# Detecting a Spectrally Variable Subject in Color Infrared Imagery Using Data-Mining and Knowledge-Engine Methods

Patricia G. Foschi, Gary Fields II, and Huan Liu

***Abstract*—To classify *Egeria densa*, Brazilian waterweed, in scan-digitized color infrared aerial photographs, we are developing automated methods based on data-mining and knowledge-engine techniques. In this paper, we present progress to date, compare the results of the two approaches, and discuss current problems and anticipated solutions.**

***Index Terms*—Active learning, class-specific ensemble, data mining, knowledge engine, rule-based classification.**

## I. Background and Problem

Airborne data collection, including aerial photography, is still needed for many applications. For example, monitoring invasive weeds and mapping wetland species frequently necessitate high-resolution data and flexible collection times. Monitoring *Egeria densa*, an invasive submergent weed, by remote sensing ideally requires 2-meter (at minimum) color infrared (CIR) imagery collected during morning low-tide conditions. *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. This exotic weed is displacing native flora, disrupting navigation and recreational uses of waterways, clogging irrigation intake trenches, and causing reservoir-pumping problems. CIR aerial photography and other airborne CIR imagery have been used to monitor the areal extent of *Egeria* [1]. The image database is at: http://romberg.sfsu.edu/~egeria.

There is a significant gap between fast routine airborne data collection and the slow interpretation and analysis of the resulting detailed and often complex data sets. Finding an efficient computerized method for detecting and mapping *Egeria* in CIR imagery would greatly enhance monitoring and support control protocols.

Manuscript received April 5, 2004. This work was supported by the California Department of Boating and Waterways.

P. G. Foschi is with the Romberg Tiburon Center for Environmental Studies, San Francisco State University, Tiburon, CA 94920-1205 USA (phone: 415-338-7508; fax: 415-338-6243; e-mail: tfoschi@sfsu.edu).

G. Fields II was with the Romberg Tiburon Center for Environmental Studies, San Francisco State University, Tiburon, CA 94920-1205 USA. He is now with Engineering Management Concepts, Camarillo, CA 93010-8519 USA (e-mail: gary.fields@emc-inc.com).

H. Liu is with the Department of Computer Science & Engineering, Arizona State University, Tempe, AZ 85287-5406 USA (email: huan.liu@asu.edu).

Classifying *Egeria* in CIR airphotos by automated methods, however, presents a challenge due to a number of unfavorable conditions including variable imaging conditions, problems associated with water-related subjects, and other environmental changes. Digital analyses also indicate that subtle changes (e.g., in *Egeria* canopy density, film vignetting, or water turbidity) produce overlapping spectral response patterns. In addition, the spectral response patterns for *Egeria* do not separate well from those of other 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 pinkish and is confused with terrestrial vegetation. Fig. 1 illustrates these problems. Clearly, computer-assisted multispectral classification methods are problematic under these conditions, and visual/manual image interpretation and analysis are time-consuming and costly.

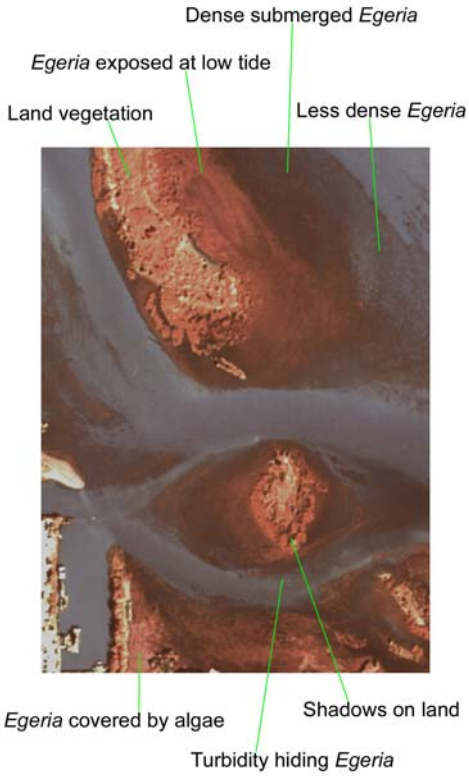


Fig. 1. Scan-digitized CIR aerial photography showing spectral variations in *Egeria* and lack of spectral separation between *Egeria* and extraneous classes.

In the next sections, we describe the data used for current

algorithm experiments, the methods most recently developed, and the results of our experiments. We conclude with a discussion of current problems and anticipated solutions to these problems.

## II. Data and Experiments

The experiments reported here used CIR aerial photographs of the Sacramento-San Joaquin Delta flown in October 2000 at 1:24,000 scale. The airphotos were scan-digitized and color separated to create 3-band (G, R, and near infrared) digital imagery at a nominal 1-meter spatial resolution. Subsets were selected to represent cases in which *Egeria* is readily interpretable to an image analyst yet problematic to classify by computerized methods due to significant spectral variability. The subsets were geometrically corrected. Algorithms were trained using one of the subsets and tested using the others.

Because prior experiments had shown that confusion between submerged *Egeria* and shadows on land was a significant source of error, land/water masks were created for each of the five training/testing sites and acted as control structures during processing.

Two approaches were developed and tested: one using active learning with class-specific ensembles (the data-mining method) and one using traditional multispectral classification followed by a rule-based knowledge engine. The data-mining method only processes data where the land/water masks indicate water is present. The knowledge-engine method processes all of the imagery and then uses the masks to reassign apparently misclassified pixels to more appropriate classes (e.g., Land Vegetation present in water area is reassigned to *Egeria*).

To keep the tests comparable, both approaches were bound by these constraints:

- Using the same and only one training image.
- Using the same land/water masks for focus and/or cleanup.
- Using the same four test sites for accuracy assessment.
- Using the same *Egeria* coverage maps (interpreted from the imagery by an independent researcher) to generate accuracy statistics.

### *A. Active Learning with Class-Specific Ensembles*

The data-mining algorithm consists of techniques for feature extraction, automated classification, and active learning with dual ensembles [2], [3]. This prototype system uses 13 features: three texture features per spectral band, three color features, and one edge feature. To calculate these features, the original image is divided into blocks of 8 x 8 pixels, each block overlapping each of its neighbors by four pixels. The texture features represent three types of textures: texture containing pure submerged *Egeria* patches, texture containing *Egeria* in shallower water, and texture containing *Egeria* adjacent to landmasses. The texture features are derived from the Discrete Cosine (DC) Transform matrix described in [4]. The DC value of each image block is then compared with the DC values for all textures and bands. The color features are generated by the YCbCr method [4]; Y, Cb, and Cr are the three features. The average intensity of each block is compared with these YCbCr components. Sobel's edge detection algorithm [5] is used to create the edge feature. To avoid identifying all edges, many of which are irrelevant, we use only edges that have *Egeria* color.

An active-learning algorithm using dual ensembles was employed to increase generalization and to decrease human expert involvement. A classification learning algorithm uses the training data to "learn" the mapping between the inputs and the output classes. It then predicts the classes for the test data. Active learning [6], [7] is a supervised learning algorithm that combines the results from multiple classification algorithms. An ensemble consists of a set of classifiers that generate a committee decision. In these experiments, we used class-specific ensembles.

We used all classification algorithms available in the machine-learning package WEKA [8] that can be applied to image data. These algorithms are efficient and powerful in learning, yet distinct in underlying principles. They include: Alternating Decision Trees, Decision Tables, Hyper Pipes, Kernel Density, and PRISM. An association-rule algorithm [9] was employed to find the best combination of classifiers associated with each class (*Egeria* and non-*Egeria*). In other words, this algorithm outputs the dual ensembles. Finally, the ensemble information was combined following the rule of majority. Further details of the method may be found in [3].

Preliminary experiments with data-mining methods revealed three identifiable sources of error: (1) significant spectral confusion between submerged *Egeria* and shadows on land; (2) misclassifications of spectrally obvious *Egeria* due to poor representation during training of the various *Egeria* types; and (3) problems with feature descriptions for small objects. Consequently, the experiments reported here include three components not found in the earlier papers: (1) land/water masks to function as control structures limiting processing at various stages to only data where the masks indicate water is present; (2) a larger training site to represent more diverse types of *Egeria*; and (3) aerial photography scan-digitized to a finer resolution (1-meter rather than 2-meter) to allow the 8 x 8 pixel blocks to better describe smaller objects.

### *B. Multispectral Classifier with Knowledge Engine*

The second method reported here uses a traditional multispectral classifier—namely, the Euclidean minimum distance classifier—followed by a rule-based knowledge engine. By itself, the minimum distance classifier performs poorly since *Egeria* is spectrally confused with other subjects, particularly land vegetation (at low tide conditions) and shadows on land (when well submerged). In addition, as mentioned earlier, variations in imaging conditions and environmental changes may cause subtle changes that affect class separability. Such variations would normally require retraining for changing conditions.

To address these problems, a knowledge engine was written within ERDAS Imagine© using the Knowledge Engineer utility [10]. This utility allows the creation of a decision tree for rule-based classification by identifying variables, rules,

and output classes of interest. From the initial classes output by the minimum distance classifier, the knowledge engine creates a five-class map containing the following classes: *Egeria densa*, Water, Land Vegetation, Shadow/Inland Water, and Soil/Urban. The knowledge engine uses the land/water masks to reassign apparently misclassified pixels to more appropriate classes. This process allows spectrally identical areas within the land and water to be classified differently. For example:

- Land Vegetation present in water is reassigned to *Egeria*.
- Dark *Egeria* on land is reassigned to Shadow/Inland Water.

Since the goal is to monitor and map *Egeria*, this coarse classification is satisfactory. For accuracy assessment, all maps were recoded to *Egeria* and non-*Egeria*.

## III. Results, Discussion, and Conclusions

To evaluate the test results, three criteria were used: precision, recall, and the F measure. They are defined as:

- Precision (P): the fraction of the relevant information over the retrieved information or the fraction of the classification that is correctly identified; TP/(TP + FP) where TP = true positives and FP = false positives.
- Recall (R): the fraction of the relevant information that is retrieved over all relevant information; TP/(TP + FN) where FN = false negatives.
- F measure (F): an overall combination of P and R criteria; (2*P*R)/(P + R).

Table I summarizes the evaluation results, and Fig. 2 illustrates the maps produced. In general, the data-mining method excelled in recall while the knowledge-engine method excelled in precision. Overall, the data-mining method yielded a somewhat larger F measure. However, visual inspection of the output classification maps compared to the *Egeria* coverage maps, in Fig. 2, reveals that the data-mining method produced more obvious classification anomalies.

Whereas these results are satisfactory for preliminary trials, more study is needed. Two apparent problems need to be addressed: (1) geometric correction and creation of land/water masks were the most labor-intensive and time-consuming procedures, and reuse of the masks requires very meticulous geometric correction of any new imagery; and (2) the higher (1-meter) resolution has not entirely improved the feature descriptions for small objects and is probably the cause of many misclassifications. As trials proceed, we intend to refine processing in two ways: (1) by finding an automated means of creating the land/water masks that will not require meticulous geometric correction and (2) by replacing the blocks of 8 x 8 pixels with shapes more descriptive of the spatial information in the imagery.

TABLE I
Summary of Results

| **Data Mining** | **P** | **R** | **F** |
|---|---|---|---|
| Test site 1 | 0.76 | 0.85 | 0.80 |
| Test site 2 | 0.61 | 0.77 | 0.68 |
| Test site 3 | 0.23 | 0.82 | 0.36 |
| Test site 4 | 0.68 | 0.68 | 0.68 |
| Test Average | 0.57 | 0.78 | 0.63 |
| Training site | 0.88 | 0.94 | 0.91 |
| Overall Average | 0.632 | 0.812 | 0.686 |
| | | | |
| **Knowledge Engine** | **P** | **R** | **F** |
| Test site 1 | 0.85 | 0.80 | 0.82 |
| Test site 2 | 0.55 | 0.45 | 0.50 |
| Test site 3 | 0.40 | 0.65 | 0.50 |
| Test site 4 | 0.63 | 0.54 | 0.58 |
| Test Average | 0.608 | 0.61 | 0.60 |
| Training site | 0.96 | 0.86 | 0.91 |
| Overall Average | 0.678 | 0.66 | 0.662 |

## Acknowledgments

Deepak Kolippakkam, Amit Mandvikar, and Jigar Mody at Arizona State University and Yukari Matsumoto and Mami Odaya at San Francisco State University have contributed to this research.

## References

[1] P. G. Foschi, “Egeria densa acreage and percent coverage in the Sacramento-San Joaquin Delta,” Dept. of Boating and Waterways, Sacramento, CA, *Egeria densa Control Program*, vol. 2, Rep. 4, 2000.

[2] P. G. Foschi, D. Kolippakkam, and H. Liu, "Feature selection for image data via learning," *Proc. Int. Workshop on Intelligent Multimedia Computing and Networking*, *Joint Conf. on Information Sciences (JCIS 2003-IMMCN 2003)*, Cary, NC, Sep. 2003, pp 1299-1302.

[3] H. Liu, A. Mandvikar, P.G. Foschi, and K. Torkkola, “Active learning with ensembles for image classification,” *Int. Joint Conf. on Artificial Intelligence (IJCAI2003)*, Acapulco, Mexico, Aug. 2003, pp. 1435-1436.

[4] R. Gonzalez and R. Woods, *Digital Image Processing*. Addison-Wesley Publication Co., 1992.

[5] *MATLAB Image Processing Toolbox User's Guide*, Vers. 3, Mathworks, Inc., 2001. Available: http://www.mathworks.com.

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

[7] N. Roy and A. McCallum, “Toward optimal active learning through sampling estimation of error reduction,” *Proc. 18th Int. Conf. on Machine Learning*, 2001.

[8] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers, 2000.

[9] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” *Proc. 20th Int. Conf. on Very Large Databases*, 1994, pp. 487-499.

[10] *ERDAS Field Guide*, 6th ed., ERDAS LLC, Atlanta, GA, 2002, pp. 237-240.

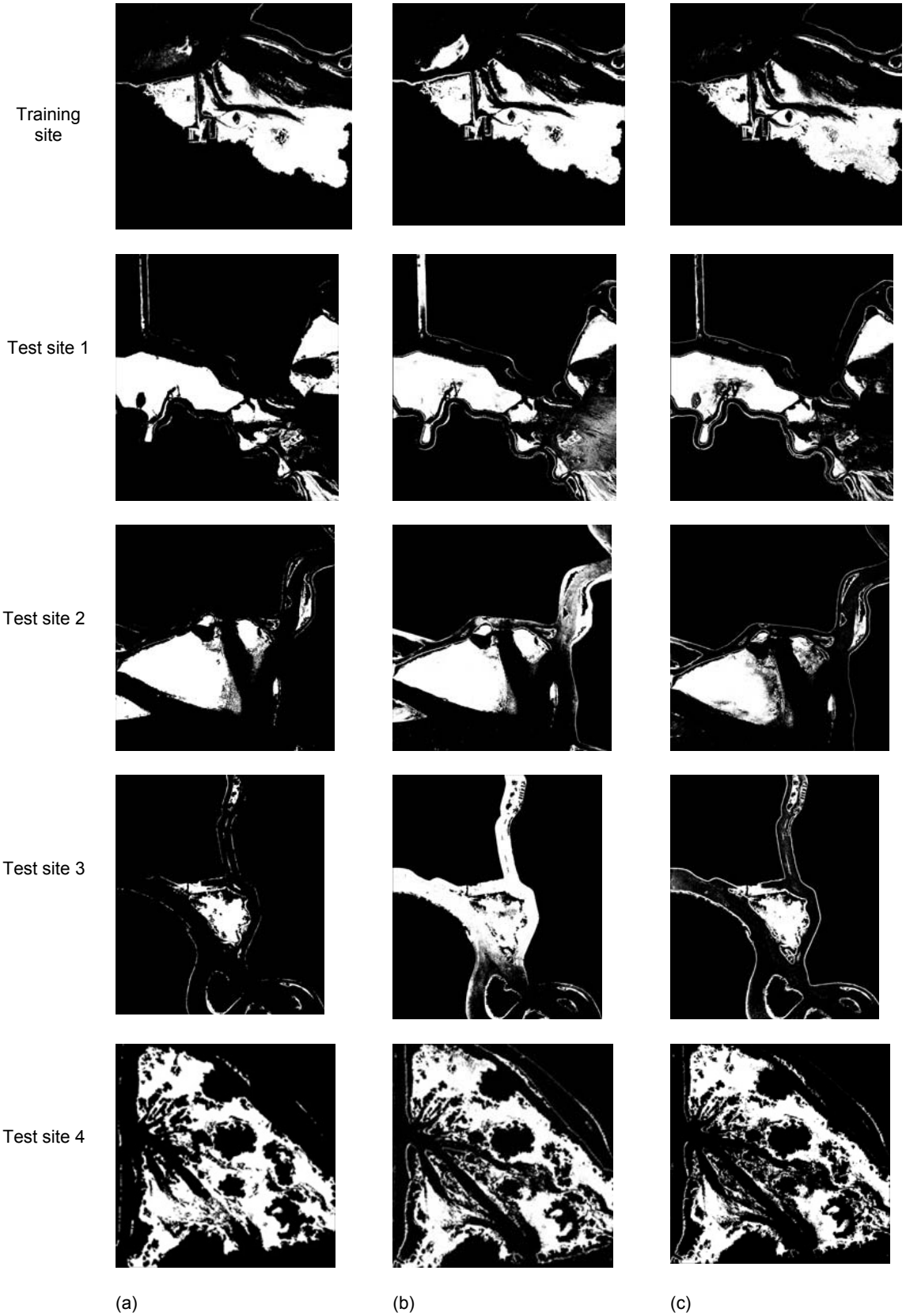


Fig. 2. *Egeria* maps for training and test sites with white indicating locations of *Egeria*: (a) *Egeria* coverage maps interpreted from imagery by independent researcher, (b) output maps from data-mining method, (c) output maps from knowledge-engine method. Relative image sizes have been altered to fit on page.