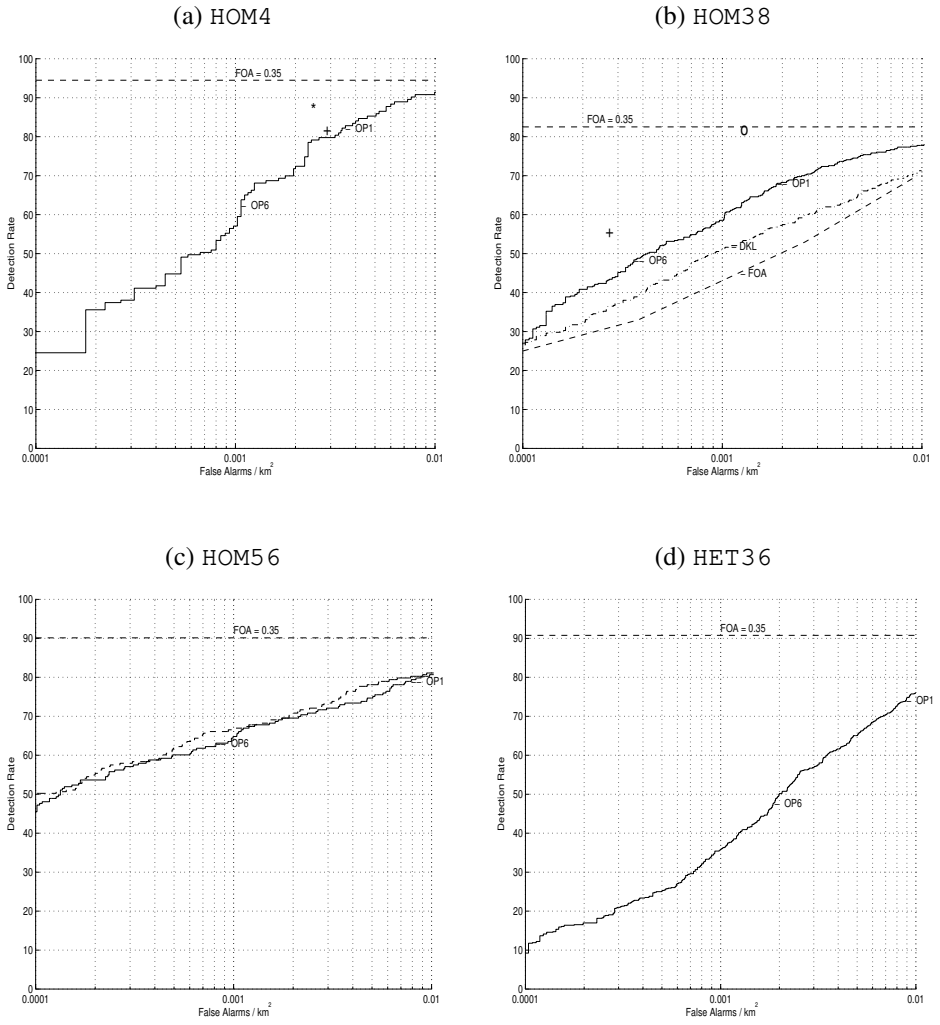


Figure 9. FROC curves showing the performance of the baseline algorithm on four image sets. Each figure shows the trade-off between detection rate and the number of false alarms per area. Note that the algorithm performs considerably better on the homogeneous image sets (a,b,c) than on the heterogeneous set (d).The discrete symbols (*,+,O) in (a) and (b) show the performance of human labelers.

the homogeneous images. From these results it appears that human labelers are much more robust with respect to image inhomogeneity.

(a) HOM4

Operating point:	OP1	OP2	OP3	OP4	OP5	OP6
Threshold:	0.75	0.80	0.85	0.90	0.95	0.99
Detected Category 1 (%)	88.9	88.9	86.1	80.6	72.2	63.9
Detected Category 2 (%)	89.7	89.7	86.2	79.3	72.4	65.5
Detected Volcanoes (%)	82.2	81.0	79.8	74.9	69.3	62.6
False Alarms per image	19.5	18.5	15.0	13.0	9.5	6.0
False Alarms per 10^4 km^2	34.7	32.9	26.7	23.1	16.9	10.7

(b) HOM38

Operating point:	OP1	OP2	OP3	OP4	OP5	OP6
Threshold:	0.75	0.80	0.85	0.90	0.95	0.99
Detected Category 1 (%)	92.0	92.0	88.0	84.0	84.0	76.0
Detected Category 2 (%)	80.8	78.2	75.6	71.8	65.4	50.0
Detected Volcanoes (%)	68.0	65.0	63.5	60.3	54.6	48.4
False Alarms per image	10.6	8.8	7.3	5.8	3.9	2.1
False Alarms per 10^4 km^2	18.8	15.6	13.0	10.3	6.9	3.7

(c) HOM56

Operating point:	OP1	OP2	OP3	OP4	OP5	OP6
Threshold:	0.75	0.80	0.85	0.90	0.95	0.99
Detected Category 1 (%)	100.0	100.0	91.7	91.7	91.7	91.7
Detected Category 2 (%)	84.2	84.2	81.6	79.0	79.0	60.5
Detected Volcanoes (%)	79.0	77.7	75.1	73.4	70.4	63.1
False Alarms per image	42.8	35.5	28.7	21.4	13.2	5.0
False Alarms per 10^4 km^2	76.1	63.1	51.0	38.0	23.5	8.9

(d) HET38

Operating point:	OP1	OP2	OP3	OP4	OP5	OP6
Threshold:	0.75	0.80	0.85	0.90	0.95	0.99
Detected Category 1 (%)	90.3	87.1	87.1	83.9	79.0	64.5
Detected Category 2 (%)	84.6	82.4	80.2	78.7	75.0	62.5
Detected Volcanoes (%)	74.1	72.0	69.8	66.2	60.9	47.8
False Alarms per image	50.2	43.7	37.1	29.7	20.5	10.6
False Alarms per 10^4 km^2	89.2	77.6	65.9	52.7	36.4	18.8

Figure 10. Performance of the baseline system at various operating points along the FROC curve.

3.4. Follow-Up Analysis

To better understand the decreased performance on heterogeneous images, we conducted follow-up experiments on a smaller set of images (HET5). Performance on this set was also poor, and our initial hypothesis was that the degradation occurred because the volcanoes were somehow different from image to image. To investigate this possibility, we performed experiments with two different training paradigms: (1) cross-validation in which one *image* was left out of the training set and (2) cross-validation in which one *example* was left out of the training set. The first method will be referred to as LOI for "leave-out-image"; the second method will be referred to as LOX for "leave-out-example."

Two nearest-neighbor classification algorithms were evaluated in addition to the baseline Gaussian classifier. The nearest-neighbor algorithms were applied directly to the pixel-space regions of interest (ROIs) identified by the FOA algorithm. To allow for some jitter in the alignment between ROIs, we used the peak cross-correlation value over a small spatial window as the similarity measure. One nearest-neighbor algorithm was the standard two-class type in which an unknown test example is assigned to the same class as its nearest neighbor in the reference library. The other was a one-class version in which the reference library contains only positive examples (volcanoes); an unknown example is assigned to the volcano class if it is similar enough to some member of the reference library.

Performance was evaluated on both the HOM4 and HET5 datasets using the three classifiers (baseline, 1-class nearest neighbor, and 2-class nearest neighbor) and two training paradigms (LOI and LOX). The results are shown in Figure 11. For computational reasons, the baseline method was trained and tested on the same data (TTS) rather than leaving out an example. The effect of including the test example in the training set is minimal in this case since one example has little effect on the class-conditional mean and covariance estimates.

The following is the key observation: on HET5 the baseline and 2-class nearest-neighbor algorithms work significantly better under the LOX training paradigm than under the LOI training paradigm; however, the 1-class algorithm works the same under both training paradigms. If having knowledge about the local volcanoes were the critical factor, the 1-class algorithm should have worked significantly better under LOX than under LOI. Instead we conclude that access to the local non-volcanoes is the critical factor. The 1-class algorithm *completely* ignores the non-volcanoes and hence does not show any difference between LOI and LOX. The other methods do use the non-volcanoes, and these show a dramatic improvement under LOX.

On HOM4 there is little difference between the LOI and LOX results. Since these images are from the same area of the planet, the appearance of the non-volcanoes is similar from image to image. Thus, leaving out one example or leaving out one image from the training set does not have much effect. The non-volcanoes in HET5 and other heterogeneous image sets vary considerably from image to image and this may be the source of the degradation in performance (the training data is inadequate for learning the non-volcano distribution).

4. Project Status

Participating in the development of the JARtool system were two planetary geologists (Aubele and Crumpler) who were members of the Volcanism Working Group of the Mag-

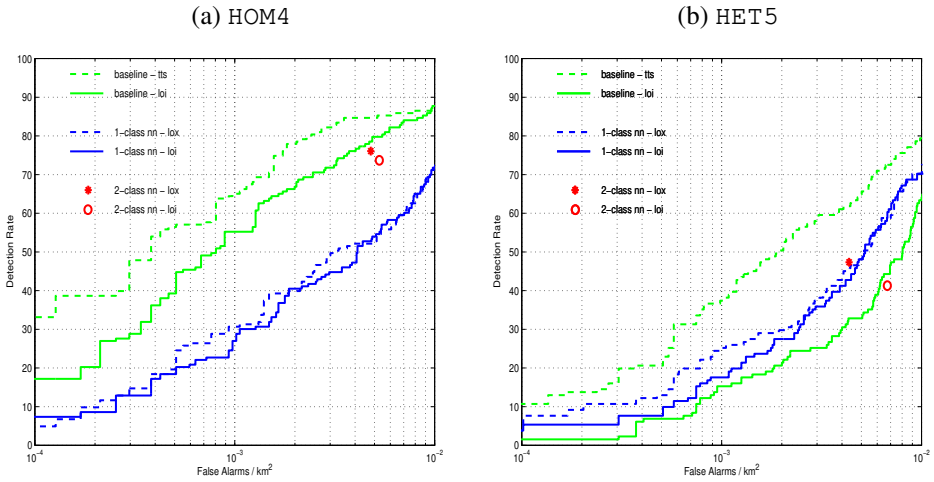


Figure 11. Performance results of svd-gauss, 1-class nearest neighbor, and 2-class nearest-neighbor algorithms under the leave-out-image (LOI) and leave-out-example (LOX) training paradigms. (a) Results on a set of four homogeneous images from one area of the planet. (b) Results on a set of five heterogeneous images selected from different areas of the planet. Refer to the text for an interpretation of the results.

ellan Science team and principal investigators in NASA's Planetary Mapping and Venus Data Analysis Programs.

The geologists have been evaluating the JARtool approach both in terms of the scientific content provided by the analyzed images and as a tool to aid in further cataloging. From the planetary geologists' point of view, the primary goal was to achieve annotation of 80% or more of the small volcanoes in the analyzed datasets. A secondary goal was to obtain accurate diameter estimates for each volcano. Locating different morphologic types of small volcanoes was also of interest. However, it was recognized up front that some of the types would be easy to detect and some would be difficult (both for human experts and for algorithms). To the geologists, the system should be considered a success if it detects a high percentage of the "easy" volcanoes (category 1 and 2). Our test results indicate that this level of performance is achieved on homogeneous image sets. However, we have not succeeded in developing a reliable method for measuring volcano diameters. Hence, sizing capability is not included in the fielded system.

Our experiments and those of the scientists have indicated that the choice of operating point will vary across different areas of the planet, dependent on factors such as terrain type and local volcano distributions. Hence, the operating point is left "open" for the scientists to choose. Although the original intent was for the JARtool system to provide a fully-automated cataloging tool, it appears that the system will be most useful as an "intelligent assistant" that is used in an interactive manner by the geologists.

The capabilities of the system were recently expanded through integration of the *Postgres* database (Stonebraker and Kemnitz, 1991). A custom query tool supports arbitrary SQL

queries as well as a set of common "pre-canned" queries. JARtool is also being evaluated for use in other problem domains. Researchers or scientists who are interested in the software can direct inquiries to `jartool@aig.jpl.nasa.gov`.

5. Lessons Learned and Future Directions

Real-world applications of machine learning tend to expose practical issues which otherwise would go unnoticed. We share here a list of "lessons learned" which can be viewed as a "signpost" of potential dangers to machine learning practitioners. In addition, for each "lesson learned" we discuss briefly related *research* opportunities in machine learning and, thus, provide input on what topics in machine learning research are most likely to have practical impact for these types of large-scale projects in the future.

1. Training and testing classifiers is often only a very small fraction of the project effort. On this project, certainly less than 20%, perhaps as little as 10% effort was spent on classification aspects of the problem. This phenomenon has been documented before across a variety of applications (Langley and Simon, 1994; Brodley and Smyth, 1997). Yet, this directly contradicts the level of effort spent on classification algorithms in machine learning research, which has traditionally focused heavily on the development of classification algorithms. One implication is that classification technology is relatively mature and it is time for machine learning researchers to address the "larger picture." A difficulty with this scenario is that these "big picture" issues (some of which are discussed below) can be awkward to formalize.

2. A large fraction of the project effort (certainly at least 30%) was spent on "feature engineering," i.e., trying to find an effective representation space in which to apply the classification algorithms. This is a very common problem in applications involving sensor-level data, such as images and time-series. Unfortunately, there are few principled design guidelines for "feature engineering" leading to much trial-and-error in the design process. Commonly used approaches in machine learning and pattern recognition are linear projection techniques (such as PCA) and feature selection methods. Non-linear projection techniques can be useful but are typically computationally complex. A significant general problem is the branching factor in the search space for possible feature representations. There are numerous open problems and opportunities in the development of novel methods and systematic algorithms for feature extraction. In particular, there is a need for robust feature extraction techniques which can span non-standard data types, including mixed discrete and real-valued data, time series and sequence data, and spatial data.

3. Real-world classification systems tend to be composed of multiple components, each with its own parameters, making overall system optimization difficult if not impossible given finite training sets. For JARtool, there were parameters associated with FOA, feature extraction, and classification. Joint optimization of all of these parameters was practically impossible. As a result many parameters (such as the window size for focus of attention) were set based on univariate sensitivity tests (varying one parameter while keeping all others fixed at reasonable values). Closer coupling of machine learning

algorithms and optimization methods would appear to have significant potential payoffs in this context.

4. In many applications classification labels are often supplied by experts and may be much noisier than initially expected. At the start of the volcano project, we believed the geologists would simply tell us where all the volcanoes were in the training images. Once we framed the problem in an ROC context and realized that the resolution of the images and other factors led to inherent ambiguity in volcano identification, we began to understand the noisy, subjective nature of the labeling process. In fact, the geologists were also given cause to revise their opinions on the reliability of published catalogs. As real-world data sets continue to grow in size, one can anticipate that the fraction of data which is accurately labeled will shrink dramatically (this is already true for many large text, speech, and image databases). Research areas such as coupling unsupervised learning with supervised learning, cognitive models for subjective labeling, and active learning to optimally select which examples to label next, would appear to be ripe for large-scale application.

5. In applying learning algorithms to image analysis problems, spatial context is an important factor and can considerably complicate algorithm development and evaluation. For example, in testing our system we gradually realized that there were large-scale spatial effects on volcano appearance, and that training a model on one part of the planet could lead to poor performance elsewhere. Conversely, evaluating model performance on images which are spatially close can lead to over-optimistic estimates of system performance. These issues might seem trivial in a machine learning context where *independence* of training and test sets is a common mantra, yet the problem is subtle in a spatial context (How far away does one have to go spatially to get independent data?) and widely ignored in published evaluations of systems in the image analysis and computer vision communities. Thus, there is a need to generalize techniques such as cross-validation, bootstrap, and test-set evaluation, to data sources which exhibit dependencies (such as images and sequences).

A common theme emerging from the above "lessons" is that there is a need for a systems viewpoint towards large-scale learning applications. For example, in retrospect, it would have been extremely useful to have had an *integrated* software infrastructure to support data labeling and annotation, design and reporting of experiments, visualization, classification algorithm application, and database support for image retrieval. (For JARtool development, most of these functions were carried out within relatively independent software environments such as standalone C programs, Unix shell scripts, MATLAB, SAOimage, and so forth). Development of such an integrated infrastructure would have taken far more resources than were available for this project, yet it is very clear that such an integrated system to support application development would have enabled a much more rapid development of JARtool.

More generally, with the advent of "massive" data sets across a variety of disciplines, it behooves machine learning researchers to pay close attention to overall systems issues. How are the data stored, accessed, and annotated? Can one develop general-purpose techniques for extracting feature representations from "low-level" data? How can one best harness the

prior knowledge of domain experts? How can success be defined and quantified in a way which matches the user's needs?

6. Conclusion

The Magellan image database is a prime example of the type of dataset that motivates the application of machine learning techniques to real-world problems. The absence of labeled training data and predefined features imposes a significant challenge to "off the shelf" machine learning algorithms.

JARtool is a learning-based system which was developed to aid geologists in cataloging the estimated one million small volcanoes in this dataset. The system is trained for the specific volcano-detection task through examples provided by the geologists. Experimental results show that the system approaches human performance on homogeneous image sets but performs relatively poorly on heterogeneous sets in which images are selected randomly from different areas of the planet. The effect on system performance of a particular classification algorithm was found to be of secondary importance compared to the feature extraction problem.

Acknowledgments

The research described in this article has been carried out in part by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. Support was provided by the NASA Office of Advanced Concepts and Technology (OACT - Code CT), a JPL DDF award, NSF research initiation grant IRI 9211651, and a grant from the Swedish Foundation for International Cooperation in Research and Higher Education (Lars Asker).

We would like to thank Michael Turmon for his help and for performing some of the experiments. We would also like to thank Saleem Mukhtar, Maureen Burl, and Joe Roden for their help in developing the software and user-interfaces. The JARtool graphical user interface is built on top of the SAOTNG image analysis package developed at the Smithsonian Astrophysical Society (Mendel et al., 1997).

Appendix: Default Settings for Algorithm Parameters

In the experiments reported in this paper, all *miscellaneous* parameters were set to default values determined manually from experiments conducted on the HOM4 data set.

- In training, all volcanoes are treated equally, i.e., the categories 1–4 are not used to weight the training in any way.
- The window size for the FOA filter was 15×15 spoiled pixels. (Each spoiled pixel is an average of a 2×2 block of image pixels.)
- The threshold value for the detector was 0.35.

- The window size for the examples provided to the PCA procedure was 15×15 spoiled pixels.
- The threshold for the detection clustering algorithm was 4 pixels.
- The number of principal components (features) used for classification was 6.
- The classification method used was a maximum-likelihood Gaussian classifier, with independent full-covariance matrices for each class.
- Let $r_{0.5}$ be half the estimated radius (according to the reference list) of a volcano close to a detected location. A region was declared a detection if the Euclidean distance d between the location of the detection and the location of the volcano on the reference list, was less then $r_{0.5}$, unless $r_{0.5} < 5$ pixels in which case $r_{0.5}$ is replaced by 5, or $r_{0.5} > 15$ pixels in which case $r_{0.5}$ is replaced by 15. Thus, the criterion for a detection was that the detected location be within half the radius of the reference volcano unless the radius is extremely small or extremely large. Empirically it has been found that volcanoes rarely overlap thus effectively eliminating the problem of detecting multiple volcanoes which are very close together.

Notes

1. The nominal pixel spacing in the highest resolution Magellan data products is 75m, but this image was resized slightly.
2. One difference is that the area under an FROC curve cannot be interpreted in the same way as for a true ROC curve.

References

Asker, L. & Maclin, R. (1997a). Ensembles as a sequence of classifiers. *Fifteenth International Joint Conference on Artificial Intelligence*.

Asker, L. & Maclin, R. (1997b). Achieving expert performance on real-world problems using machine learning. *Fourteenth International Conference on Machine Learning*.

Aubele, J.C. & Slyuta, E. N. (1990). Small domes on Venus: characteristics and origins. *Earth, Moon and Planets*, 50/51:493–532.

Baldi, P. (1994). Personal communication.

Brodley, C.E. & Smyth, P. (1997). Applying classification algorithms in practice. *Statistics and Computing*.

Bunch, P.C., Hamilton, J.F., Sanderson, G.K. & Simmons, A.H. (1978). A free-response approach to the measurement and characterization of radiographic-observer performance. *J. Appl. Photo. Eng.*, 4(4):166–171.

Burl, M.C., Fayyad, U.M., Perona, P., Smyth, P. & Burl, M.P. (1994a). Automating the hunt for volcanoes on Venus. *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 302–309). Los Alamitos, CA:IEEE Computer Society Press.

Burl, M.C., Fayyad, U.M., Perona, P. & Smyth, P. (1994b). Automated analysis of radar imagery of Venus: handling lack of ground truth. *IEEE International Conference on Image Processing*, volume III, pp. 236–240.

Burl, M.C., Fayyad, U.M., Perona, P. & Smyth, P. (1996). Trainable cataloging for digital image libraries with applications to volcano detection. (Technical Report CNS-TR-96-01). Pasadena, CA: California Institute of Technology, Dept. of CNS.

Chakraborty, D.P. & Winter, L.H.L. (1990). Free-response methodology: alternate analysis and a new observer-performance experiment. *Radiology*, 174:873–881.

Cherkauer, K. (1996).Personal communication.

Chesters, M.S. (1992). Human visual perception and ROC methodology in medical imaging. *Physics in Medicine and Biology* 37(7):1433–1476.

Cooke, R.M. (1991). *Experts in Uncertainty*. New York,NY:Oxford University Press.

Cross, A.M. (1988). Detection of circular geological features using the Hough transform. *International Journal of Remote Sensing*, 9(9):1519–1528.

Crumpler, L.S., Aubele, J.C., Senske, D.A., Keddie, S.T., Magee, K.P. & Head, J.W. (1997). Volcanoes and centers of volcanism on Venus. In S. W. Bougher, D. M. Hunten, and R. J. Phillips, (Ed.), *Venus II*. University of Arizona Press.

Duda, R.O. & Hart, P.E. (1973). *Pattern Classification and Scene Analysis*. New York, NY:John Wiley and Sons, New York.

Fayyad, U.M., Smyth, P., Burl, M.C. & Perona, P. (1996). A learning approach to object recognition: applications in science image analysis. In S. Nayar and T. Poggio, (Ed.), *Early Visual Learning*. New York, NY:Oxford University Press.

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D. et al. (1995). Query by image and video content – the QBIC system. *Computer*, 28(9):23–32.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*. Second Edition, Academic Press.

Green, D.M. & Swets, J.A. (1966). *Signal Detection Theory and Psychophysics*. New York:Wiley.

Guest, J.E., Bulmer, M.H., Aubele, J., Beratan, K., Greeley, R., Head, J., Michaels, G., Weitz, C. & Wiles, C. (1992). Small volcanic edifices and volcanism in the plains of Venus. *Journal of Geophysical Research Planets*, 97(E10):15949–15966.

Kubat, M., Holte, R. & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, 30, 195–215.

Langley, P. & Simon, H.A. (1995). Applications of machine learning and rule induction. *Communications of the ACM*, 38(11):55–64.

MacMillan, N.A. & Creelman, C.D. (1991). *Signal Detection Theory: A User's Guide*. New York:Cambridge Univesity Press.

Mendel, E., et al. (1997). SAOimage: the next generation. Smithsonian Astrophysical Society, version 1.7.

Moghaddam, B. & Pentland, A. (1995). Maximum likelihood detection of faces and hands. *International Workshop on Automatic Face and Gesture Recognition* (pp. 122–128). Zurich, Switzerland.

Pentland, A., Picard, R.W. & Sclaroff, S. (1996). Photobook - content-based manipulation of image databases. *International Journal of Computer Vision*, 18(3):233–254.

Pettengill, G.H., Ford, P.G., Johnson, W.T.K., Raney, R.K. & Soderblom, L.A. (1991). Magellan: Radar Performance and Data Products. *Science*, 252:260–265.

Picard, R.W. & Pentland, A.P. (1996). Introduction to the special section on digital libraries: representation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):769–853.

Poulton, E.C. (1994). *Behavioral Decision Theory: A New Approach*. New York, NY:Cambridge University Press.

Provost, F. & Fawcett, T. (1997). Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* (pp. 43–48). Newport Beach, CA:AAAI Press.

Richards, J.A. (1986). *Remote Sensing for Digital Image Analysis*. Berlin:Springer-Verlag.

Saunders, R.S., Spear, A.J., Allin, P.C., Austin, R.S., Berman, A.L., Chandlee, R.C., Clark, J., Decharon, A.V. & Dejong, E.M. (1992). Magellan Mission Summary. *Journal of Geophysical Research Planets*, 97(E8):13067–13090.

Simard, P., le Cun, Y. & Denker, J. (1993). Efficient pattern recognition using a new transformation distance. *Advances in Neural Information Processing Systems 5* (pp. 50–58).

Sirovich, L. & Kirby, M. (1987). Low dimensional procedure for the characterization of human faces. *Journal of Optical Society of America*, 4(3):519–524.

Skingley, J. & Rye, A.J. (1987). The Hough transform applied to SAR images for thin line detection. *Pattern Recognition Letters*, 6:61–67.

Spackman, K.A. (1989). Signal detection theory: valuable tools for evaluating inductive learning. *Proceedings of the Sixth International Workshop on Machine Learning* (pp. 160–163). San Mateo, CA:Morgan Kaufman.

Stofan, E.R., Sharpton, V.L., Schubert, G., Baer, G., Bindschadler, D.L., Janes, D.M. & Squyres, S.W. (1992). Global distribution and characteristics of coronae and related features on Venus – Implications for origin and relation to mantle processes. *Journal of Geophysical Research Planets*, 97(E8):13347–13378.

Stonebraker, M. & Kemnitz, G. (1991). The POSTGRES next generation database-management system. *Communications of the ACM* 34(10):78–92.

Stough, T. & Brodley, C. (1997). Image feature reduction through spoiling: its application to multiple matched filters for focus of attention. *Proceedings of the Third Annual Conference on Knowledge Discovery and Data Mining* (pp. 255–259). Newport Beach, CA:AAAI Press

Swets, D.L. & Weng, J. (1996). Using discriminant eigenfeatures for image retrieval. *Pattern Analysis and Machine Intelligence*, 18(8):831–836.

Treitel, S. & Shanks, J. (1971). The design of multistage separable planar filters. *IEEE Transactions on Geoscience and Electronics*, 9(1):10–27.

Turk, M. & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86.

Turmon, M. (1996). Personal communication.

Wiles, C.R. & Forshaw, M.R.B. (1993). Recognition of volcanoes using correlation methods. *Image and Vision Computing*, 11(4):188–196.

Received March 4, 1997
Accepted September 18, 1997
Final Manuscript November 15, 1997