Knowledge Discovery in Large Image Databases: Dealing with Uncertainties in Ground Truth

Padhraic Smyth[†], Michael C. Burl[†§], **Usama** M. **Fayyad**[†], and **Pietro Perona**[‡]
[†] Jet Propulsion Laboratory
California Institute of Technology
MS 525-3660 — Pasadena, CA 91109
{pjs,fayyad}@aig.jpl.nasa.gov
[‡] Electrical Engineering
California Institute of Technology
MS 116-81 — Pasadena, CA 91125
{burl,perona}@systems.caltech.edu

Abstract

This paper discusses the problem of knowledge discovery in image databases with particular focus on the issues which arise when absolute ground truth is not available. It is often the case that a user exploring a large database is in search of items that are not easy to define completely. One way to circumvent this problem is to ask the user to specify examples of the item of interest. However, labelling of specific items may in itself be inconsistent, whether it is multiple labels from a single labeller at different times or labels from different labellers. The paper discusses issues which arise in terms of elicitation of subjective probabilistic opinion, estimation of basic scientific parameters of interest given probabilistic labels, learning from probabilistic labels, and effective evaluation of both user and algorithm performance in the absence of ground truth. The problem of searching the Magellan image data set in order to automatically locate and catalog small volcanoes on the planet Venus is used as a case study. In the absence of calibrated ground truth, planetary scientists provide subjective estimates of ground truth based on visual inspection of Magellan images. Data from the Magellan volcano detection project is used to illustrate the various techniques which we have developed to handle these issues. The primary conclusion of the paper is that knowledge discovery methodologies can be modified to account for lack of absolute ground truth provided the sources of uncertainty in the data are carefully handled.

1 Introduction

At last year's KDD workshop we presented initial results on building art automated system for locating and cataloging small volcanoes on the surface of Venus, using radar images returned by the Magellan spacecraft [Fayy93]. In particular we discussed some of the issues which are unique to the problem of knowledge discovery in *image* databases. This paper tackles the problem of developing and evaluating knowledge discovery systems for image databases when ground truth is not available. Remote-sensing image analysis problems where there is a complete absence of ground truth are increasingl y common: the Magellan volcano detection project is used as an illustrative example. Indeed, the problem of uncertainties in ground truth is a pervasive one that affects not only image applications. For example, in medical diagnosis absolute diagnoses may not be known with certainty, yet may be assumed to be certain. In general, if one is given a set of labeled cases in a database, it maybe the case that the labels contain "noise" for various reasons. In practical applications, this factor may become so significant, that simply ignoring it is not a realistic option, The volcano detection problem certainly falls in this category, as we shall illustrate.

The phrase "lack of ground truth" requires some comment: what is typicall y available is not a complete lack of ground truth, but rather, subjective estimates of ground truth. In other words, a domain expert (or group of same) examines the available data (an image) and provides a subjective estimate of the class labels for particular locations within the image. Hence, there is an additional source of noise in the data, namely *the noisy estimates from the expert.* It is critical that this noise source be modelled and calibrated as far as possible. The alternative is to ignore the noisy nature of the labelling process, assume that the labels

are correct, and condition all algorithm design, parameter estimation, and performance evaluation on this premise, If the labelling process is not very noisy this is often the practical approach.

In this paper we focus on the case where there is considerable visual ambiguity in the images, such that there will be significant differences on the same data between the labelings of the same expert at different times and between different experts. Ignoring this source of noise is likely to lead to a significantly miscalibrated system. For example, in the volcano detection problem, the local density of volcanoes in a given planetary region is a parameter of significant geological relevance: as will be discussed later, ignoring the subjective uncertainty in the labelling would lead to a systematic bias in terms of over-estimating local volcano densities. Since volcano counts and distributions are used primarily to infer the gelogical history/evolution of the planet, this systematic bias in turn would bias other scientific inferences made based on it, possibly leading to misguided theories about the underlying phenomena.

First the general background to the problem is described, namely the Magellan mission and the scientific importance and relevance of investigating volcanism on Venus. We then review our overall philosophy behind developing "user-trainable" tools for knowledge discovery in databases, focusing in particular on the development of machine learning and pattern recognition tools which allow a scientist to train a search algorithm based on sample objects of interest. This sets the stage for the main discussion of the paper: the modelling and treatment of subjective label information. We outline the experimental methodology and basic principles of subjective elicitation, using data obtained from the participating scientists as examples. The following issues are then discussed in some detail: noise models to relate probabilistic labels to ground truth, performance evaluation metrics which incorporate probabilistic labels, and learning algorithm modifications. We note that previous work in the pattern recognition literature has dealt with some of the general theoretical aspects of this problem [Lug92, Silver80]; the originality of the work described here lies in the handling of the ground truth ambiguity problem in the context of a large-scale, real-world, image analysis problem.

2 Application Domain: Finding Volcanoes on Venus

Both in planetary science and astronomy, image analysis is often a strictly manual process and much investigative work is carried out using hardcopy 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.* The Magellan Venus data set is a typical instance of the now familiar data glut situation in science, medicine, industrial applications, as well as security and defense contexts.

The background to this work is the notion of a trainable image analysis system; a scientist trains the system to find certain geological features by giving it examples of features to be located. The scientist can thus customize the tool to search for one type of feature versus another simply by providing positive and negative examples. In addition to automating laborious and visually-intensive tasks, the system provides an objective, examinable, and repeatable process for detecting and classifying objects in images. This allows scientists to base their analysis results on uniformly consistent data, free from subjective variations that invariably creep in when a visually exhausting task requiring many months or years is undertaken.

The Magellan spacecraft transmitted back to earth a data set consisting of over 30,000 high resolution synthetic aperture radar (SAR) images of the Venusian surface. This data set is greater than that gathered by all previous planetary missions combined — planetary scientists arc literally swamped by data [Fayy94]. The stud y of volcanic processes is essential to an understanding of the geological evolution of the planet [Head9 1]. Central to volcanic studies is the cataloging of each volcano location and its size and characteristics. There are estimated to be on the order of 106 visible volcanoes scattered throughout the 30,000 images [Aubele90]. Furthermore, it has been estimated that manually locating all of these volcanoes would require on the order

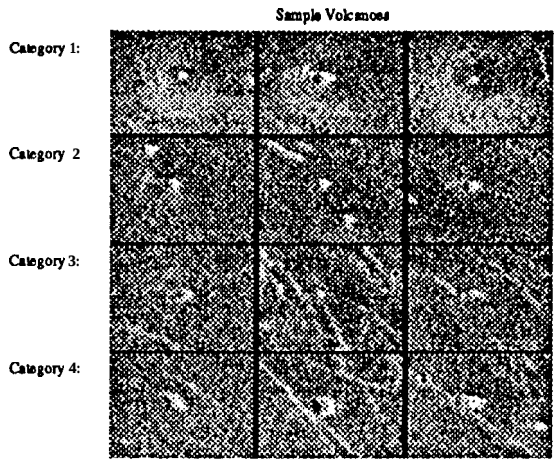

Figure 1: A small selection of volcanoes from four categories as labeled by the geologists.

of 10 man-years of a planetary geologist's time to carry out.

Empirical results using spatial eigenrepresentations (combined with supervised classification algorithms) have demonstrated that a trainable image analysis system can be roughly competitive with humans in terms of classification accuracy [Burl94, Fayy94]. The system uses a matched filter (for example, the mean of locally windowed training examples of volcanoes) to initially focus attention on local regions of interest. The detected local regions are projected into a subspace consisting of significant principal directions of the training data — the subspace itself is determined by selecting the most significant components produced by a singular value decomposition of the training data. Supervised learning is used to produce a model which can discriminate between volcano and non-volcano local regions in the projected subspace. A simple maximum-likelihood Gaussian classifier with full covariance matrices has been found to perform as well as alternative non-parametric methods such as neural networks and decision trees for the problem of discriminative learning in the projected eigenspace [Burl94].

3 Eliciting Ground Truth Estimates from Scientists

In the volcano location problem, as in many remote sensing applications, real ground truth data dots not exist, No one has ever actually been to the surface of Venus (apart from a Russian robotic lander which melted within a fcw minutes), and despite the fact that the Magellan data is the best imagery ever obtained of Venus, scientists cannot always determine with 100% certainty whether a particular image feature is indeed a volcano.

In principle, for a given local region of interest, a scientist can provide a subjective probability that a volcano exists at that point given the local intensity values. It can been shown [Smyth94] that eliciting subjective probabilities is preferable to forcing a "yes/no" decision — in particular, it allows more accurate estimation of the underlying discriminant surfaces (compared to learning from''ycs/no" labels) for a given training sample size. However, this result is conditioned on the assumption that the scientists arc providing perfect unbiased subjective probability estimates, It is well known that accurate elicitation of subjective probabilities from humans is quite difficult and subject to various calibration errors and biases.

3.1 Defining Sub-Categories of Volcanoes

A more effective approach in practice is to have the scientists label training examples into quantized probability bins, where the probability bins correspond to visually distinguishable sub-categories of volcanoes. In particular, we have used 5 bins: (i) summit pits, bright-dark radar pair, and apparent topographic slope, all clearly visible, probability 0.98, (ii) only 2 of the 3 criteria in category (i) are visible, probability 0.80, (iii) no summit pit visible, evidence of flanks or circular outline, probability 0.60, (iv) only a summit pit visible, probability 0.50, (v) no volcano-like features visible, probability 0.0. The probabilities correspond to the mean probability for a particular bin (the mean probability that a volcano exists at a particular location given that it has been identified as belonging to a particular bin) and were elicited based on considerable discussions with the participating planetary geologists. How we use these probabilities for both training and evaluation will be discussed in more detail in the next few sections.

Figure 1 shows some typical volcanoes from each category. The use of quantized probability bins to attach levels of certainty to subjective image labelling is not new: the same approach is routinely used in the evaluation of radiographic image displays to generate subjective ROC (receiver operating characteristic) curves [Chest92]. However, this paper extends the basic approach by defining the notion of probabilistic ROC curves (see Section 5).

3.2 Methodologies for Collecting Subjective Label Information

Participating in the development of the detection algorithm are planetary geologists from the Department of Geological Sciences at Brown University. We are fortunate to have direct collaboration with two members of this group who are also members of the Volcanism Working group on the Magellan Science team (Jayne Aubele (JA) and Larry Crumpler (LC)). Both of these scientists have extensive experience in studying both earth-based and planetary volcanism and have published some of the standard reference works on Venus volcanism [Aubele90, Head91]. Hence, their collective subjective opinion is (roughly speaking) about as expert as one can find given the available data and our current state of knowledge about the planet Venus.

It is an important point that, in the absence of absolute ground truth, the goal of our work is to be as comparable in performance as possible to the scientists in terms of labelling accuracy. Absolute accuracy is not measurable for this problem. Hence, the best the algorithm can do is to emulate the scientist's performance — this point will become clearer when we discuss performance metrics later in the paper.

A standard Magellan image consists of 1000 x 1000 pixels, where the pixels are 75m in resolution for the results referred to in this paper. Small volcano diameters are typically in the 2–3km range, i.e., 30 to 50 pixels wide. Volcanoes are often spatially clustered in volcano fields. As a consequence, most of the volcanoes are expected to be found in about 10-20% of the total number of images, and within these images there may number as many as 100 or more volcanoes, although typicall y the number is in the 10-50 range.

The standard manner in which we obtain labels is to have a labeller interact with an X-windows software tool whereby he or she uses mouse-clicks to locate candidate volcanoes. Starting with an initially blank image, the labeller proceeds to sequentially click on the estimated centers of the volcanoes. The labeller is then prompted to provide a subjective label estimate from a choice of categories 1–4 as described above — by default, locations which are not labelled are considered to have label "O" (non-volcano). Clearly it is possible that based on the visual evidence, for the same local image patch, the same label may not be provided by different labellers, or indeed by the same labeller at different times. In addition to labels, the labeller can also provide a fitted diameter estimate by fitting a circle to the feature. Figure 2 shows a typical image labelled in this manner,

After completing the labelling, the result is an annotation of that image which can be stored in standard database format — the unique key to the image is a label event, which corresponds to a particular latitude/longitude (to the resolution of the pixels) for a particular labeller at a particular time (since the same

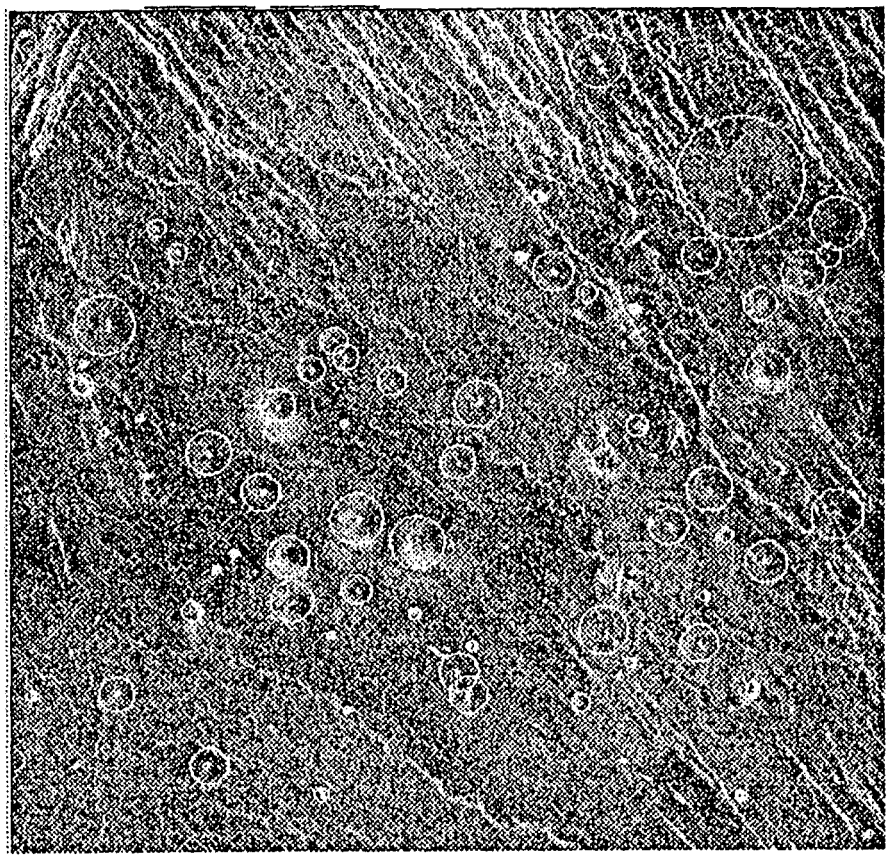

Figure 2: Magellan SAR image of Venus with consensus ground truth showing size and locations of small volcanoes.

labeller may relabel an image multiple times). It is this database which provides the basic reference framework for deriving estimates of geologic parameters, training data for the learning algorithms, and reference data for performance evaluation. A simple form of spatial clustering is used to determine which label events (from different labellers) actually correspond to the same geologic feature (volcano). It is fortunate that volcanoes tend not to overlap each other spatial] y and thus maintain a separation of at least a few kilometers, and also that different scientists tend to bc quilt consistent in their centring of the mouse-clicks — mean differences of about 2.5 pixels (Euclidean distance) have been found in cross comparisons of label data from scientists JA and LC, which is reasonable considering the precision onc can expect from mouse location on a screen. Hence, accurate location of the volcanoes is not in itself much of problem.

Table 1 shows the confusion matrix between labellers JA and LC for a set of 4 images. The (i, j)th clement of the confusion matrix counts the number of label events which corresponded to labeller A generating label i and labeller B generating label j, where both labels were considered to belong to the same visual feature, i.e., were within a few pixels of each other. The (i, O) entries count the instances where labeller A provided label i, but labeller B did not provide any label — entry (0,0) is always defined to be zero by default. Ideally, the confusion matrix would have all of its entries on the diagonal if both labellers agreed completely on all events. Clearly, however there is substantial disagreement, as judged by the number of off-diagonal counts in the matrix. For example, label 3's arc particularly noisy, in both "directions." Note that on the order of 50% of the label 3's detected by either labeller arc not detected at all by the other labeller. On the other hand only on the order of 10% of the label 1's of either labeller arc missed by the other. The

matrix underlines the importance of modelling probabilistic labels for this particular problem.

Table 1: Confusion Matrix of Scientist A Vs. Scientist B.

		Scientist A				
		Label 1	Label 2	Label 3	Label 4	Not Detected
Scientist B						
Label	1	19	8	4	1	3
Label	2	9	8	6	5	5
Label	3	13	12	18	1	37
Label	4	1	4	5	24	15
Not Detected		4	8	29	16	0

4 Relating Probabilistic Labels to Ground Truth

Before we describe the particular methods used for training, estimation, and evaluation it is informative to look at a relatively simple model for the noise introduced into the data by the subjective labelling process.

We will use the shorthand v and $\bar{v}$ to denote the events "volcano present" and "volcano not present", respectively, and l to denote a particular label, $0 \leq l \leq l_{max}$ ($l_{max} = 4$ for the labelling problem). Let V be a binary random variable taking values v and $\bar{v}$, and let L be another discrete random variable taking values l, $1 \leq l \leq l_{max}$. The shorthand notation of "v" for "V = v," etc., will be used. Note that we assume that labelling is *stochastic* rather than *deterministic* in the sense that presented multiple times with the same local image region, a scientist may not always provide the same label. The relevant probabilities we arc interested in are conditional probabilities of the form $p(\text{volcano}|\text{label}) = p(v|l)$. In particular, marginal probabilities such as $p(\text{volcano}) = p(v)$ are not well-defined without reference to a particular region of a particular size.

Figure 3: Causal Model 1 of Volcano Labelling Process.

Consider Figure 3 which identifies a simple causal model: volcanoes arc mapped to an image intensity $\underline{i}$, which in turn is mapped to a label by the scientists, There is an implicit conditionalization on local pixel regions of fixed size, i.e., the labelling process is effectively treated as a series of decisions on such local regions. From Figure 3 we arc ultimately interested in determining the probability of a volcano given $\underline{i}$. To train and evaluate our models wc need to estimate terms such as $p(v|l)$. If we expand this out, we have to condition on all possible realizations of the image intensity $\underline{i}$;

$$p(v|l) = \sum_{\underline{i}} p(v|\underline{i}, l)p(\underline{i}|l) \tag{1}$$

Given the dimensionality of $\underline{i}$ (all possible intensities of a local region), this method of estimating $p(v|l)$ is clearly impractical. Note that the above equation can be rewritten as:

$$p(v|l) = \sum_{\underline{i}} p(v|\underline{i})p(\underline{i}|l) \tag{2}$$

since by the causal model of Figure 3, V *is* conditionally independent of L.

It is convenient to assume that the volcanoes correspond to visually distinguishable categories, namely "types." In addition, "type O" will be used to identify all local images not covered by the "well-distinguished" types (i.e., non volcanoes in general). "Type" will be treated as another random variable T', taking values $1 \leq t \leq t_{max}$, where $t_{max} = l_{max}$ typically. Conceptually it is useful to imagine that there is an Oracle who can unambiguously identify types given intensity information; the main point is that we do not have access to such an Oracle, but rather have access only to fallible scientists who provide labels, noisy estimates of types. In other words, T is an unobserved, hidden variable, while L *is* observed directly.

Figure 4: Causal Model 2 of Volcano Labelling Process: volcanoes to types to labels.

To circumvent the problems of estimating probabilities conditioned on intensity values the following simplification of the model is proposed: replace the high dimensional intensity $\underline{i}$ with the low-dimensional T in the causal model, T can be considered a quantization of the intensity map. The effect is to remove any dependency on intensity values in the model, which can now be written as

$$p(v|l) = \sum_t p(v|t)p(t|l) \tag{3}$$

The dependence of types on volcanoes will be assumed given by the scientists as a general piece of prior information — in particular, $p(v|t)$, for $t = 1,2,3,4$ are the subjective probabilities we have elicited from the scientists which described the mean probability that a volcano exists at a particular location, given that it belongs to a particular type. These subjective probabilities are not conditioned on *labels* per se, but on the *types*, i.e., $p(v|t) \in \{0.98,0.80,0.60,0.5,0.0\}$, $t \in \{1,2, 3,4,0\}$.

The $p(t|l)$ terms in Equation (3) represent the estimation noise resulting from the fact that scientists are unable to specify, with 100% certainty, the particular "type" of a volcano. Determination of these probabilities is rendered non-trivial by the fact that the true types t arc never directly observed, and thus some assumptions about the relationship between T and L must be made in order to infer their dependence. At this point in time, estimating the $p(t\ |l)$ terms from multiple labellings of the same data represents work in progress — one proposed method is outlined in Appendix 1, but we will not discuss the $p(t|l)$ estimation in any more depth. Appendix 2 outlines the procedures for handling data which has been labelled by multiple experts.

Note that the overall effect of the above models will be to reduce our overall confidence that a typical local region is a volcano, given some labelling information — this has direct implications for estimating the overall numbers of volcanoes in a particular region, and so forth. For example, in accordance with the models described in Appendix 2, local regions which which have label disagreements between labelings will be down-weighted compared to volcanoes which receive unanimous labellings.

5 Performance Evaluation: Probabilistic Free-Response ROC Analysis

Given that the scientists cannot classify each object with 100% confidence, how can we assess how well our algorithms arc performing? We have investigated the idea of "consensus ground truth": a consensus-based probabilistic labelling is generated by multiple scientists working together in labelling images. The individual labelings and the results of the automated detection system described earlier are then evaluated in terms of performance relative to the consensus. The performance of an algorithm is considered to be

satisfactory if, compared to consensus ground truth, its performance is as good as that of an individual scientist,

As a performance evaluation tool we use a variation of the well-known receiver operator characteristic (ROC) methodology. The purpose of the ROC is to determine the complete range of performance of a decision system in terms of its estimated detection rate versus false alarm rate. Consider a binary hypothesis testing problem (equivalently a binary classification or discrimination problem): the 2 mutually exclusive and exhaustive hypotheses are denoted as ω_1 and ω_2. Assume there exists a certain fixed cost $c_{ij}, 1 \leq i, j, \leq 2$, which is incurred when ω_i is the chosen hypothesis and ω_j is true. The observed data (the features) correspond to a d-dimensional random variable $\underline{X}$ taking values $\underline{x}$. Standard Bayesian decision theory [VanTrees68] shows that the optimal decision rule (in terms of minimum cost) must be of the form:

$$\frac{p(\omega_1|\underline{x})}{p(\omega_2|\underline{x})} \underset{<}{\geq} \frac{c_{10} - c_{00}}{c_{01} - c_{11}} \tag{4}$$

In general, when the costs are not known exactly (as is often the case in practice), one can treat the term on the right as a varying decision threshold *t*.

The probability of detection, $p_d(t)$, is defined as the probability that the decision algorithm chooses ω_1 when ω_1 is the correct hypothesis; similarly, the probability of false alarm, p_{fa} (t) is the probability that the decision algorithm chooses ω_1 when ω_2 is the correct hypothesis. Both probabilities are implicit functions oft, the decision threshold: as t increases the system becomes more conservative in its decisions, reducing the false alarm rate; as t decreases the system will increase its detection rate, but at a cost of increasing the false alarm rate. When the conditional densities $p(\omega_1|\underline{x})$ and $p(\omega_2|\underline{x})$ are known exactly one can determine Pal(t) as a function of p_{fa} (t); this plot is known as the ROC and provides the characteristic signature of a decision system over the entire range of possible detection/false alarm operating points.

Since in practical applications $p(\omega_1|\underline{x})$ and $p(\omega_2|\underline{x})$ are not known, the ROC must be estimated directly from data. This is straightforward provided the decision system is producing either a direct estimate of the ratio $r = \frac{p(\omega_1|\underline{x})}{p(\omega_2|\underline{x})}$, or some monotonic function of r. The estimation procedure is to vary r (or a monotonic function of same) as a decision threshold on a labelled training data set and count the resultant numbers of detections and false alarms for each value of r. A training set of size N produces N + 1 operating points (including the end points of (0.0,0.0) and (1.0, 1.0)). One converts the number of detections and number of false alarms to probabilities by dividing by the total number of training examples of class ω_1 and class ω_2 respectively. Thus, one can plot an empirical ROC, the estimated probability of detection versus estimated probability of false alarm.

For the volcano. detection problem, the reference labels are taken from the consensus labelling, i.e., this is in effect treated as ground truth. Class ω_1 corresponds to volcanoes, class ω_2 to non-volcanoes. False alarms correspond to label events which are categorized by the detection system as being of class volcano, when the consensus labelling indicates a non-volcano event, i.e., a local region which was not labelled. There is a problem in defining the probability of false alarm, since it is difficult to define the prior probability of class ω_2. For example, should the prior probability be proportional to the number of pixels in the image which were not labelled as volcanoes ? This definition does not make much intuitive sense, since it would be a function of the number of pixels in a given image (one wants the ROC to be invariant to changes in such parameters) and also since it would result in an astronomical y high prior in favour of non-volcanoes.

Hence, instead of plotting detection versus false alarm rates, we usc detection rate versus *false alarms per total number of detections* — this normalized false alarm rate is a much mom useful parameter since it is invariant to the size and resolution of the images used to determine the ROC. This plot is no longer directly interpretable as a standard ROC since the false alarm rate axis can now run from O% to some arbitrary percentage greater than 100%, i.e., there may have been more false alarms detected than true detections in total for some threshold operating points. This slightly modified ROC methodology is essentially the same

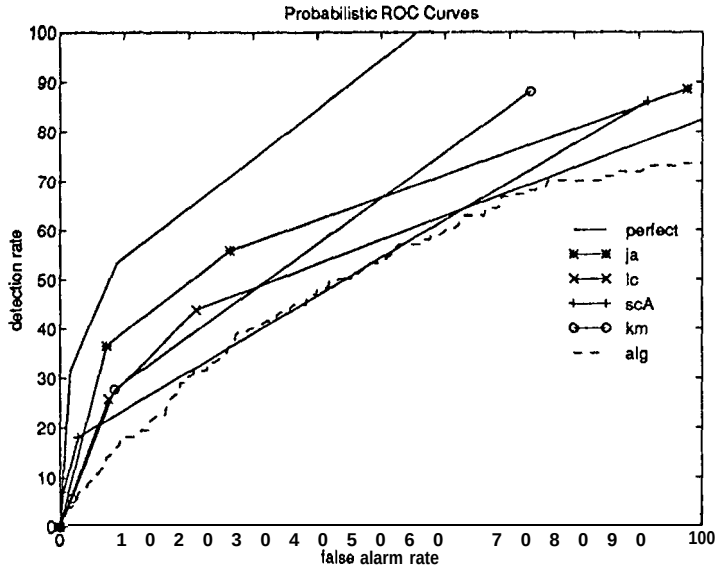

Figure 5: Probabilistic ROC curves showing the performance of four planetary geologists and our automated algorithm.

as the free-response ROC (FROC) used in evaluation of radiology display systems [Bunch78, Chakra90].

Furthermore, standard ROC and FROC approaches assume that ground truth is known. When ground truth is known only in probabilistic form as described earlier, one must allow for the fact that each detected local region is only a detection with probability p: there is an associated probability of $1 - p$ of it being a fatse alarm. These probabilities are determined with reference to the consensus labelling: for example, for a given threshold value, if the detection system being evaluated detects a particular local region which has been labelled by the consensus as category 2 (probability of volcano = 0.8), then this is counted as 0.8 of a detection and 0.2 of a false alarm. The overall effect is to drag the non-probabilistic ROC (where no allowance is made to for the probabilistic effect) towards the "center" of the plot, away from the ideal "false alarm rate 0.0, detection rate 1.0" operating point. Furthermore, the ideal "perfect" operating point is no longer achievable by *any* system, since the reference data is itself probabilistic. Hence, an effective optimal ROC is defined by exactly matching the probabilistic predictions of the consensus — one can do no better. Note that the actual probability values used to weight detections and false alarms can be determined by the estimation methods described in the Appendices; for the results reported here, only consensus labels were used as reference, and they were assumed to be noiseless in terms of estimation error (i.e., $p(t|l) = 1.0$, $t = l$ and $p(t|l) = 0.0$, $t \neq l$).

We denote the probabilistic method with the normalized false alarm rate as the probabilistic FROC (PFROC) (for problems where the base rate of class ω_2 is well-defined one could also define the probabilistic ROC (PROC)). Within this framework, the performance of a human labeller can only be determined within the resolution of the quantized probabilistic bins used in the subjective labelling process, With k bins, one can determine the location of k operating points on the PFROC, including the (0.0, 0,0) point. Again, the relative probability weights for each category of volcano correspond to the probabilities assigned to the consensus labels.

Figure 5 shows a PFROC curve for four images, comparing the performance of 4 planetary geologists (JA, LC, and two others) and the current version of the automated detection system. A consensus of 2 planetary geologists (JA and LC) was used as reference. The consensus labelling was determined some time *after the* individual labelings by JA and LC. The algorithm was as described in Section 2 (and in more

detail in [Burl94]) and was evaluated in cross-validation mode (trained on 3 of the images, and tested on the 4th, repeated 4 times). In total, the consensus labelling produced 163 volcanoes, which correspond to the 100% point on the y-axis: as described above, the false alarm rates are determined relative to the 163 "true" detections.

The upper curve is the optimal achievable performance on these 4 images relative to the consensus labelling. The other curves largely parallel this curve but arc 10 to 50% less accurate in terms of detection rate at a fixed false alarm rate. Note that the algorithm performance is comparable with the scientists over the 20 to 80% false alarm range.

The performance of the individual scientists relative to the consensus is not as good as might have been expected a priori. Nevertheless, we feel these results provide a true picture of the relative accuracy with which volcanoes can be detected in the Magellan images. This underlying ambiguity in volcano detectability should be recognized and factored into any scientific inferences made based upon labelings by individuals or machine algorithms. The consensus labelings themselves arc also probably noisy (to a lesser degree), but we have not quantified this yet.

6 Other Aspects of Probabilistic Labels

The acknowledgement of uncertainty in the labelling can have other significant impacts on overall image analysis methodologies. For example, as described in detail in [Smyth94] and [Burl94], the matched filter generation, SVD subspace generation, and discriminant learning procedures can all be modified to account for probabilistic labels. The general approach is based on the notion of assigning fractions of a training data sample to each class in proportion to the subjective label weight: for example, a category 2 volcano might be treated as 0.8 of a sample for the volcano class and 0.2 of a sample for the non-volcano class. When the estimation noise can be calibrated and there are labels from multiple experts, the methods of Appendix 1 and 2 can be used to determine more accurate relative class weighings. While this weighted treatment of probabilistic labels leads to improved performance in theory [Smyth94], in our experiments to date we have found no improvement in performance by learning from probabilistic labels as compared to the default approach of treating all labelled items as examples of class volcano. Investigation of the data revealed that the subspace projection technique was destroying any probabilistic structure which existed in the data at the level of the intensity maps, i.e., category 1 's, 2's, 3's and 4's were all being projected into the same region of feature space (as revealed by 2-d scatterplots of various feature pairs) and completely overlapped each other without any structure. If the probabilistic structure had been preserved, one would expect to see the 1's to be further away from the non-volcano class than the 2's and so forth. This is an example of a learning algorithm dealing with a feature space (SVD filter responses) which is different than that on which the labelling is performed (local intensity maps), with the result that the probabilistic labels do not relate in any useful way to the space in which learning is taking place. As a consequence, a detection algorithm based only on SVD filter responses cannot reproduce accurate posterior probability estimates which match those of the scientists subjective labels. A current direction of investigation is to seek projections which preserve the probabilistic label information, which in turn should result in better PFROC performance.

Estimation of various spatial statistics can also bc conditioned on the probabilistic nature of the labels — for example, non-parametric kernel density estimates of the volcano diameters (an important geological "signature") can bc modified to take probabilistic labels into account as described in [Smyth94]. Densities which arc not unimodal are particularly sensitive to probabilistic labels: incorrect treatment of the labels can lead to oversmoothing of real modes, or the introduction of spurious ones. Once again, the actual values of the probabilistic labels area function of the particular noise model one chooses to use as described in the Appendices. Estimation of spatial statistics in this manner is a topic of current investigation.

7 Conclusion

The major focus of this paper is the treatment of uncertainty in the training data when designing and evaluating knowledge discovery systems for image databases. The net effect of ground truth ambiguity is to propagate an extra level of subjective noise into processes such as training learning algorithms, performance evaluation methodologies, and estimation of spatial statistics of scientific interest. Handling this uncertainty requires the introduction of special techniques such as the probabilistic free-response ROC (PFROC) methodology discussed in this paper. If issues of ground truth ambiguity are simply ignored, a knowledge discovery system may appear to perform inaccurately, while in fact it is not possible for it to perform any better. By characterizing user performance relative to a standard (such as consensus labelling in our case) one can target the more realistic, and possibly achievable, goal of matching the discovery algorithm's performance with that of an individual user compared to the chosen standard.

The techniques described in this paper provide a framework for accurate estimation and evaluation of basic image quantities of interest for applications where absolute ground truth is not available. Such applications arc becoming increasingly common as remote-sensing platforms provide orders of magnitude more data and well-calibrated ground truth constitutes a tiny (and perhaps even zero) fraction of the overall data set.

References

[Aubele90] J. C. Aubele and E. N. Slyuta, "Small domes on Venus: characteristics and origins," in *Earth, Moon and Planets, 50/51, 493-532,1990.*

[Chakra90] Chakraborty, D. P., and Winter, L. H. L. (1990), " Free-Response methodology: alternate analysis and a new observer-performance experiment; *Radiology, 174,873-881.*

[Bunch78] Bunch, P. C., Hamilton, J. F., Sanderson. G. K., and Simmons, A. H., (1978), "A Free-Response approach to the measurement and characterization of radiographic-observer performance," *J. Appl. Photo. Eng.*, vol.4, no. 4, pp.166-171.

[Burl94] Burl, M.C., Fayyad, U.M., Perona, P., Smyth, P., and Burl, M.P. (1994), "Automating the hunt for volcanoes on Venus," to appear *Computer Vision and Pattern Recognition Conference, CVPR-94.*

[Chest92] Chesters, M. S., (1992), "Human visual perception and ROC methodology in medical imaging," *Phys. Med. Biol.*, vol.37, no.7, pp. 1433-1476.

[Fayy93] Fayyad, U. M., and Smyth, P., (1993) "Image database exploration: progress and challenges; *Proceedings of the 1993 AAAI Workshop on Knowledge Discovery in Databases.*

[Fayy94] Fayyad, U. M., P. Smyth, N. Weir, and S. Djorgovski (1994), "Automated analysis and exploration of large image databases: results, progress, and challenges, " *Journal of Intelligent Information Systems,* in press.

[Lug92] Lugosi, G., (1992) "Learning with an unreliable teacher," *Pattern Recognition,* vol. *25,* no.1, pp.79-87.

[Head91] Head, J. W., et al. (1991). "Venus volcanic centers and their environmental 'settings: recent data from Magellan," EOS 72, p. 175, American Geophysical Union Spring meeting abstracts.

[McCon81] McConway, K. J., (1981), "Marginalization and linear opinion pools," *J. Amer. Statist. Assoc.,* vol.76, pp.410–414.

[Science91] *Science, special issue on Magellan data,* April 12,1991.

[Silver80] Silverman, B., (1980), "Some asymptotic properties of the probabilistic teacher," *IEEE Trans. Info. Theory,* IT-26, *no.2,* pp.246–249.

[Smyth94] Smyth, P., (1994), "Learning with probabilistic supervision," in *Computational Learning Theory and Natural Learning Systems 3,* T. Petsche, M. Kearns, S. Hanson, R. Rivest (eds), Cambridge, MA: MIT Press, to appear.

[VanTrees68] Van Trees, H. L., (1968), *Detection, Estimation, and Modulation Theory: Part 1,* John Wiley and Sons, New York, pp.23-31.

Appendix 1: Estimating the $p(t|l)$ terms from multiple labellings

Consider that we have a database of N labelled local regions. Assume that each local region has been examined m times, either by m different scientists or groups of same, or the same scientist multiple times, or some combination of same (the extension to the case where some subsets of local regions have been examined by different numbers of labellers or groups of labellers is trivial and will be omitted to keep notation simple). Hence, each local region has been labelled as one of the 4 labels 1/2/3/4 by at least one of the m labellers.

For each label event assign a "vote" of $1/m$ to each label 1/2/3/4 each time that a labeller assigns that label, and a "vote" of $1/m$ to the label O if a labeller did not label it at all. Implicit here is an assumption that each labeller is being weighted equally — extensions to the case of non-equal weighting are straightforward and are not dealt with here. We can interpret the sum of the votes for a particular label (from different labellers) as the probability that local intensity $\underline{i}$ will be assigned label 1. More formally, we deftne the estimator

$$\hat{p}(l|\underline{i}) = \frac{1}{m}\sum_{k=1}^{m}\delta(l, v_k(\underline{i})) \tag{5}$$

where $\delta(x,y)$ = O unless x = y, and $v_k(\underline{i})$ is the label provided by the kth labeller for the local intensity map $\underline{i}$.

We can now estimate the marginal probability that an arbitrary labeller will assign label 1 to a local region, by summing over all intensities:

$$\hat{p}(l) = \sum_{j=1}^{N}\hat{p}(l|\underline{i})\hat{p}(\underline{i}^j) \tag{6}$$

$$= \frac{1}{N}\sum_{j=1}^{N}\hat{p}(l|\underline{i}^j) \tag{7}$$

where j is an index over the N local regions in the database. To estimate $p(t|l)$ by Bayes' rule we first need to estimate $p(t,l)$. The following estimator is defined:

$$\hat{p}(t,l) = \sum_{\underline{i}}\hat{p}(t,\underline{i},l) \tag{8}$$

$$= \sum_{\underline{i}}\hat{p}(t|l,\underline{i})\hat{p}(l|\underline{i})p(\underline{i}) \tag{9}$$

$$= \frac{1}{N}\sum_{j=1}^{N}\hat{p}(t|l,\underline{i}^j)\hat{p}(l|\underline{i}^j) \tag{lo}$$

$$= \frac{1}{N}\sum_{j=1}^{N}\hat{p}(t|\underline{i}^j)\hat{p}(l|\underline{i}^j) \tag{11}$$

since the type t is independent of the label l given the local intensity $\underline{i}$. If we define the estimator for $\hat{p}(t|\underline{i}^j)$ *to be the same as* for $\hat{p}(l|\underline{i}^j)$ (as in Equation 7 above), the estimation process is complete, since all necessary terms are now defined and can be estimated directly from the database. Finally, we have that

$$\hat{p}(t|l) = \frac{\hat{p}(t,1)}{\hat{p}(l)} \tag{12}$$

As a simple example, this method was applied to two labellings of the same 9 images with the following results:

$$\hat{p}(T=1|L=1) = 0.80, \quad \hat{p}(L=1) = 0.11 \tag{13}$$

$$\hat{p}(T=2|L=2) = 0.68, \quad \hat{p}(L=2) = 0.22 \tag{14}$$

$$\hat{p}(T=3|L=3) = 0.70, \quad \hat{p}(L=3) = 0.29 \tag{15}$$

$$\hat{p}(T=4|L=4) = 0.78, \quad \hat{p}(L=4) = 0.22 \tag{16}$$

$$\hat{p}(T=0|L=0) = 0.5, \quad \hat{p}(L=\mathrm{O}) = 0.16 \tag{17}$$

Labelling of 1's and 4's appears to be the most accurate, labelings of 2's and 3's less so. Furthermore, it is estimated that 16% of the local regions identified (out of 330 which were labelled by at least one labeller in the 9 images) are truly non-volcanoes.

At this point of our research, we have not settled on using a particular estimation technique for the $p(t|l)$ terms — we plan to have more detailed results to report by the time of the workshop.

Appendix 2: Handling Data from Multiple Labellers

There are two distinct ways in which data from multiple labellers can be combined in the context of this model. In the first method, a single $p(t|l)$ matrix of probabilities can be derived to characterize the mean estimation noise of all of the labellers; then, given a particular set of labels for the same local image region, one can estimate

$$\hat{p}(v|\underline{i}) = \sum_{l=1}^{l_{max}} \hat{p}(v|l)\hat{p}(l|\underline{i}) \tag{18}$$

where $\hat{p}(l|\underline{i})$ is simply the proportion of labellers which voted for label 1 as described in Appendix 1, assuming a linear weighting of experts. Linear weighting of multiple experts (using a set of coefficients which sum to 1) is the preferred choice when combining probabilities due to the fact that it is the only weighting scheme to satisfy the marginalization property [McCon81] — a weighting scheme satisfying this property is invariant to the manner in which the event space is partitioned.

The second *method* utilizes a different $p(t|l)$ matrix for each labeller and then uses a linear weighted sum of these estimates to arrive at composite estimates for the posterior p(t) terms — the posterior $p(v)$ terms are then calculated in the standard manner using $p(v|t)$.

As with the estimation of the $p(t|l)$ terms themselves, both of these schemes represent work in progress on which we plan to report more details at the workshop.