Active Learning for Classifying a Spectrally Variable Subject

Patricia G. Foschi
Romberg Tiburon Center for Environmental Studies
San Francisco State University

Huan Liu
Department of Computer Science and Engineering
Arizona State University, Tempe

Abstract—Classifying *Egeria densa*, Brazilian waterweed, in scan-digitized color infrared (CIR) airphotos by automated methods presents a challenging problem due to a number of variable and unfavorable conditions: changes in imaging conditions, problems associated with water-related subjects, and other environmental changes as well as expected lack of spectral separation between *Egeria* and other land cover classes in CIR imagery. To address these challenges, we are developing an interactive computer system based on data mining techniques with Active Learning capabilities. The key components of this system are: feature extraction, automatic classification, active learning, and experimental evaluation. We anticipate creating an interactive learning system that can learn from human analysts by relating results to extracted objects and that can learn analytic rules for classification. In this paper, we report the concept of the system, preliminary experimental results, and anticipated future work.

I. Background and Problem

Rapid advances in remote sensing and in data storage have made huge quantities of image data available for analysis. The spatial, spectral, and radiometric resolutions of digital remotely-sensed imagery have greatly improved in recent years. However, automated classification of high-resolution imagery has not kept pace with developments in data collection and storage. The need for data collection by aerial photographic surveys also still exists for many applications. Improvements in orthorectification techniques have allowed such photography, recent and historic, to be digitized, geometrically corrected, and integrated into a database more rapidly and at reduced cost. Such progress has particularly aided applications that require high-resolution spatial data, for example: precision agricultural monitoring, individual-tree forest inventory, wetland and small waterway mapping, and invasive species monitoring.

In many cases, human experts are heavily involved in the interpretation and analysis of high-resolution imagery. Visual/manual image interpretation and analysis procedures are often time-consuming and costly, not repeatable, and dependent on the varying abilities of the interpreters. Individual expertise is hard to transfer from one interpreter to another, which contributes to high training costs. There is a significant gap between fast routine data collection and the slow interpretation and analysis of complex and detailed images in multidimensional (spatial and temporal) space. Monitoring *Egeria densa*, an invasive submergent weed, by remote sensing represents such a case.

Egeria densa, commonly called Brazilian waterweed, has grown uncontrolled in the Sacramento-San Joaquin Delta of Northern California for over 35 years and now covers about 2400 hectares of waterways. The presence of this exotic weed is disrupting navigation

and recreational uses of waterways, clogging irrigation intake trenches, and causing reservoir-pumping problems. The cost of pumping Clifton Court Forebay reservoir alone has increased by over US$1 million per year. The *Egeria* invasion has displaced native flora and probably affected native fauna.

In 1997, the California Department of Boating and Waterways (DBW) started developing a control program to manage *Egeria* in the Delta. Research scientists at the Romberg Tiburon Center for Environmental Studies (RTC) at San Francisco State University were hired to assess the effects of control protocols on fish and other fauna and to estimate the areal extent of the *Egeria* [1]. Monitoring the areal extent of *Egeria* using scan-digitized color infrared (CIR) aerial photographs has occurred every year since then. The database of photographs is online at: http://romberg.sfsu.edu/~egeria.

Classifying *Egeria* in CIR airphotos by automated methods presents a challenging problem due to a number of variable and unfavorable conditions. These include changes in imaging conditions (e.g., film exposure, vignetting, scanning anomalies), problems associated with water-related subjects (e.g., turbidity, sun glint, surface reflectance due to wind), and other environmental changes (e.g., exposure of *Egeria* at extremely low tide, shadows falling upon the water, algae cover over *Egeria*). Figure 1, a detail of a scan-digitized CIR aerial photograph, illustrates the spectral variations in *Egeria* that may occur even within a short distance. Digital analyses also indicated that subtle changes—for example, in *Egeria* canopy density, film vignetting, or water turbidity—produced overlapping spectral response patterns. In addition, the spectral response patterns for *Egeria* do not separate well from those of other land cover classes in CIR imagery. For example, dense well-submerged *Egeria* appears black and is confused with shadows on land; *Egeria* exposed during very low tide appears reddish and is confused with terrestrial vegetation. Clearly, traditional computer-assisted

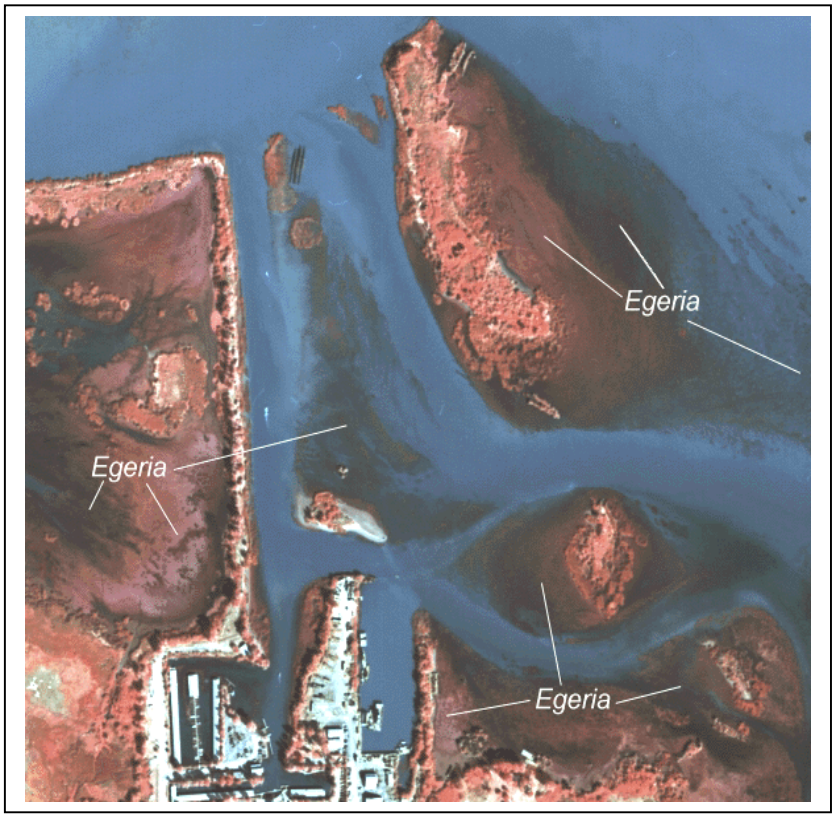

Figure 1. Scan-digitized CIR aerial photograph showing spectral variations in *Egeria*.

multispectral classification methods are problematic under these conditions, and visual/manual image interpretation and analysis procedures are daunting. Post-classification processing is frequently needed to clean up noisy patches in classification maps.

To address these challenges, we are exploiting the latest developments in data mining, investigating novel combinations of effective methods in feature extraction, classification, and machine learning, and proposing a computer system that implements the novel combinations. Such a system that integrates feature extraction, classification, and machine learning is expected to accomplish routine tasks, highlight areas for verification or further analysis, minimize expert intervention, and consequently allow experts more time to concentrate on more vexing problems. In this paper, we report our research addressing some challenges outlined above by developing an interactive computer system with Active Learning. The key components of the system are as follows:

- Feature extraction - to derive relevant and effective features for automatic classification [2];
- Automatic classification – to achieve speed, accuracy, generalization, and automation;
- Active learning - to use machine learning techniques to minimize expert intervention without performance deterioration; and
- Experimental evaluation – to compare the performance with and without the newly developed systems.

We anticipate creating an interactive learning system that can learn from human analysts by relating results to extracted objects and that can learn analytic rules for classification. In the following, we introduce the system, elaborate on the key components, present preliminary results, and conclude with future work.

II. Integrated Interactive System

Human experts have been heavily engaged in the problem of detecting *Egeria* in CIR imagery for practical applications. Due to the variability of this landscape class, a significant amount of time and interpreter training has been needed to discriminate *Egeria* in this imagery. Characteristic and unusual colors, shapes and textures have had to be verified by ground surveys. Typical *Egeria* sites and associations have been cataloged over time. Many problems have only been identified by multitemporal information.

Well-trained image analysts are very good at analyzing a new image. Manual processing, though, is inherently unscalable. Experts can process only a small portion of the available unprocessed images (unlabeled data) at any given time. As a first step, we aim to reduce the amount of involvement of human experts. Therefore, the new system uses Active Learning and is interactive in defining features and resolving uncertainties. The images are divided into two groups: a training set and a testing set. Ideally, we wish to have a very small training set because the smaller training set (i.e., a few subsets of images) will require less involvement of human experts. Also, we are trying to minimize the involvement of experts during testing.

A flow chart of the interactive system is presented in Figure 2. The following sections describe the components of this system in more detail.

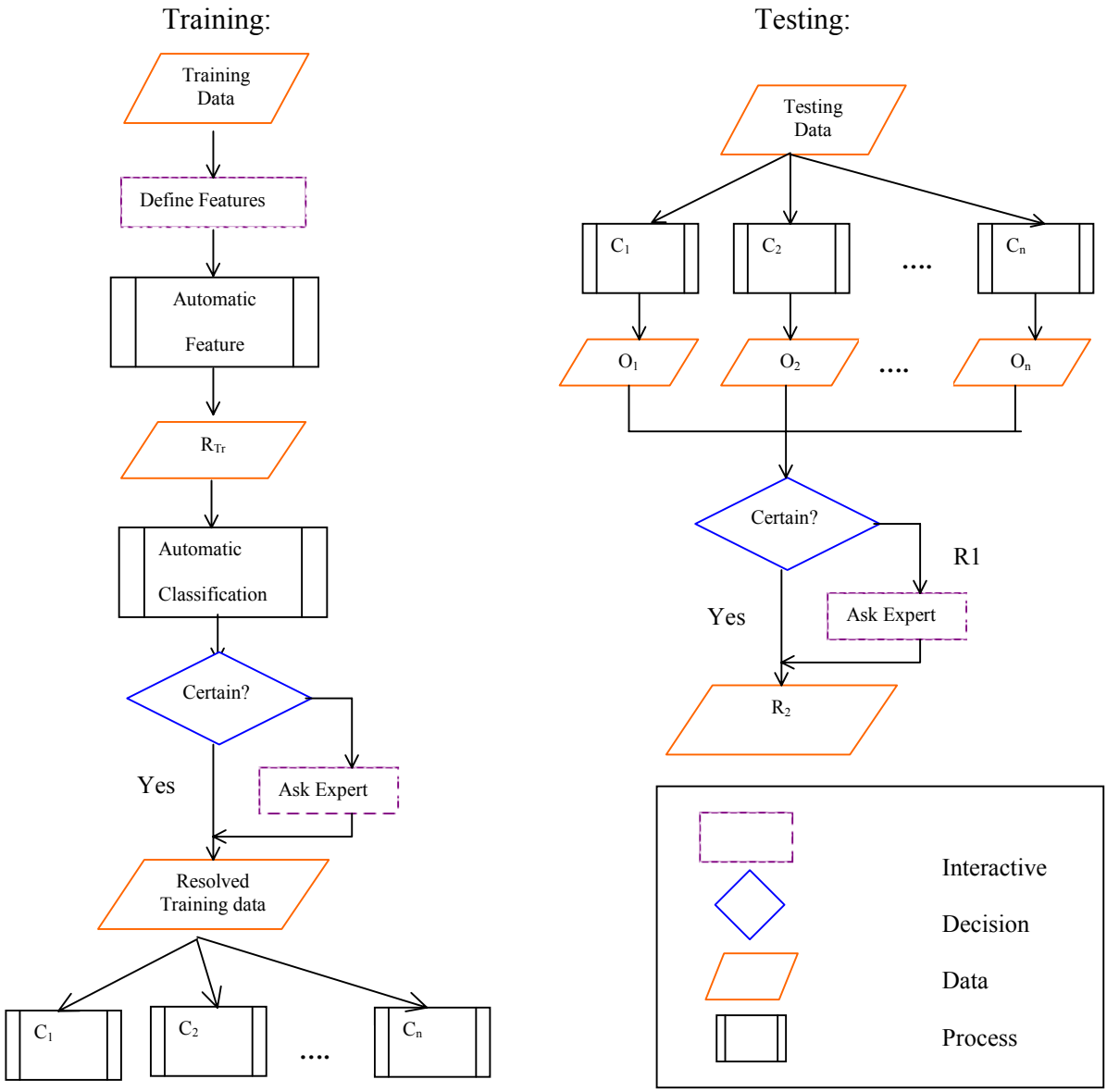

Figure 2. Flow chart of interactive system for detecting *Egeria* in digital imagery.

A. Feature Extraction

The primary step of the system for detecting *Egeria* in digital imagery is feature selection and extraction. The domain under study is scan-digitized high-resolution aerial photographs. In order to effectively detect *Egeria*, we must select or extract regions of interest from the entire image. Features are needed that can describe regions and that can separate the interesting from the rest [3].

There are numerous features, which can be used to select regions of interest. The color feature alone has proven ineffective in detecting *Egeria*. Consequently, spatial information or features—like context, shape, and texture—are more likely to aid automated classification. In our trials, we are very conservative in introducing new features. We have used two textural features and one color feature to start.

To calculate a particular texture, a template is created using a standard size of 10 pixels x 10 pixels. The original image is divided into blocks of 10 x 10 pixels. Each block is

compared with the template using histograms and the results are limited with a threshold value. Then, depending on these values, a region matrix is created. The same process is repeated for a second texture. There may be common regions identified by both textures, which correspond to the presence of *Egeria* with greater certainty. Uncertain regions will require expert interpretation.

The texture templates describe: texture containing *Egeria* in its purest form and texture having both *Egeria* in a waterbody and some landmass adjoining it. The first texture was chosen to identify those regions that are greatly affected by *Egeria*. The second texture was chosen since *Egeria* occurs near the banks of waterbodies. Hence, these two textures represent a significant part of the occurrences of *Egeria*. More textures will be added as the analysis proceeds.

For the color feature, we chose: the shades of black that are the most common color of *Egeria* in the images under consideration. The average color of each 10 x 10 block is compared to threshold values. This process effectively captures all areas where the color matches that of a shade of black. This also captures some areas that are not *Egeria*. The color feature provides one more dimension to describe *Egeria*. In combination with the texture features, classification results are expected to improve. The efficiency of the extraction depends mainly on the template textures that are chosen and on the threshold values for color values that allow or disallow borderline areas.

B. Automatic Classification Using an Interactive System

An automatic classification system, called I-Class, was built using these three features. Table I shows the decision rules in I-Class. In the last column, 1 means most probably *Egeria* (called Certain), 2 means probably *Egeria* (called Uncertain), and 0 means not *Egeria*. A 2 requires resolution by human experts. Therefore, we have designed and developed an interactive system facilitating the resolution of 2s.

TABLE I
RULES FOR DETERMINING IF *EGERIA* OR NOT

	Texture 1	Texture 2	Color 1	*Egeria*
Rule 1	1	1	1	1
Rule 2	1	1	0	1
Rule 3	1	0	0	1
Rule 4	1	0	1	1
Rule 5	0	0	0	0
Rule 6	0	0	1	2
Rule 7	0	1	0	2
Rule 8	0	1	1	1

To date, we have developed a prototype version of the interactive system that is currently workable in the MATLAB environment. This preliminary system allows the user to resolve doubtful areas and determine the presence or absence of *Egeria*. The human expert knows best in these situations. The system allows the user to open images that have already been classified, either for viewing or for resolving conflicts.

The interactive system can be used in both training and testing (or classification) phases. During the training phase, the expert resolves those areas labeled 2s by I-Class. A resolved image is an image with only areas labeled as 0 or 1. After an image is resolved, it is

analyzed by the Active Learning subsystem (to be illustrated in the next section). During the testing phase, the Active Learning subsystem outputs labels for regions in the test images. Some regions are very probably *Egeria* and some are not; some are uncertain. Uncertain regions are labeled by the expert using the interactive system. After either phase, the output of the interactive system is expected to be a completely resolved image.

It is reasonable to assume that I-Class would work well; however, this is only partially true. It works well in the training phase. The problem with I-Class is that it is predefined and does not generalize from the training data and the labeled regions resolved by the expert. Therefore, I-Class will probably fail when it is applied to test images. In order to generalize from the data, some data mining algorithms were employed. Since they are also used in Active Learning, we will discuss them next.

C. Active Learning

Classification is a predictive process in which applications or systems are automated to predict outputs without, or with as little intervention as possible from, human experts. The job of a learning algorithm is to relieve the experts from making all decisions and to automate the process of making predictions of class labels of new instances.

A classification learning algorithm first "learns" the mapping between the inputs and the output classes by using the training data. It then uses this to predict the classes for the test data. Classification learning is sometimes referred to as Supervised Learning because the actual classes for the training examples have to be provided as the algorithm learns the mapping. Active Learning [4, 5] is a Supervised Learning algorithm that combines the results from multiple classification algorithms.

In Active Learning, initially multiple algorithms are trained using the same training data to learn the patterns in the data. Then each of the algorithms is tested with the test data. The outputs for each sample are collected from all the classifiers and compared. The set of all the samples for which all the algorithms agree is selected and labeled appropriately. The rest of the samples with Uncertain output classes (i.e., different learning algorithms identified different classes) are then referred to the experts for resolution. This resultant set is denoted by R1 in Figure 2. The interactive system introduced earlier prompts the expert to manually label the classes by regional blocks. The resultant set of samples, denoted by R2 in Figure 2, can then be appended to the training set to retrain the algorithms. This procedure is repeated each time there is a new set of testing data made available. As the procedure is repeated with more resolved and correct training data, it is expected that the interactive system will be needed less and less. This happens because the algorithms at each stage will learn more and will be able to resolve more over time.

The Active Learning Algorithm, shown below as Table II, can start with a small set of labeled data. At each step, the samples that have been learned by the algorithms do not have to be labeled again. Only those samples not learned previously by the algorithms need to be labeled. Thus, Active Learning saves relabeling time as well as time required by the domain experts to label the data. Advantages of Active Learning are:

- In most of the classification problems, the training data available are not labeled (without class labels). Active Learning can start by using only a small part of the data being labeled to train the algorithms.
- Labeling done by the domain experts is usually an expensive process. Active Learning reduces the load on the domain experts by classifying most of the samples automatically.

- As the process is repeated, the amount of work to be done by the domain experts continues to decrease.

TABLE II
ACTIVE LEARNING ALGORITHM

- Given: L labeled instances, U unlabeled instances, E expert
- Loop until all instances in U are labeled
  - Use L to train h1
  - Use L to train h2
  - Pick one instance i from U
  - if h1(i) != h2(i) then
    Ask E to label i and
    Add i to L and remove i from U
    else label i according to h1

Many machine learning algorithms could be used in Active Learning. Since the comprehensibility of learning results is crucial, we are experimenting with two learning algorithms proven to be efficient and powerful in learning, yet distinct in underlying principles: decision tree induction (DTI) and the Naïve Bayes Classifier (NBC) [6]. DTI aims to build a compact decision tree with pure leaf nodes as soon as possible. One way to achieve that is to find the most promising feature that can split the data and to do so recursively until some criterion is met. The degree of a feature's promise can be measured by the information gained or the weighted difference of entropy before and after splitting the data. The entropy of 2-class data with n+ number of positive instances and n– negative instances (n is the sum of n+ and n–) is calculated as:

$$E(n+, n-) = -\frac{n+}{n}\log\frac{n+}{n} - \frac{n-}{n}\log\frac{n-}{n}.$$

NBC adopts a different approach and tries to maximize:

$P(C_i \mid x)$ by using Bayes Rule to maximize $P(x \mid C_i)P(C_i)$

As more images are classified by human experts, the machine learning algorithms are repeatedly strengthened and improved. The key is to search for patterns using relevant objects repeatedly identified by various features in different images.

III. DATA AND EXPERIMENTS

The preliminary trials are being tested using CIR aerial photographs of the Sacramento-San Joaquin Delta flown in October 2000 at 1:24,000 scale. The flight was in the morning during a low-tide period. The low-tide period was selected to optimize the detection of *Egeria* in the photography. The time of day for the flight was a trade-off between a low enough sun angle to provide more reflected sunlight and a high enough sun angle to minimize shadows from objects on the banks falling upon the water. Aquatic submergent species are usually best discriminated in either near IR (NIR) data (shallow depths) or visible data (deeper depths). Since color IR film detects reflected radiation in visible (green and red light) and NIR spectral regions, this type of film was expected to detect

vegetation within a range of water depths. The airphotos were scan-digitized and color separated to create 3-band digital imagery at a nominal 2-meter spatial resolution. In other words, each pixel represents approximately a 2m x 2m area on the ground. Small patches of *Egeria* can be resolved and mapped using this fine resolution. The imagery was not geometrically corrected. Subsets were selected from 12 of the airphotos to represent cases in which *Egeria* is readily interpretable to an image analyst. However, *Egeria* does not appear the same in all subsets. The system was trained using one of subsets, and it was tested using the other 11 subsets.

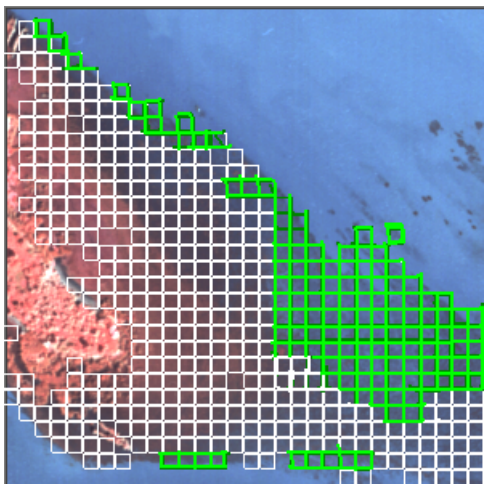

Image 1000_2m_bb1
Certain (white) and Uncertain (green)
blocks before Active Learning

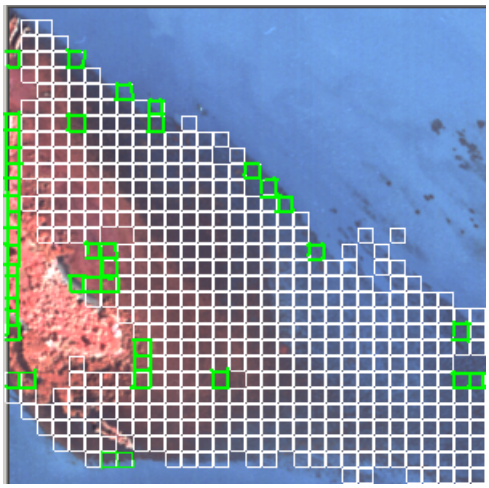

Image 1000_2m_bb1
Certain (white) and Uncertain (green)
blocks after Active Learning

Image 1000_2m_di1
Certain (white) and Uncertain (green)
blocks before Active Learning

Image 1000_2m_di1
Certain (white) and Uncertain (green)
blocks after Active Learning

Figure 3. Output of the Data Mining System before and after Active Learning.

The results for two of the testing images 1000_2m_bb1 and 1000_2m_di1 are shown in Figure 3. The images on the left show the Certain blocks (shown in white) and Uncertain blocks (shown in green) before Active Learning. The images on the right display the Certain and Uncertain blocks after Active Learning. It is clear that the number of Certain blocks has

increased and the number of Uncertain blocks has decreased after Active Learning. The Uncertain blocks are referred to the experts to be resolved. These experiments suggest that Active Learning, by reducing the number of Uncertain blocks, can decrease expert interaction in the classification process.

To evaluate the results of the testing, we used three criteria: the reduction in expert interaction, precision, and recall. These criteria are defined as:

- Reduction: the reduction in the number of Uncertain blocks.
- Precision: the fraction of the retrieved information that is relevant; the fraction of the classification which is correctly identified.
- Recall: the fraction of the relevant information that is retrieved; the fraction of the information sought which is correctly identified.

Table III summarizes the evaluation results for all testing images used in the experiments.

TABLE III
SUMMARY OF RESULTS

Image Name	Before Active Learning				After Active Learning			
	Certain	Uncertain	Precision	Recall	Certain	Uncertain	Precision	Recall
1000_2m_bb1	338	191	0.9595	0.6614	469	60	0.9391	0.8982
1000_2m_di1	301	133	0.7181	0.7461	339	95	0.6591	0.7719
1000_2m_hr1	248	342	0.2956	0.9309	503	87	0.1478	0.9441
1000_2m_ft1	847	29	0.3778	0.9795	872	4	0.3743	0.9990
1000_2m_lps1	227	426	0.2032	0.9056	317	336	0.1470	0.9150
1000_2m_ls1	281	306	0.3180	0.6807	587	0	0.1845	0.8249
1000_2m_lvi1	150	505	0.2863	0.8909	499	156	0.0896	0.9280
1000_2m_orh1	209	466	0.1901	0.7571	449	226	0.1010	0.8638
1000_2m_qi1	235	217	0.6099	0.8471	452	0	0.3583	0.9572
1000_2m_vc1	213	76	0.8931	0.6383	223	66	0.8736	0.6537
1000_2m_wi1	305	539	0.5393	0.8186	344	500	0.5142	0.8803

In Table III, the reduction in expert interaction can be seen in the reduction in the number of Uncertain blocks before and after applying Active Learning. The recall values for all test images should be large enough so that the instances retrieved are comparable to the actual (relevant) instances. However, recall and precision are inversely related. Therefore, the disadvantage of having very high recall values is that precision can decrease rapidly. While these results are satisfactory for preliminary trials, more study is needed to determine the overall evaluation and accuracy of the system.

As trials proceed, we intend to refine the choice of texture and color templates and to add new features that describe other spatial information indicative of the presence of *Egeria*. For example, a new template is needed to eliminate wet agricultural land, a class known to cause misclassification.

IV. SUMMARY AND CONCLUSIONS

To address the challenges posed by a spectrally variable subject, we are exploiting the latest developments in data mining, investigating novel combinations of effective methods—including feature extraction, automatic classification, and machine learning—and proposing a computer system that implements the novel combinations. Such a system is

expected to accomplish routine tasks, highlight areas for verification or further analysis, minimize expert intervention, and consequently allow experts more time to concentrate on more difficult problems. To date, we have created a prototype interactive system that uses Active Learning. We anticipate further developments and refinements.

Although we use images of *Egeria densa* as our testbed, many underlying principles and design methodologies can be extended to other domains of wetland/waterway monitoring. Providing an automated solution to such monitoring alone would have far-reaching applications since rapidly changing conditions occur in all tidal marshland and in wetlands in general. In addition, the automatic data extraction algorithms can be deployed in other cases where data are massive and complex [7]. The Active Learning systems can be applied in many cases in which labeling images is time-consuming and, more often than not, images are unlabeled.

ACKNOWLEDGMENTS

This research has been supported by a California Department of Boating and Waterways contract awarded to Patricia Foschi at San Francisco State University. Part of this award was subcontracted to Huan Liu at Arizona State University (ASU). Deepak Kolippakkam and Amit Mandvikar at ASU have participated in developing systems outlined in the paper and helped draw figures and run experiments. Gary Fields, Yukari Matsumoto, and Mami Odaya at SFSU have supported this research by participation in ground data collection, image interpretations, and training set selections.

REFERENCES

[1] P.G. Foschi, "Egeria densa acreage and percent coverage in the Sacramento-San Joaquin delta," in *Egeria densa Control Program*, CA Department of Boating and Waterways, Vol. II, Report 4, March 2000.

[2] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, 1998.

[3] H. Liu and H. Motoda, Eds., *Feature Extraction, Selection, and Construction for Data Mining*. Kluwer Academic Publishers, 1998.

[4] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Machine Learning*, vol. 15, pp. 201-221, 1994.

[5] N. Roy and A. McCallum, "Toward optimal active learning through sampling estimation of error reduction," *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

[6] I.H. Witten and E. Frank, *Data Mining*. Morgan Kaufmann Publishers, 2000.

[7] H. Liu and H. Motoda, Eds., *Instance Selection and Construction for Data Mining*. Kluwer Academic Publishers, 2001.