A New Generation of Intelligent Trainable Tools for Analyzing Large Scientific Image Databases

Usama M. Fayyad, Padhraic Smyth, and David J. Atkinson
Jet Propulsion Laboratory,
California Institute of Technology
4800 Oak Grove Drive
Pasadena, CA 91109-8099
Tel: 818-306-6197 Fax: 818-306-6912
Email: Fayyad@aig.jpl.nasa.gov, pjs@galway.jpl.nasa.gov, Atkinson@isd.jpl.nasa.gov

KEY WORDS AND PHRASES
Machine learning, Automated Data Analysis, Intelligent Analysis Tools, Large Image Databases.

1. INTRODUCTION
In a variety of scientific disciplines two-dimensional digital image data is now relied on as a basic component of routine scientific investigation. The proliferation of image acquisition hardware such as multi-spectral remote-sensing platforms, medical imaging sensors, and high-resolution cameras has led to the widespread use of image data in fields such as atmospheric studies, planetary geology, ecology, agriculture, glaciology, forestry, astronomy, diagnostic medicine, to name but a few. Across all of these disciplines there is a common factor: the image data for each application, whether it be a Landsat image or an ultrasound scan, is but a means to an end in the sense that the investigator is only interested in using the image data to infer some conclusion about the physical properties of the target being imaged. In this sense, the image data serves as an intermediate representation to facilitate the scientific process of inferring a conclusion from the available evidence.

In the past, in planetary science for example, image databases were analyzed in a careful manual manner and much investigative work was carried out using hard copy photographs. However, due to the sheer enormity of the image databases currently being acquired, simple manual cataloging is no longer a practical consideration if all of the available data is to be utilized.

A currently familiar pattern in the remote-sensing and astronomy communities is the following: a new image data set becomes available but the size of the data set precludes the use of simple manual methods for exploration. Scientists are beginning to express a need for automated tools which can assist them in navigating through large sets of images. A commonly expressed wish is the following: "is there a tool where I could just point at an object on the screen (or even draw a caricature of it) and then have the algorithm find similar items in the database?"

Note that in this paper the type of problem being addressed differs from the types of problems typically addressed by classical work in machine vision. Machine vision work has focused primarily on image understanding, parsing, and segmentation, with a particular emphasis on detecting and analyzing *human-made* objects in the scene of interest. The focus of this paper is on the detection of *natural*, as opposed to *human-made*, objects. The distinction is important because, in the context of image analysis, natural objects tend to possess much greater variability in appearance than human-made objects. Hence, we shall focus primarily on the use of algorithms that "learn by example" as the basis for image exploration. The "learn by example" approach is potentially more generally applicable compared to model-based vision methods since domain scientists find it relatively easier to provide examples of what they are searching for versus describing a model.

1.1 TWO ILLUSTRATIVE CASE STUDIES
Using ongoing JPL projects as case studies, this paper is intended to provide motivation for the need to develop automated image analysis techniques as well as report on our initial success in the application of pattern recognition and machine learning technology to the general problem of image database exploration. The first project, the Sky Image Cataloging and Analysis Tool (SKICAT), represents an already successful application of

decision-tree learning to classification in the context of a well-understood image analysis problem in astronomy. The second project represents ongoing work which targets a more ambitious problem of dealing with domains where the basic image processing itself is not straightforward: The JPL Adaptive Recognition Tool (JARtool) is being developed for use by planetary geologists on the automated analysis of the Magellan Synthetic Aperture Radar (SAR) images of the planet Venus.

2. SKICAT: AUTOMATED SKY SURVEY CATALOGING

The first case study consists of an application of machine learning techniques to the automation of the task of cataloging sky objects in digitized sky images. SKICAT has been developed for use on the images resulting from the 2nd Palomar Observatory Sky Survey (POSS-II) conducted by the California Institute of Technology (Caltech). The photographic plates collected from the survey are being digitized at the Space Telescope Science Institute (STScI). This process will result in about 3,000 digital images of roughly 23,000 x 23,000 pixels[1] each. The survey consists of over 3 terabytes of data containing on the order of 10^7 galaxies, 10^9 stars, and 10^5 quasars.

The first step in analyzing the results of a sky survey is to identify, measure, and catalog the detected objects in the image into their respective classes. Once the objects have been classified, further scientific analysis can proceed. For example, the resulting catalog may be used to test models of the formation of large-scale structure in the universe, probe galactic structure from star counts, perform automatic identification of radio or infrared sources, and so forth. The task of reducing the images to catalog entries is a laborious time-consuming process. A manual approach to constructing the catalog implies that many scientists need to expend large amounts of time on a visually intensive task that may involve significant subjective judgment. The goal of our project is to automate the process, thus alleviating the burden of cataloging objects for the scientist and providing a more objective methodology for reducing the data sets. Another goal of this work is to classify objects whose intensity (isophotal magnitude) is too faint for recognition by inspection, hence requiring an automated classification procedure. Faint objects constitute the majority of objects on any given plate. We target the classification of objects that are at least one magnitude fainter than objects classified in previous surveys using comparable photographic material.

The learning algorithms used in SKICAT are the GID3* [4] and O-Btree [5] decision tree generation algorithms. In order to overcome limitations inherent in a decision-tree approach, we use the RULER [6] system for deriving statistically cross-validated classification rules from multiple (typically > 10) decision trees. The details of the learning algorithms are beyond the scope of this paper and are therefore not covered here. For details of how rules are generated from multiple decision trees, and for other algorithmic details, the reader is referred to [6,7].

A manual approach to classifying sky objects in the images is infeasible. Existing computational methods for processing the images will preclude the identification of the majority of objects in each image since they are at levels too faint (the resolution is too low) for traditional recognition algorithms or even methods based on manual inspection or analysis. Low-level image processing and object separation are performed by the public domain FOCAS image processing software developed at Bell Labs [11,14]. In addition to detecting the objects in each image, FOCAS also produces basic attributes describing each object. These attributes are standard in the field of astronomy and represent commonly measured quantities such as area, magnitude, several statistical moments of core intensity, ellipticity, and so forth. Additional normalized attributes were measured later to achieve accuracy requirements and provide stable performance over different plates. In total, 40 attributes are measured by SKICAT for each detected object.

2.1 FAINT SKY OBJECT CLASSIFICATION

In addition to the scanned photographic plates, we have access to CCD images that span several small regions in some of the plates. The main advantage of a CCD image is higher resolution and signal-to-noise ratio at fainter levels. Hence, many of the objects that are too faint to be classified by inspection of a

[1] Each pixel consists of 16 bits and represents the intensity in one of three colors.

photographic plate, are easily classifiable in the corresponding CCD image (if available). We make use of the CCD images in two very important ways: CCD images enable us to obtain class labels for faint objects in the photographic plates, and CCD images provide us with the means to reliably evaluate the accuracy of the classifiers obtained from the decision-tree learning algorithms.

In order to produce a classifier that classifies faint objects correctly, the learning algorithm needs training data consisting of faint objects labeled with the appropriate class. The class label is therefore obtained by examining the CCD frames. Once trained on properly labeled objects, the learning algorithm produces a classifier that is capable of properly classifying objects based on the values of the attributes provided by FOCAS. Hence, in principle, the classifier will be able to classify objects in the photographic image that are simply too faint for an astronomer to classify by inspection of the survey images. Using the class labels, the learning algorithms are basically being used to solve the more difficult problem of separating the classes in the multi-dimensional space defined by the set of attributes derived via image processing. This method allows us to classify objects at least one magnitude fainter than objects classified in photographic sky surveys to date.

2.2 RESULTS

We were able to achieve a stable classification accuracy of 94% in classification of sky objects into four classes: *star, galaxy, star-with-fuzz*, and *artifacts* [15]. The latter class represents non-sky objects in the photographs due to film problems, satellite or airplane traces, or other problems. It is noteworthy that using the learning algorithms, we are able to classify objects that are at least one magnitude fainter than objects classified in previous comparable surveys. The SKICAT system is expected to speed up catalog generation by one to two orders of magnitude over traditional manual approaches to cataloging. This should significantly reduce the cost of cataloging survey images by the equivalent of tens of astronomer workyears. In addition, SKICAT classifies objects that are at least one magnitude fainter than objects cataloged in previous surveys. We have exceeded our initial accuracy target of 90%. This level of accuracy is required for the data to be useful in testing or refuting theories on the formation of large structure in the universe and on other phenomena of interest to astronomers.

The catalog generated by SKICAT will eventually contain about a billion entries representing hundreds of millions of sky objects. For the first survey (POSS-I) conducted over 4 decades ago, without the availability of an automated tool like SKICAT, only a small percentage of the data was used and only specific areas of interest were studied. In contrast, we are targeting a comprehensive sky catalog that will be available on-line for the use of the scientific community. Because we can classify objects that are one magnitude fainter, the resulting catalog will be significantly richer in content, containing three times as many sky objects as would have been possible without using SKICAT.

3. JARTOOL: VOLCANO DETECTION IN MAGELLAN-VENUS DATA

The Magellan-Venus data set constitutes an example of the large volumes of data that today's instruments can collect, providing more detail of Venus than was previously available from Pioneer Venus, Venera 15/16, or ground-based radar observations put together [13]. Venus is an extremely volcanic planet (volcanoes are by far the single most visible geologic feature in the Magellan data set); hence, the study of basic volcanic processes is essential to a basic understanding of the geologic evolution of the planet [10]. Central to volcanic studies is the cataloging of each volcano location and its size and characteristics. We are initially targeting the automated detection of the "small-shield" volcanoes (less than 15 km in diameter) that constitute the most abundant visible geologic feature [8] in the more than 30,000 SAR images of the surface of Venus. It is estimated, based on extrapolating from previous studies and knowledge of the underlying geologic processes, that there should be on the order of 10^6 of these volcanoes visible in the Magellan data [1,10].

Identifying and studying these volcanoes is fundamental to a proper understanding of the geologic evolution of Venus. However, locating and parameterizing them in a manual manner is forbiddingly time-consuming. Hence, we have undertaken the development of techniques to partially automate this task. The primary constraints for this particular

problem are that the method must be reasonably robust and fast.

3.1 The Approach

There has been little prior work on detecting naturally occurring objects in remotely-sensed images. Most pattern recognition algorithms are geared towards detecting straight edges or large changes in texture or reflectivity. While this works well for detecting *human-made* objects, approaches such as edge detection and Hough transforms deal poorly with the variability and noise present in typical remotely sensed data [3,12].

We are developing a system that consists of three distinct components: focus of attention, feature extraction, and classification learning. Figure 1 gives a block diagram of the approach. The focus of attention component is designed primarily for computational efficiency. Its function is to quickly scan an input image and roughly determine regions of interest (regions potentially containing objects similar to those specified by the scientist). Given a set of detected regions of interest, the remaining task is to discriminate between the volcanoes and false alarms. A current focus of the research is to find a useful feature-representation space --- although nearest neighbor classifiers can provide reasonably accurate results, a representation based purely on pixels will tend to generalize poorly. For the purposes of incorporating prior knowledge, the ideal feature set would be expressed in the form of expected sizes, shapes, and relative geometry of slopes and pits, namely, the same features as used by the scientists to describe the volcanoes. However, due to the low signal-to-noise ratio of the image, it is quite difficult to gain accurate measurements of these features, effectively precluding their use at present. The current focus of our work is on a method which automatically derives robust feature representations. The current method is based on performing a singular value decomposition of training images (15 x 15 pixel vectors centered at volcanoes) to find the eigenvectors of the data. In turn, the dominant eigenvectors (principal components) provide the means to translate pixels into a low-dimensional feature space. In the latter, classification learning is used to distinguish between true volcanoes and focus of attention (FOA) false alarms.

3.2 Status and Preliminary Results

We have constructed several training sets using 75-m/pixel resolution images labeled by the collaborating geologists at Brown University to get an initial estimate of the performance of the system. The FOA component typically detects more than 80% of all the volcanoes, while generating 5-6 times as many false alarms. Using features derived from both segmentation and principal component methods [2] has resulted in accuracies of the order of 85% of the volcanoes detected by FOA. It is important to clarify that these are initial results and with further effort we hope to be able to significantly improve the accuracy. Demonstrating the general applicability of this approach to the detection of other Venusian features as well as images from other missions will be the next step. So far the emphasis has been placed mainly on developing the computer tools to allow scientists to browse through images and produce training data sets (as well as partial catalogs) within a single integrated workstation environment.

4. Concluding Remarks

Natural object detection and characterization in large image databases is a generic task which poses many challenges to current scientific analysis tasks. The SKICAT and Magellan SAR projects are typical examples of the types of large-scale image database applications which will become increasingly common --- for example, the NASA Earth Observing System Synthetic Aperture Radar (EOS SAR) satellite will generate on the order of 50 GBytes of remote sensing data per hour when operational. In order for scientists to be able to effectively utilize these extremely large amounts of data, basic image database navigation tools will be essential. Our existing JPL projects have so far demonstrated that efficient and accurate tools for natural object detection are a realistic goal provided there is strong prior knowledge about how pixels can be turned into features and from there to class categories. With the astronomy problem there was sufficient strong knowledge for this to be the case: with the volcano data, the knowledge is much less precise and consequently the design of effective object detection tools is considerably more difficult.

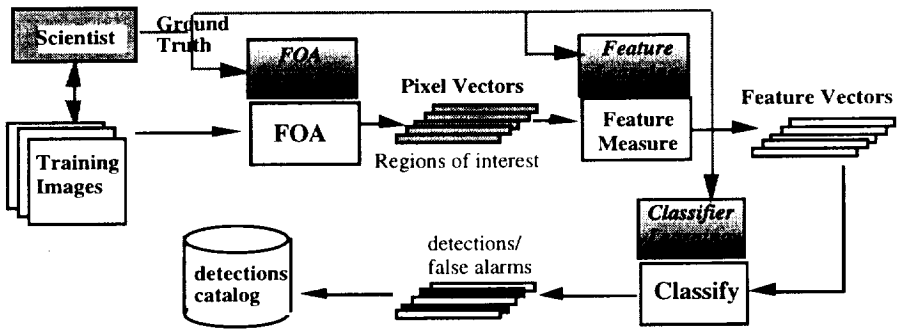

Figure 1. Block Diagram of the JARtool System

We believe that trainable tools for object recognition/cataloging will soon become a necessity. The alternative of writing purpose specific programs customized to individual problems is simply unrealistic and too constrained. The alternative of manual analysis by the scientists is no longer feasible due to the large database sizes.

ACKNOWLEDGMENTS

SKICAT work is a collaboration between Fayyad (JPL) and N. Weir and S.G. Djorgovski of Caltech's Astronomy Department. JARtool work is a collaboration between Fayyad and Smyth (JPL) and M.C. Burl and P. Perona (Electrical Engineering Department, Caltech); the domain scientists are J. Aubele and L. Crumpler (Dept. of Geological Sciences, Brown University). The research described in this paper was carried out in part by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

REFERENCES

[1] J. C. Aubelle and E. N. Slyuta, ``Small domes on Venus: characteristics and origins," in Earth, Moon and Planets, 50/51, 493--532, 1990.

[2] M.C. Burl, U.M. Fayyad, Perona, P., Smyth, P. and Burl, M.P. (1994) "Automating the Hunt for Volcanoes on Venus", *Proc. of The Computer Vision and Pattern Recognition Conference (CVPR'94)*, IEEE Press.

[3] A.M. Cross, *Int. J. Remote Sensing*, 9,no.9, 1519-1528, 1988.

[4] U.M. Fayyad (1991). *On the Induction of Decision Trees for Multiple Concept Learning*. Ph.D. Dissertation, The University of Michigan.

[5] U.M. Fayyad and K.B. Irani (1992) "The attribute selection problem in decision tree generation." *Proceedings of the Tenth National Conference on Artificial Intelligence, AAAI-92*. San Jose, CA.

[6] U. Fayyad, N. Weir, and S.G. Djorgovski (1993). "SKICAT: a machine learning system for automated cataloging of large scale sky surveys." *Proc. of 10th Int. Conf. on Machine Learning*, Amherst, MA.

[7] U.Fayyad, P.Smyth N. Weir, & S.Djorgovski (1994). "Automated Analysis and Exploration of Image Databases", *J.Intelligent Info. Systems* (in press).

[8] J. E. Guest et al. (1992) *Journal Geophys. Res.*, 97, E10, 15949.

[9] J. W. Head et al. (1991) ``Venus volcanic centers and their environmental settings: recent data from Magellan," EOS 72, p.175, *American Geophysical Union Spring meeting abstracts*.

[10] J. W. Head et al. (1992) *Journal Geophysical Res.*, 97, E8, 13,153-13,197.

[11] J.Jarvis and A.Tyson (1981) *Astronomical Journal* 86:41.

[12] S. Quegan et al, (1988) *Trans. R. Soc. London*, A 324, 409-421.

[13] Science, special issue on Magellan data, April 12, 1991.

[14] Valdes (1982) Instrumentation in Astronomy IV, SPIE vol. 331, no. 465.

[15] N.Weir, S.G.Djorgovski, U.M.Fayyad, et al (1992) "SKICAT: A system for the scientific analysis of the Palomar-STScI Digital Sky Survey." *Proc. Astronomy from Large databases II*, p. 509, Munich, Germany:Euro. Southern Observatory.