Knowledge Discovery in an Object-Oriented Oceanographic Database System

Final Technical Report
September 1, 1996 - December 31, 1999

ONR EPSCoR Grant N00014-96-1-1276
PR Number 96PR07924-00

Principal Investigators:
Julia Hodges, Susan Bridges

Student Contributors to Report:
George Brannon Smith, Sean Taylor, Andrew Watkins, Bruce Wooley,

Department of Computer Science
Mississippi State University
Box 9637
Mississippi State, MS 39762-9637
<hodges,bridges>@cs.msstate.edu

Abstract

The rate at which scientific data is collected today has overwhelmed the ability of scientists to store and analyze the data. This report describes the results of a three year effort in the development of a knowledge discovery system for use by oceanographers at the Naval Oceanographic Office (NAVOCEANO) at the Stennis Space Center in the identification of provinces of interest in the ocean floor from acoustic imagery. The system is composed of a knowledge discovery component built to interact with a database system currently in use at Stennis Space Center. The knowledge discovery system applies machine learning techniques to features extracted from sonar images to identify provinces of the ocean floor based on visual texture. This requires that the images be segmented into regions of homogeneous texture using a region-growing technique, that features describing the texture of these regions be extracted, that machine learning techniques be applied to classify the regions, that classified images be constructed for visualizing the results, and that the classified images be combined and geo-referenced using a mosaic procedure. NAVOCEANO is currently supporting efforts to integrate the software developed from this project with their image analysis system.

1. Introduction

Fayyad, Piatetsky-Shapiro, and Smyth (1996) describe the knowledge discovery problem as the need to develop "a new generation of techniques and tools with the ability to intelligently and automatically assist humans in analyzing the mountains of data for nuggets of useful information." The goal of this project is to develop a knowledge discovery system consisting of an oceanographic database and the tools needed to support the automated extraction of

information from the database. Such a system will aid oceanographers in the analysis of complex oceanographic data sets that are too large to be analyzed manually. This work involves the development of knowledge discovery tools that aid oceanographers in their data analysis tasks and mechanisms for storing the discovered knowledge in an existing oceanographic database. This work is being done in collaboration with scientists at the Naval Oceanographic Office at the Stennis Space Center. The scientists at NAVOCEANO at Stennis who have been our primary collaborators are:

Martha Head	Oceanographer
Steve Lingsch	Supervisory Oceanographer (Division Head NAVOCEANO N95)
Dr. Peggy Haeger	Oceanographer (Specialty in Acoustics and Hydrography) N95
Mollie Haynes	Geologist N531
Chris Robinson	Senior Scientific Applications Computer Scientist N95
Chuck Martin.	Computer Scientist N95.

Geologists at NAVOCEANO currently manually classify regions of the ocean floor from very high resolution sonar images. The work is very tedious and error-prone. The scientists at NAVOCEANO are interested in a knowledge discovery system that can aid in this "provincing" process and that can be used as the images are collected at sea. We are working with them to develop a prototype knowledge discovery system that uses acoustic imagery and other data to province the ocean floor.

2. The Oceanographic Database

Although we initially established the design and implementation of the object-oriented oceanographic database as the first major task in this project, this approach was abandoned for two major reasons. First, the delay in ordering the object-oriented DBMS tool that we reported in the first year caused us to concentrate on the knowledge discovery aspects of the project. This delay, along with the encouraging results of our first year's work on the knowledge discovery component, have caused us to focus our efforts on the KD component. Second, we, in consultation with the NAVOCEANO scientists, determined that it would be better to extend the geographical information system already in use at NAVOCEANO to handle the new data types rather than design a new database system. We have obtained and installed this system at our site and have extended it to accommodate meta data and discovered knowledge.

3. The Knowledge Discovery System

The overall goal of the knowledge discovery system we are building is to aid the scientists at NAVOCEANO in the analysis of large sets of complex oceanographic data. More specifically, the scientists at NAVOCEANO wish to have a system that can use acoustic imagery and other data to province the ocean floor. Geologists currently do this job manually. Our system extracts texture statistics from acoustic images (collected from a 100 kHz Chirp Side-Scan Sonar using a Data Sonics SIS1000) and applies clustering algorithms to identify classes of textures. In our first year's report, we described:

1) the knowledge discovery process that we are using including both texel-based and region-based extraction of texture statistics,

2) sets of experiments that we conducted to select a parameters such as texel size and number of classes for the texel-based approach,
3) the results of feature selection experiments,
4) the results of experiments comparing the texel-based approach and the region-growing approach, and
5) techniques used to visualize the results of the classification.

Because the scientists at NAVOCEANO found the results from the region-growing approach much more satisfactory than those from the texel-based approach, we concentrated our second-year's effort on refining and extending this approach. In the second year of our research, our major focus was the application of the knowledge discovery process based on region-growing to a much larger number of images. We conducted experiments to parameterize the region-growing algorithm to make it applicable to a wider variety of images. We also compared clustering results based on a Bayesian classifier (Autoclass) and a decision-tree clustering algorithm (Cobweb), and investigated methods for parallelizing the knowledge discovery process. Details of the work from the first two years was provided in the annual reports. In this final report we will describe the current status of the knowledge discovery process, we will discuss techniques that we have investigated for classification of sand wave regions, and we report on-going related research efforts. We also briefly describe our development efforts that have been supported by NAVOCEANO to integrate our software into their UNISIPS image processing system.

3.1 Current Status of the Knowledge Discovery Process

We have developed a knowledge discovery system that classifies provinces of the ocean floor based on visual texture of acoustic images (Hodges et al. 1997; Karpovich 1998; Wooley and Smith 1998; Bridges et al. 1999). The current system takes a set of acoustic images as input and "clusters" regions of the images into classes based on similarity of visual texture of the scan-line images. The classified scan-line images are then geo-referenced and combined using a mosaic process as shown in Figure 1. Figure 2 gives a more detailed view of the provincing process. The steps in this process are outlined below:

1. Preprocessing. Many of the acoustic images have a gray-scale range that is very limited. In addition, the method of texture analysis that we use requires that the images be quantized to a small number of gray-scale values. We use a Kohonen self-organizing map to reduce the number of pixel values (quantization) while ensuring that the range used adequately represents the variability found in the images (stretching). Figure 3 shows an acoustic image and the same image after preprocessing.

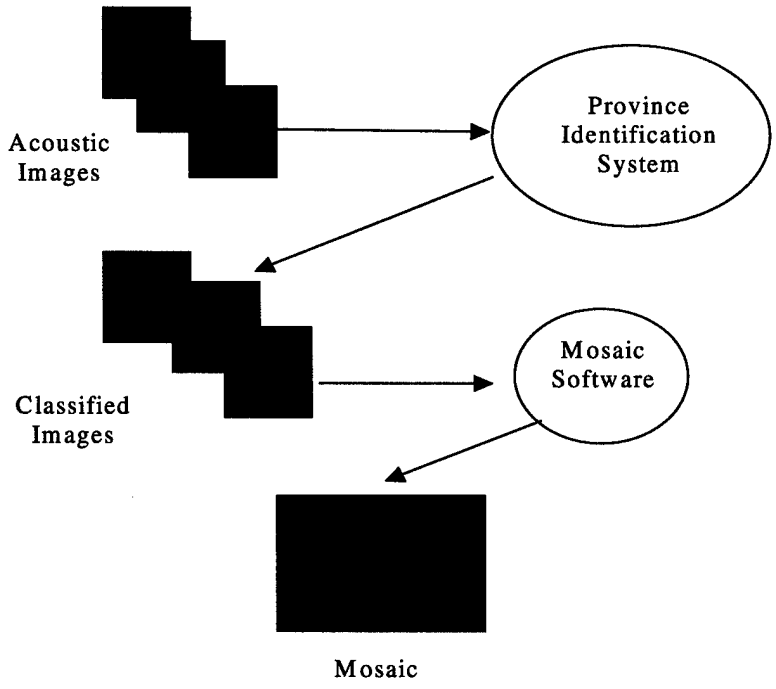

Figure 1. Knowledge Discovery Process for Provincing a Set of Acoustic Images

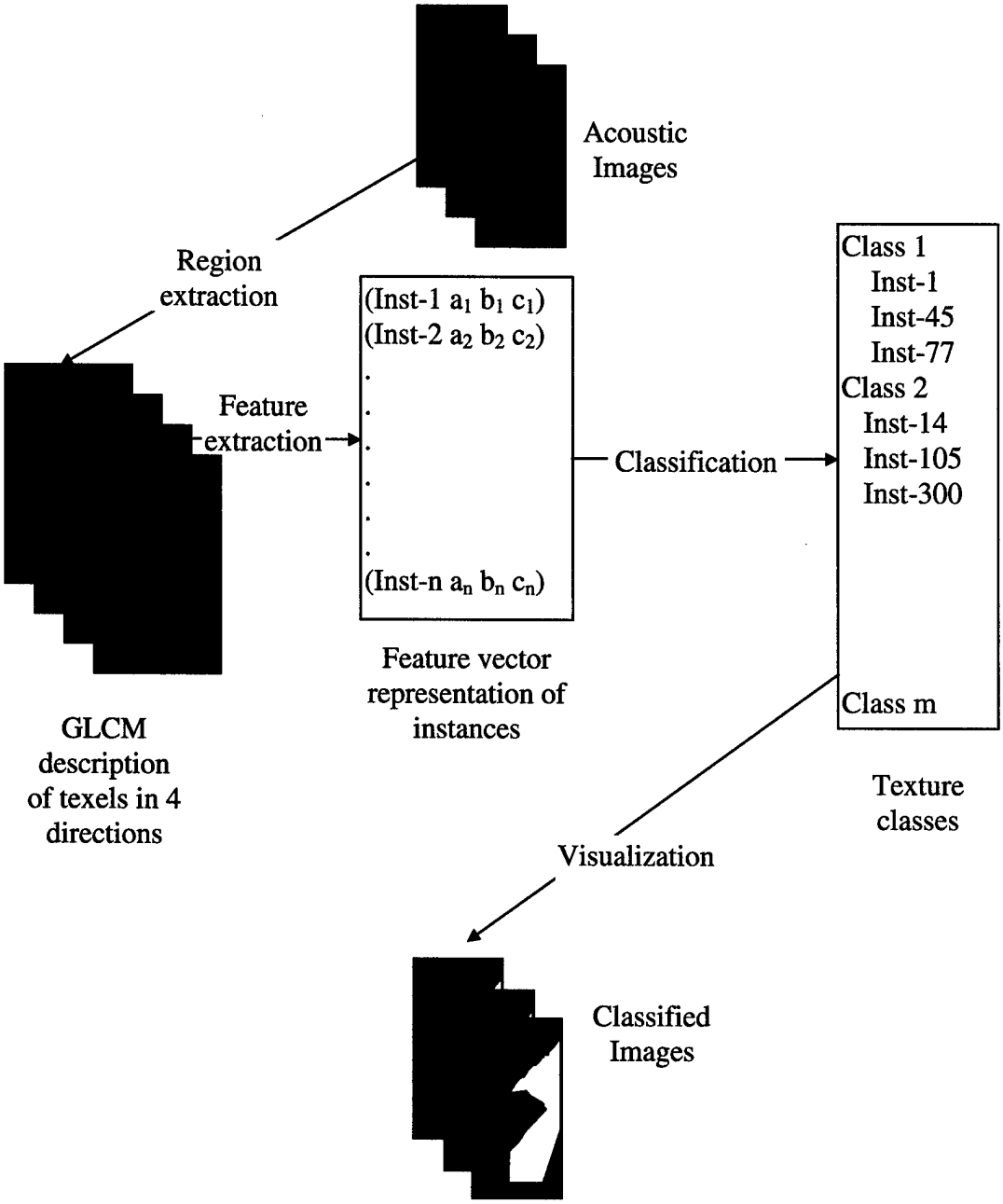

Figure 2. Detailed View of the Provincing Process.

2. Region growing. A region growing process is used to identify irregularly shaped images of homogeneous texture. Regions are "grown" by analyzing the homogeneity of small blocks called "cells." If a cell is not sufficiently homogeneous, it is labeled a "boundary cell" and will not be used in the clustering process. Otherwise the cell is either combined with adjacent cells to grow an existing region or it is used to start a new region. At the completion of the process, each region thus identified will be considered one object to be classified (each referred to as an instance). Figure 4 shows the results of region-growing for a portion of one example image. This region-growing process has two parameters (a boundary parameter and a region parameter) whose values must be adjusted based on properties of the images and preferences of the user for the granularity of the classification. We are currently investigating the use of Bayesian networks to aid in the selection of good sets of parameters (see On-Going Work).

3. Gray-Level Co-occurrence Matrices. A gray-level co-occurrence matrix is computed for each region for each of 4 different directions. These matrices provide information about the co-occurrence of different pixel values. We typically quantize our images to 16 gray level values using a Kohonen map and thus use 16×16 GLCMs. To date, we have only computed GLCMs for a distance of 1. We plan to investigate other distances as part of our work to distinguish different roughness categories.

4. Second order textural statistics (feature extraction). There are a number of second order statistics that can be computed from the gray-level co-occurrence matrices for the regions (Haralick 1973; Reed and Hussong 1989). These statistics provide quantitative summaries of textural properties of the images. The statistics that we are using are: mean pixel value, standard deviation of pixel values, angular second moment, correlation, entropy, contrast, and angular inverse difference moment. Each of these values is referred to as a feature. Each instance (region) is described by a set of feature values called a feature vector.

5. Clustering. A clustering algorithm (Autoclass) is used to group the regions based on the similarity of the textural statistics for each region. The Autoclass clustering package that we are using is distributed by NASA Ames and implements a Bayesian clustering algorithm. Figure 5 shows a visualization of the clustering results for one image. We have performed experiments comparing the performance of several different clustering algorithms and thus far have obtained the best results using AutoClass. Clustering is a two-phase process. In the first phase, a subset of the instances are used to "train" a classifier. We have found that we obtain better results if we only use instances representing relatively large regions for the training process that correspond roughly to the extent of regions of interest to the oceanographers. During this phase, AutoClass learns probabilistic descriptions of the classes it identifies in the images. After training is complete, these descriptions are used to classify all instances (including those excluded because they were too small and those representing "boundary cells").

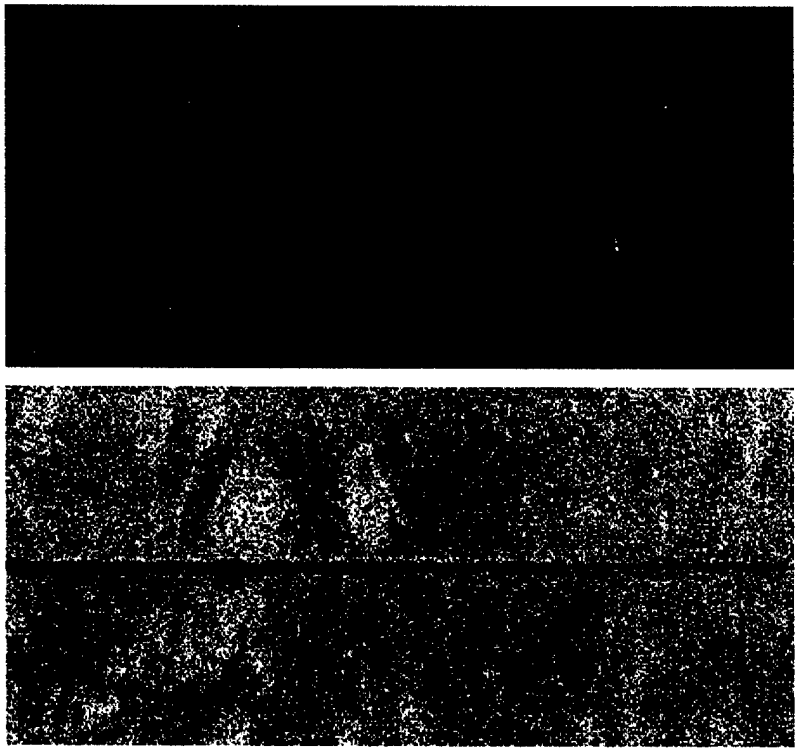

Figure 3. Original acoustic image and image after preprocessing.

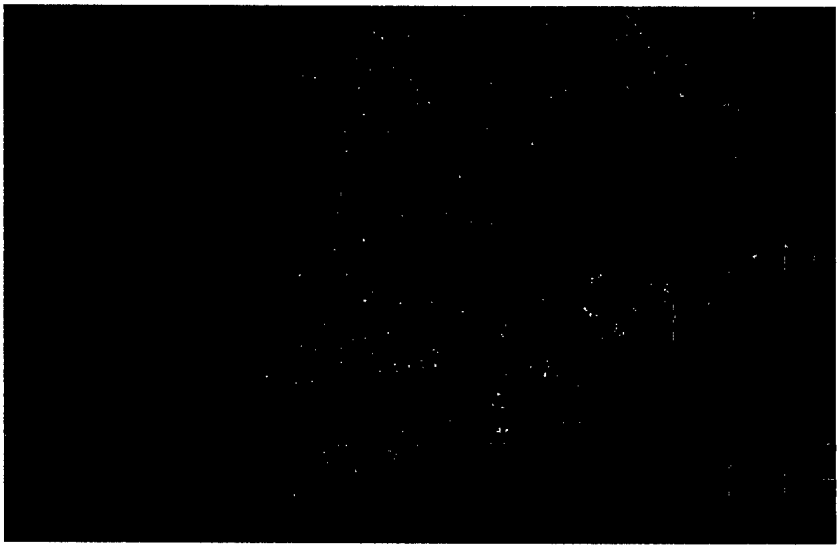

Figure 4. Visualization of Region-Growing Results for a Portion of One Image.

6. Visualization. When clustering is complete, the classified instances must be used to reconstruct classified images. This is a complex process that requires mapping from classified instances to regions, from regions to cells, and finally from cells to pixel values. Input for the process are the classified instances, the original image, and a file that provides necessary mapping information. A gray-level is assigned to each class (with gray levels distributed over the 0-255 range to assist visualization) and a gray-level classified image is reconstructed. The individual classified gray-level images are then ready to be geo-referenced and combined into a mosaic image. We have also developed an additional utility to produce color images from the gray-level images to facilitate visualization. Input for this utility is the gray-level image, the number of classes in the image, and a color map. This tool can also be used to produce color images of the mosaic images described below.

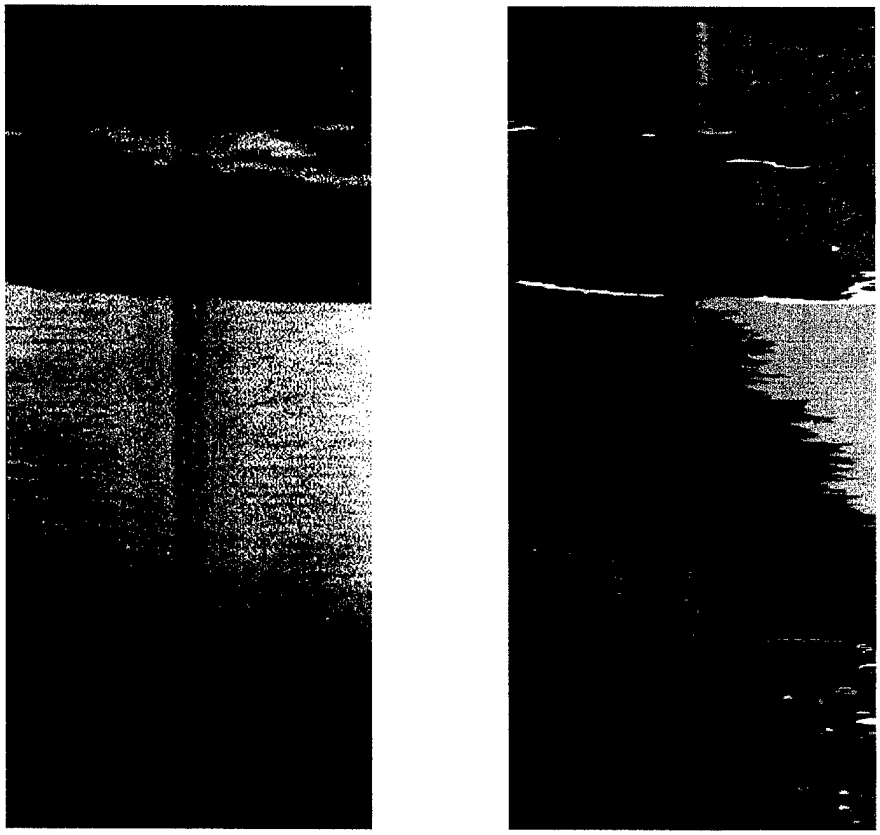

Figure 5. A Single Acoustic Image Before and After Clustering.

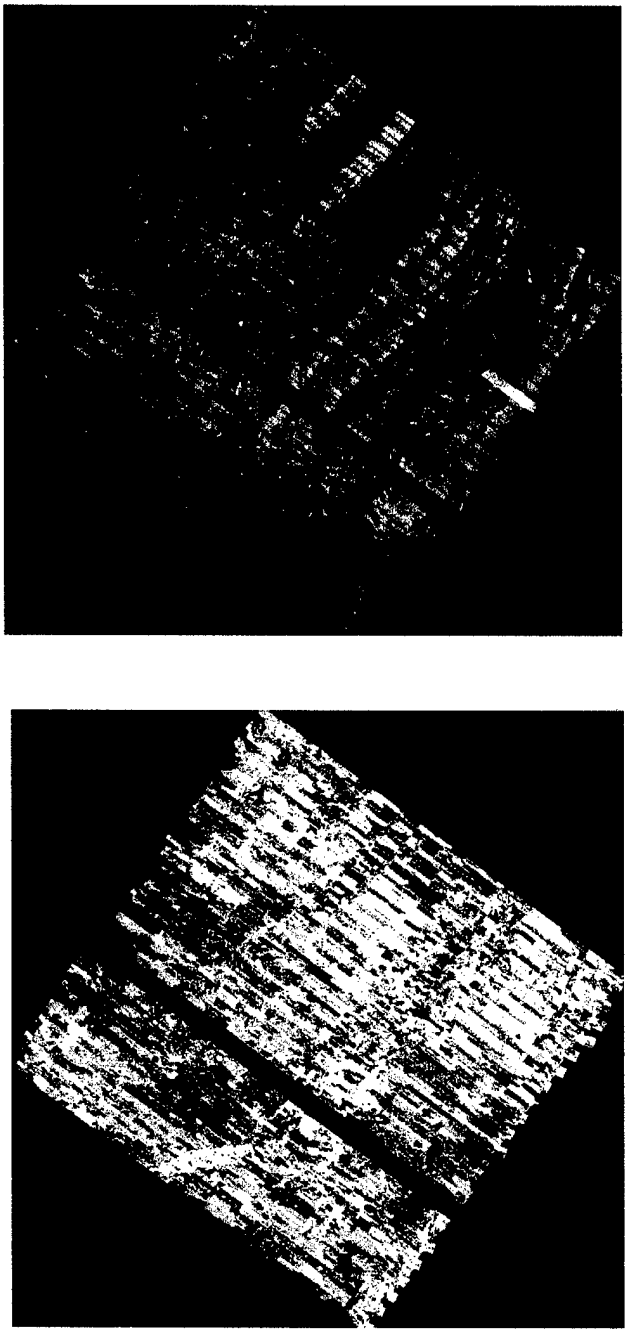

Figure 6. Mosaic of Original Images and Classified Images

After all of the images in a set have been classified they are geo-referenced and combined using a mosaic process. We have adapted the mosaic process used by NAVOCEANO for use with the classified images. Figure 6 gives the mosaic results for a set of acoustic images and the classification of these images. We are currently investigating methods to provide a color

mapping that will cause texturally similar classes to be assigned similar colors to make the color assignment more visually meaningful.

4. Identification of Sand Wave Regions

The knowledge discovery system described above finds classes of the ocean floor of visually similar texture. It makes no attempt to identify the different classes. An aspect of interest to our collaborating geologists at NAVOCEANO is the identification of sand waves in the acoustic images. Our previous work focused on the identification of distinct textural regions (provinces) in acoustic images through the use of unsupervised learning (clustering) techniques. While this work has had success in distinguishing differing textural regions in an image, the geologists must still manually classify each region as sand wave or non-sand wave—a very time-consuming process. Our current work is to automate this classification process as well.

We compare the performances of several machine learning techniques for this task. Particularly, we present results from the use of various styles of committee machines. While we have seen some improvement in our classification system through the use of the various learning techniques, we conclude that the current set of features do not accurately reflect the necessary domain knowledge to make this classification system completely successful.

4.1 The Texel Marker Tool

So far we have used the same features for our instances for this supervised learning task that we describe above for the unsupervised learning system. The supervised learning task requires that a set of training data be developed which consists of classified instances. In this case, we are dealing with a two-class problem where each instance should be classified as sand wave or non-sand wave. We have developed a tool (written in Java) for users to use to build the training and test sets for supervised learning. A screen-shot of the tool is shown in Figure 7. This tool displays an image with a grid overlay of a size specified by the user. The user can then use the mouse to select square regions of the image (called texels) and the user chooses a label for the selected region sand wave or non-sand wave. The GLCM features described previously are computed for the selected texels and these feature values along with the labels are output to a file for use as input for the supervised learning system.

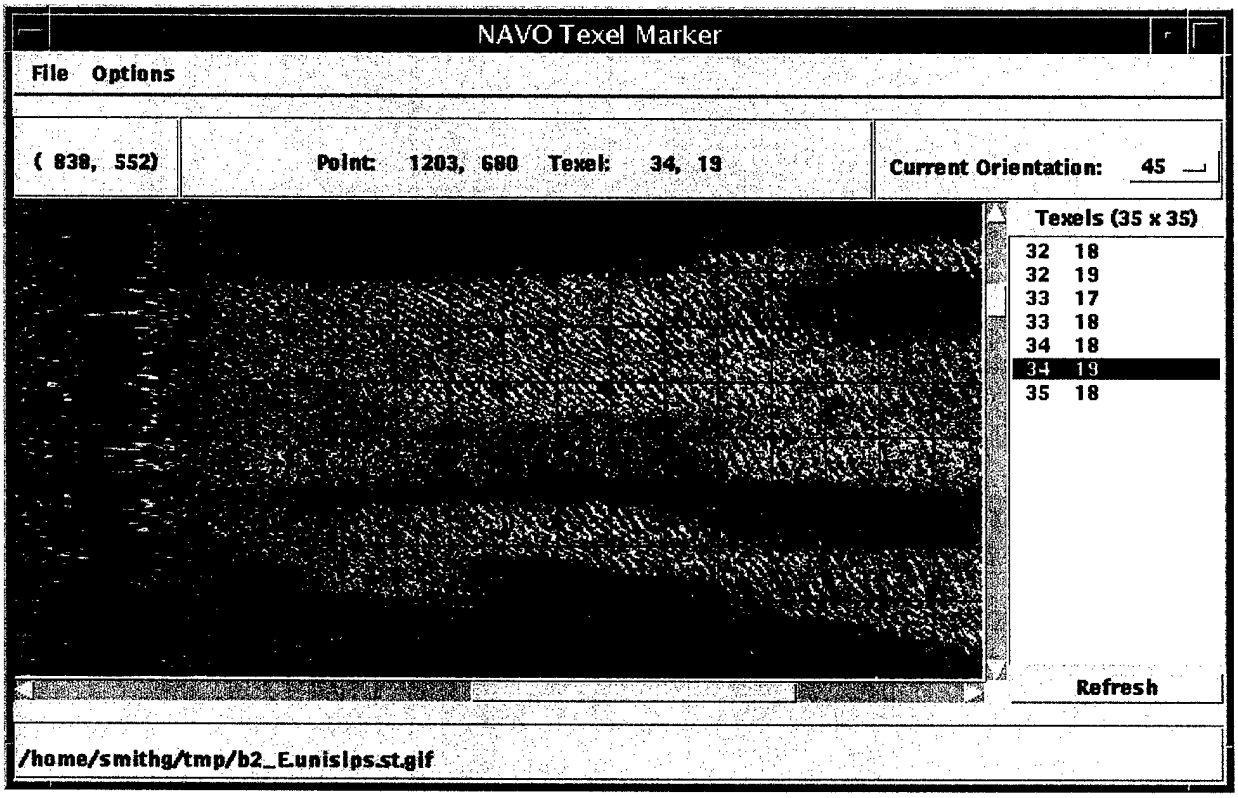

Figure 7. Graphical User Interface for the Tool for Building a Training and Test Sets for Supervised Learning.

One important point to note in this method of feature extraction is that the GLCM is only calculated for a distance of 1 for the texels. This greatly limits the amount of contextual information encapsulated within the features extracted. Through observation of human experts classifying sand waves, it became apparent that humans use much more contextual information in making their decisions than our current learning systems do.

4.2 Experimental Data Set

The data set used for the experiments presented here consists of 944 vectors, each with 22 features extracted from eight sonar images and a 1 or 0 classification for sand wave or non-sand wave, respectively. For the supervised learning experiments, this set was divided into a training set with 796 instances and a test set with 198 instances. This data was normalized to the range of -1 to 1.

4.3. Experiments, Results, and Analysis

4.3.1 Initial Explorations

For an initial exploration of the performance of traditional supervised learning techniques on this data set, Quinlan's C4.5 decision tree package (1992) was used. Unsurprisingly, C4.5 did

not perform extremely well on this data set. It was hypothesized that since the training instances consist of only continuous valued features, the general nature of a decision tree makes it unsuitable for this task. Therefore, we next explored the use of back-propagation neural networks using the NevProp system (Goodman 1998) as well as a C++ version of the back-propagation algorithm implemented by Andrew Watkins. The performance of back-propagation on the data set was better than C4.5; however, it was hoped that the accuracy could be improved even more. To explore the possibility that the high dimensionality of the feature vector was significantly affecting these various classifiers and the possibility that not all of the features were necessary to truly represent what constitutes a sand wave, several feature selection algorithms were applied to the data. These algorithms included RELIEF, SFS, SBS (Kira and Rendell 1992), and Branch and Bound (Narendra and Fukunaga 1977). While feature selection did reduce the dimensionality of the feature vector, using these reduced features did not significantly improve the accuracy of the classifier. Table 1 presents the best test set accuracy results for each of these initial classification attempts. All results were averaged over three runs.

Table 1. Test Set Accuracy for Initial Classification Attempts

Classifier Type	Decision Tree	Back-Propagation (BP)	BP with Feature Selection
Test Set Accuracy	67.2%	72.03%	72.07%

As these results indicate, each new variation we tried did improve accuracy; however these improved classifiers were still not as accurate as we had hoped.

4.3.2 Self-Organizing Maps

In an attempt to better understand the characteristics of the data set, we next turned to self-organizing maps (SOM). When meeting the experts at NAVOCEANO, we noticed that they quite often disagreed with each other about whether a given texel was a sand wave or not. With this in mind, the initial motivation for the SOM experiments was an attempt to confirm the classifications of the data made by experts. It was hoped that the union of certain clusters identified by a SOM would correspond with the classifications made by the expert. Klimek's (1998) SOM implementation was used to explore this possibility. However, it was found that the clusters did not come close to purely corresponding to the expert's classes. In fact, an individual cluster was either split evenly—responding to instances of both classes equally—or it would respond around 80% of the time to the sand wave class. None of the clusters responded strongly to the non-sand wave class. Since Klimek's implementation was slightly limited in the adjustable parameters, the more robust SOM_PAK (Kohonen, et al 1996) was explored. While this package is useful in visualizing the data and the effects various changes in the parameters have on the data, it does not allow for predictions the way that Klimek's package does. Despite this limitation, the use of SOM_PAK proved to be useful in our investigation of the data set.

We tried numerous combinations of the adjustable parameters available in SOM_PAK and found that most of the variation produced no noticeable differences. Since Kohonen (1996) asserts that a lower quantization error represents a "better" SOM, figures 8 and 9 present different visualizations of the SOM produced when using 0.5 as the initial learning rate of the

first part of the training and 0.09 as the initial learning rate of the second part of the training as using these parameter resulted in the lowest quantization error. Figure 8 gives a view of the labeled 15×10 matrix and gives a good indication of the distances between the different nodes in the system. Figure 9 gives a slightly different view of the SOM and provides a visualization of the different clusters of labels.

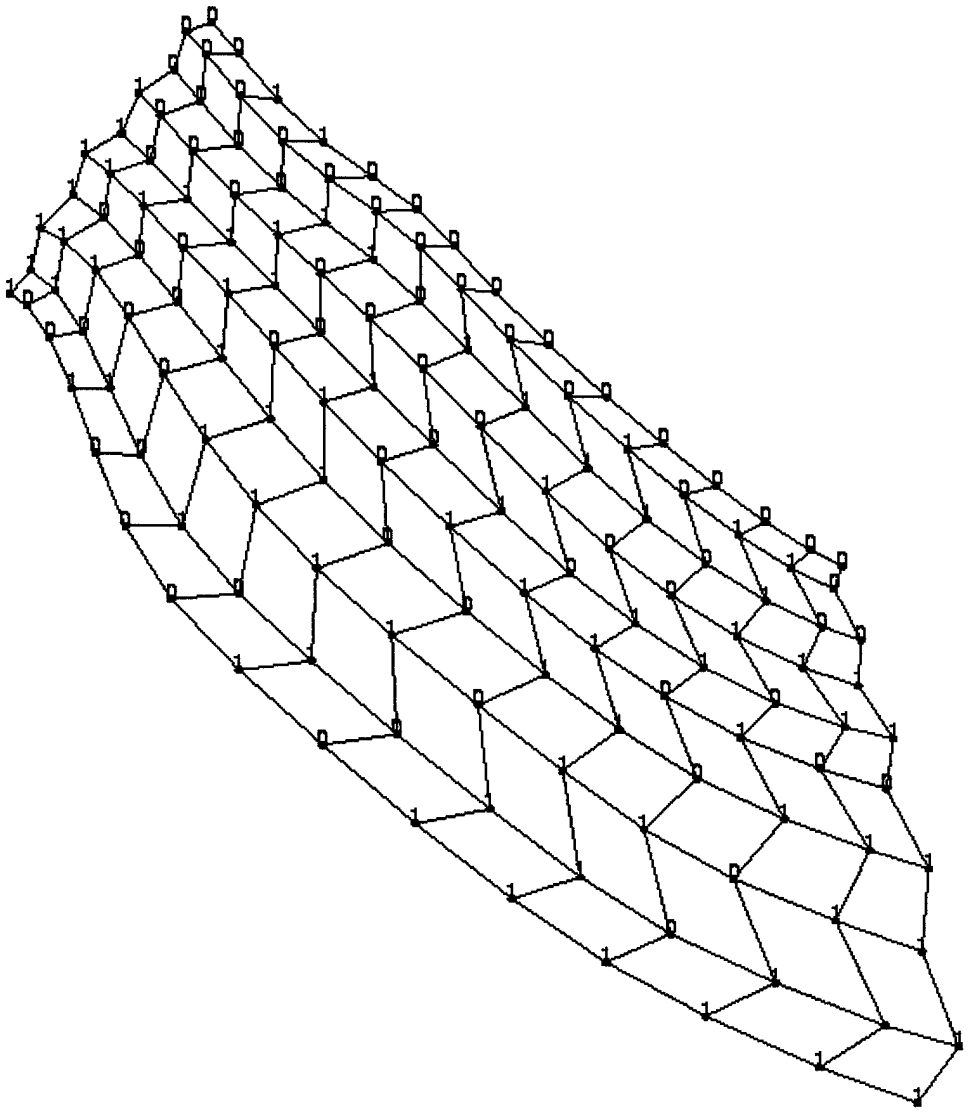

Figure 8. SOM Obtained from Using 0.5/0.09 as Learning Parameters.

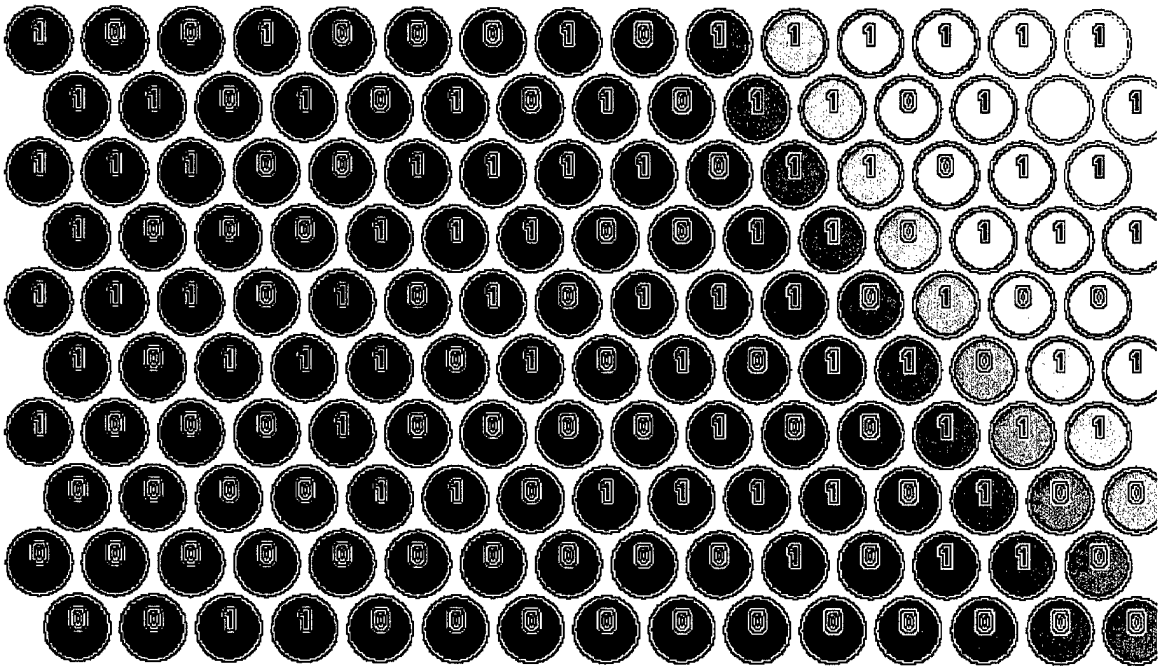

Figure 9. A Planar View of the SOM.

As both of these visualizations indicate, there are definite clusters of the two different classes. However, not only are they close together, there are also several outliers which make classification extremely difficult. Nevertheless, this phase of the investigation provided encouragement for the classification system development process. Since there are distinct clusters of similarly classified data items within the data set, we decided that it should be possible to design a classifier to correctly identify sand waves in sonar images.

4.3.3 Committee Machines

The bulk of the experiments performed to date have involved the investigation of various ways to combine base classifiers for the creation of an improved overall classifier. The basic committee machine architecture is given in Figure 3.

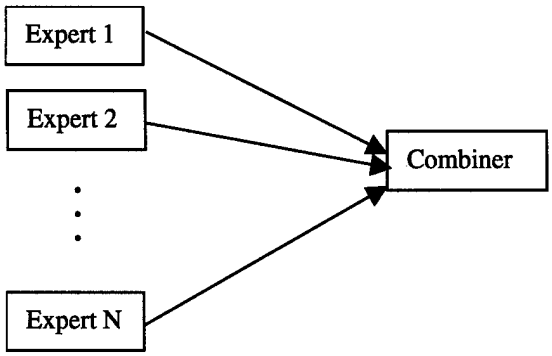

Figure 10. Basic Committee Machine Architecture

Five different methods of combining base classifiers were attempted: a simple ensemble committee machine (Haykin 1999), bagging (Breiman 1996), a mixture of the original input vector and the outputs from the base experts, the use of only the outputs of the base experts, and the use of two randomly chosen base experts as a filter for the training set for the combiner (Boggess 2000). For the simple ensemble and bagging experiments, the combiner in Figure 10 either used an averaging of the experts' outputs or employed a voting scheme with each of the experts' votes being equally weighted. For the other three styles of committee machines, the combiner was a newly trained back-propagation neural network. Except for the bagging experiments, the base 1-N experts in Figure 10 were all back-propagation neural networks trained on the same data set using different random initializations of the weights. For the bagging experiments, the training set was chosen randomly with replacement. For the base experts, the effect of varying the number of hidden nodes on the overall combined classification was investigated. The numbers of hidden nodes used in the fully connected network were 8, 10, 12, 14, and 16. These architectures were chosen so as to provide a range of hidden nodes from roughly one-third the size of the input vector to roughly two-thirds this size. For each of these networks, a learning rate of 0.01 and a momentum factor of 0.5 were used. Each network was allowed to train for 1000 epochs. Each of the networks had only a single layer of hidden nodes which were fully connected to the input vector and output node. Each back-propagation network architecture was run ten times with different initial conditions and, in the case of bagging, with different training set distributions to produce the individual experts for the committee machines.

4.3.3.1 Simple Ensemble and Bagging

Table 2 gives the best results obtained for the bagging and simple ensemble experiments. The first column indicates the number of hidden nodes in the ten experts' architecture. Multiple values indicate a combination of base experts. For example, the line with the ID value 10_14 indicates a combination of 20 experts, ten of which have 10 hidden nodes and ten of which have 14 hidden nodes. Columns two and three give the correct classification rate for the test set using the averaging scheme for the simple ensemble and bagging, respectively. The final two columns give the correct classification rate for the test set using voting types of combiners. All results represent the average of three runs.

Table 2. Results For Simple Ensemble and Bagging Style Committee Machines

ID	ENS AVE	BAG AVE	ENS VOT	BAG VOT
8	**72.73%**	72.73%	**73.91%**	72.90%
12	72.22%	73.06%	73.74%	**74.24%**
14	72.22%	**73.40%**	73.23%	73.06%
10_14	**72.73%**	72.90%	72.73%	72.39%
12_16	71.72%	**73.40%**	72.56%	73.06%
ID=Number of hidden nodes in the base experts				
ENS=Simple Ensemble; BAG=Bagging				
AVE=Averaging; VOT=Voting				

Table 2 reveals several interesting results. The most encouraging, from the view-point of the classification system development, is that both simple ensemble style committee machines and bagging style committee machines produced better results than had previously been obtained through other methods. In looking at the comparison of these two types of committee machines, bagging tends to outperform the simple ensemble. This could be due to the fact that, by randomly choosing the training set distribution, each expert is forced to focus on different areas of the input space. This new focus allows the individual experts to be more specialized, and then the combiner reduces this specialized bias by averaging or voting without losing all of the benefits of the base expert's specialization. There are two interesting general trends indicated in Table 2 that should be noted. First, the voting style of combining tends to perform better. It appears that a more accurate classifier is produced by forcing each expert to make a 0 or 1 decision on the classification rather than a classification somewhere in this 0-1 range and then choosing the majority classification of these experts. Secondly, having base experts trained on different architectures does not seem to produce better results. In fact, if all of the experts are trained with the same number of hidden nodes, then the resulting combined results are more accurate. One possible reason for this trend is that, in the current experiments, use of differently configured experts also entailed the use of more experts. This produces a greater degree of bias that the combiner must attempt to reduce. We have not investigated keeping the number of base experts consistent (i.e., at 10) but allowing each (or some fraction) of these experts to have different hidden node architectures.

4.3.3.2 Other Committee Machine Schemes

In this section we discuss three different, yet related, committee machine schemes. For the first scheme the output of the ten base experts was combined with the original input vector to create a new input vector of 32 features. Through this approach, the neural network combiner has the opportunity to learn which of the experts to pay attention to for different input patterns. The second scheme is similar. For this style of committee machines, the input for the network combiner was only the output of the ten base experts. The rationale for this approach was to ascertain if the experts' classifications alone contained enough information for the combiner to successfully differentiate between the two classes in the sonar images. The final approach used the classification of two randomly chosen base experts as a filter on the training set. The combiner was trained only on input instances where the two experts disagreed. We hoped that

this would force the combiner to pay specific attention to input patterns that were difficult to classify. Unfortunately, the resulting filtered training set was significantly smaller than the original training set, which made it very difficult for the combiner to learn any general rules applicable to the test set. The ten base experts used for these three schemes were the same ten used for the 14 hidden node experiments in the simple ensemble averaging/voting trials. The network combiner was trained across the same parameters as the original base classifiers, i.e., 8, 10, 12, 14, and 16 hidden nodes, 0.01 learning rate, 0.5 momentum rate, and 1000 epochs of training. The combiner network was then tested on the entire original test set.

Table 3 presents the results obtained using these three schemes. The first column indicates the number of hidden nodes in the network combiner. The second column gives the test set results obtained when using the 10 base expert opinions combined with the 22 original features. The third column presents the test set results obtained when using only the 10 base expert opinions for training and testing. The final column gives the test set results for the filter approach. All results represent the average of three runs.

Table 3. Results for Various Committee Machine Approaches

NUM Hidden Nodes	Features + Experts	Experts Alone	Filter
8	**74.07%**	73.57%	50%
10	**73.74%**	73.40%	50%
12	**73.91%**	73.23%	50%
14	**73.91%**	73.57%	50%
16	72.05%	**73.40%**	50%

One of the most striking results presented in Table 3 are those results obtained from the features plus experts scheme. In fact, when the network combiner had only eight hidden nodes, the system produced results comparable to the best obtained to date. While this is encouraging, it is not extremely surprising. By giving the combiner the additional information of the base experts' decisions, the system is able to produce a more robust opinion as to what constitutes a sand wave. It is surprising to see that using only the trained experts' opinions also produced comparatively high results for the test set. This implies that the combiner network is learning which experts to pay attention to based solely on their individual opinions without any additional information as to the underlying input space. Perhaps unsurprisingly, the filter approach did not perform well. This is probably due to the fact that there were only 100 training instances on which the two base experts disagreed. For this approach to be more successful a much larger training set to be filtered would be required so that the combiner would have more data items to train on.

4.3.4. Discussion of Classification Results

The current set of experiments make a significant argument for the use of committee machines in pattern classification. In general, we found that a combination of experts

consistently provided our best results in the development of our system. Of particular interest is the mixture of base experts with the original input vectors. This model allows the combiner to learn which experts are reliable in the classification of different areas of the input space and to take advantage of this additional information in the overall classification. However, despite these interesting results, our current system is unable to accurately classify more than 74% of the test set.

At this point in the development of our sonar classification system, the possibility that no new strategy will result in a more accurate classifier must be conceded. It is possible that the data is too noisy to be separated with any more accuracy than 74%, as even human experts often disagree on correct classification. Another possibility, as Burl, et al. (1998) suggest, is that the diversity of the negative examples could have an impact on the classification system. They also emphasize, however, that in the development of a classification system, feature extraction is as important, if not more so, than testing numerous styles of classifiers. After briefly meeting with experts at the NAVOCEANO, it was evident that humans use numerous contextual clues present in the image for correct classification. Therefore it seems likely that the features extracted from the original images are not sufficient to encapsulate what experts are calling a sand wave. While some contextual information is inherent in the current set of features, much more domain knowledge could be imbedded in the decision making process. Of particular interest will be expanding the GLCM distance from 1 to 5 or greater. This should provide more of the contextual data that human experts seem to rely so heavily on.

5.0 On-Going Research and Development

The research focus established with this DEPSCoR grant is on-going. NAVOCEANO at Stennis Space Center continues to fund work for 1) integrating our knowledge discovery system with their image processing software and 2) extending the capability of our system to characterization of the roughness of the provinces. In addition, three graduate students are completing degrees (one Ph.D. and two Masters) in research initiated as part of this grant.

5.1 Integration with UNISIPS

UNISIPS is the image processing system developed by NAVOCEANO to support many of their image analysis tasks including provincing. We are currently integrating a custom Graphical User Interface extension into the existing NAVO tools. Figure 11 shows a screen shot of a portion of this GUI. Our goal is to allow the geologist user base to utilize our processes without detailed knowledge of the algorithms.

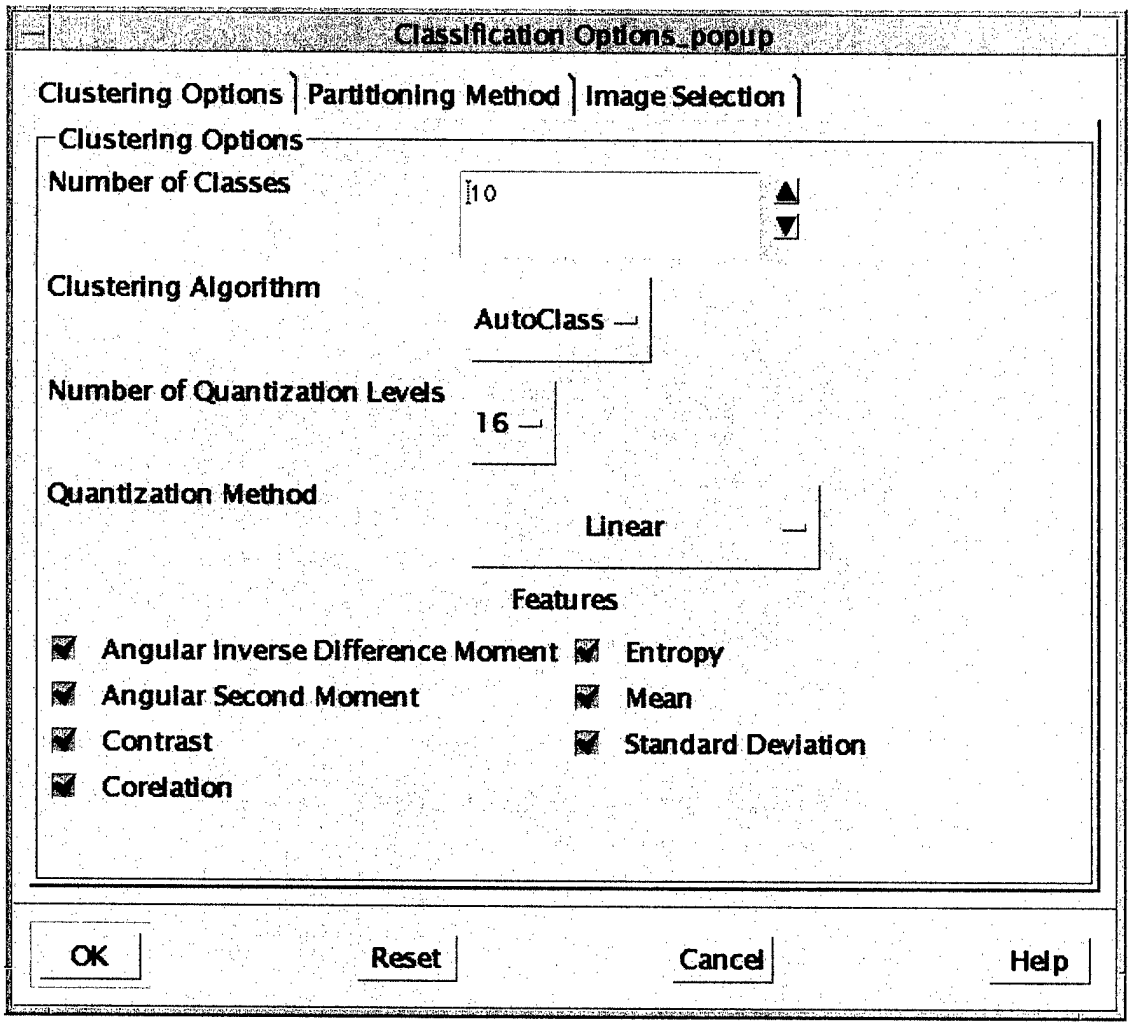

Figure 11. Screen Shot of GUI to Integrate with UNISIPS

5.2 Roughness Classification

Section 4 above discusses our efforts to date to classify sections of the images into "sand wave" and "non-sand wave" categories. We are working to improve the accuracy of this classifier. In addition, our goal is to extend this work so that our system will be able to automatically classify provinces into the roughness categories defined by Navy doctrine. Results from the work that we have completed suggests that we need to consider the use of additional features, particularly those that will provide more contextual information.

5.3 Scalable Knowledge Discovery

Our second annual report included an extensive discussion of work we have done in the area of scalable knowledge discovery. The computational requirements of the algorithms used in our knowledge discovery work and the size of the data sets involved have prompted us to

investigate methods for parallelizing the knowledge discovery process. The methods we have developed for parallelizing unsupervised learning algorithms are extensions of those proposed by Chan and Stolfo (1995; 1996) for supervised learning. Three conference papers describing our work in this area have been presented (Wooley, et al 1998; Wooley, et al. 2000a; 2000b). This work is the focus of Bruce Wooley's Ph.D. research. Further support for research in the area of scalable knowledge discovery was provided by a Mississippi State University College of Engineering Hearin Award and an NSF CISE Instrumentation Award (see section 9 below)..

5.3 Fuzzy Spatial Data Mining

The knowledge discovery work that we have described above focuses on identifying and characterizing different regions of the ocean floor. We are also interested in building on this work to identify relationships between different classes of regions after the regions have been identified. We would like to be able to extract relationships such as "regions of class 1 are typically found east of regions of class 3." In order to do this sort of data mining, one must be able to describe the directional relationships between individual regions and between groups of regions. We are using an algorithm developed by Isabelle Bloch (1999) for finding the fuzzy relative position of objects in image space. Bloch's method works by computing a fuzzy landscape for a reference object in a specified direction. This landscape gives the degree to which each pixel in the image is in the specified direction of the reference object. Then one can use fuzzy pattern matching to evaluate how well other objects match with areas in the fuzzy landscape to determine the fuzzy relative position of the objects with respect to the reference object. Bloch (1999) gives both an exact and an approximation algorithm for computing the fuzzy landscape. We have found that the much more efficient approximation algorithm provides sufficient accuracy for our data mining work.

From the fuzzy landscapes, one can identify the relationships between all possible pairs of objects in a predetermined set of directions. We define $R_a(m,n)$ to represent a binary relation that says object m is in the direction α of object n. This is a fuzzy relation and so each element has an associated membership value. We are interested only in elements of the relation whose membership is above some specified threshold. We are developing methods to efficiently compute these relations and to use the relations to extract rules that describe typical relationships found among classes of objects in images. We are initially developing the methods using synthetic data containing known relationships and will extend the work to classified image data. George Brannon Smith is developing these algorithms as part of his Masters thesis.

5.5 Parameter Optimization using Bayesian Networks

Our knowledge discovery process is complex and requires that the user have some understanding of a long sequence of algorithmic steps and how they interact in order to find the best options and parameter settings for each step. We would like to provide the oceanographers with an intelligent interface to the image-processing and data-mining software that will insulate users from its complexity by automatically directing the discovery process and providing expert guidance to assist the user. As a first step in the development of this intelligent interface, we are investigating the use of Bayesian networks to represent causal relationships between different options selected by the user. Also incorporated in the network is information about the goals of

the user and about characteristics of the images under consideration. We have found, that for a relatively simple task—selecting appropriate region and boundary thresholds for the region-growing process—selecting the set of parameters with the highest probability of yielding a "good" result only works about 50% of the time. We are exploring methods for finding a policy that will allow the user to explore other options in the most optimal order. This research is the Masters thesis of Sean Taylor. A new faculty member in the Computer Science Department at Mississippi State, Dr. Eric Hansen, is also participating in this aspect of the research.

6. Summary

In this project, we have developed a knowledge discovery system to identify regions of similar visual texture from sonar images. The work is challenging because the regions of interest to the oceanographers at NAVOCEANO are characterized by very fine grain textural differences, but the regions cover regions of large extent and the images are quite noisy. The results of this work have been well-received by our collaborators at NAVOCEANO and we are currently integrating our software with their image processing tools. We continue to investigate methods for extending the capability of the system to classify regions by roughness according to Navy doctrine. We are also extending the basic research done on the grant in several ways as described above. Publications resulting from the project, students who have been trained as part of the project, and other funding received to extend work on this project are listed below.

7. Publications Resulting From This Project

Bridges, Susan, Julia Hodges, Bruce Wooley, Donald Karpovich, and George Brannon Smith. 1999. Knowledge discovery in an oceanographic database. *Applied Intelligence* 11, 135-148.

Hodges, Julia, Susan Bridges, Bruce Wooley, Donald Karpovich, and Brannon Smith. 1997. *Knowledge Discovery in an Object-Oriented Oceanographic Database System.* October 21, 1997. Mississippi State University Technical Report #971021.

Karpovich, Donald. 1998. Choosing the optimal features and texel sizes in image categorization. In *Proceedings of the 36th ACM Southeast Conference held in Marietta, GA, April 1-3, 1998.* 104-107.

Klimek, Lee, Bruce Wooley, Susan Bridges, Julia Hodges, Andrew Watkins, Sara Smolensky. 1999. A comparison of the performance of a Bayesian algorithm and a Kohonen map for clustering texture data. In *Proceedings of the Artificial Neural Networks in Engineering Conference (ANNIE '99) held November 7-10, 1999, in St. Louis, MO.* 777-784.

Wooley, Bruce, Yoginder Dandass, Susan Bridges, Julia Hodges, and Anthony Skjellum. 1998. Scalable knowledge discovery from oceanographic data. In *Proceedings of the Artificial Neural Networks in Engineering Conference (ANNIE '98), St. Louis, MO, November 1998.*

Wooley, Bruce and George Brannon Smith. 1998. Region-growing techniques based on texture for provincing the ocean floor. In *Proceedings of the 36th ACM Southeast Conference held in Marietta, GA, April 1-3, 1998.* 99-103.

Wooley, Bruce, Susan Bridges, Julia Hodges, and Anthony Skjellum. 2000a Scaling the data mining step in knowledge discovery using oceanographic data. Accepted for presentation at and publication in the *Proceedings of the Thirteenth International Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems (IEA/AIE) June 19-22, 2000.*

Wooley, Bruce, Diane Mosser-Wooley, Anthony Skjellum, and Susan Bridges. 2000b. Machine Learning Using Clusters of Computers. In *Proceedings of MPIDC2000 Message Passing Interface and High-Performance Clusters Developer's and User's Conference, March 21-23, 2000, Cornell University.*

8. Students Trained as Part of the Project

Donald Karpovich. M.S. student. Graduated August 1999.
Lee Klemik. M.S. student. Graduated May 2000.
George Brannon Smith. M.S. student. Expected date of graduation December 2000.
Sara Smolensky. Undergraduate student. Expected date of graduation May 2001.
Sean Taylor. M.S. student. Expected date of graduation December 2000.
Andrew Watkins. M.S. student. Expected date of graduation May 2001.
Bruce Wooley. Ph.D. student. Expected date of graduation December 2000.

9. Other Research Support Resulting From This Project

1. Source: Naval Oceonographic Command, Funded by NASA
   Title: Roughness Characterization of Provinces Identified from Acoustic Imagery
   PIs: Susan Bridges and Julia Hodges
   Amount: $50, 000
   Date: March 21, 2000 – September 30, 2000.

2. Source: National Science Foundation
   Title: A Gigabit/s, VIA-Enabled Cluster Architecture for Research in High Performance Systems Software, Scalable Knowledge Discovery, Visualization, and Parallel Planning under Uncertainty
   PIs: Anthony Skjellum, Julia Hodges, Lois Boggess, Susan Bridges, Donna Reese, Raghu Machiraju, and Eric Hansen
   Amount: $214,939
   Date: July 1, 1999 – June 30, 2002.

3. Source: Naval Oceanographic Office through NASA, Stennis.
   Title: Province Identification and Classification in Acoustic Imagery
   PIs: Susan Bridges and Julia Hodges
   Amount: $97,984
   Date: August 28, 1998 – May 30, 1999.

4. Source: Mississippi State University College of Engineering Hearin Fund.
   Title: The Scalable Knowledge Discovery Initiative
   PIs: Susan Bridges, Julia Hodges, Anthony Skjellum, Raghu Machiraju
   Amount: $49,000
   Date: August 16, 1997 – August 15, 1998.

10. References

Boggess, L. 2000. Interview by author, 3 May, Mississippi State University. Personal Communication

Bloch, Isabelle. 1999. Fuzzy relative position between objects in image processing: A morphological approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21(7): 657-664.

Bridges, Susan, Julia Hodges, Bruce Wooley, Donald Karpovich, and George Brannon Smith. 1999. Knowledge discovery in an oceanographic database. *Applied Intelligence* 11, 135-148.

Burl, M.C., L. Asker, P. Smyth, U. Fayyad, P. Perona, L. Crumpler, and J. Aubele. 1998. Learning to recognize volcanoes on Venus. *Machine Learning* 30 (2/3): 165-94.

Chan, Philip K., and Salvatore J. Stolfo. 1995. Learning arbiter and combiner trees from partitioned data for scaling machine learning. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*. Edited by Usama Fayyad and Ramasamy Uthurusamy. Menlo Park, CA: AAAI Press. 39-44.

Chan, Philip K., and Salvatore J. Stolfo. 1996. Scalable exploratory data mining of distributed geoscientific data. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. Edited by Evangelos Simoudis, Jiawei Han and Usama Fayyad. Menlo Park, CA: AAAI Press. 2-7.

Cheeseman, Peter, and John Stutz. 1996. Bayesian classification (AutoClass): Theory and results. *Advances in knowledge discovery and data mining*. Edited by Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy. Menlo Park, CA: AAAI Press. 158-180.

Fayyad, Usama M., Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. From data mining to knowledge discovery: An overview. *Advances in knowledge discovery and data mining.* Edited by Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy. Menlo Park, CA: AAAI Press. 1-36.

Fisher, Doug, Lin Xu, James R. Carnes, Yoran Reich, Steven J. Fenves, Jason Chen, Richard Shiavi, Gautam Biswas, and Jerry Weinberg. 1993. Applying AI clustering to engineering tasks. *IEEE Expert* 8(6): 51-60.

Goodman, PH. 1998. *NevProp software, version 4.* Reno, NV: University of Nevada. http://www.scs.unr.edu/nevprop/.

Haralick, R.M. 1979. Statistical and structural approaches to texture. *Proceedings of the IEEE* 62(5): 786-804.

Haykin, Simon. 1999. *Neural networks: A comprehensive foundation.* Upper Saddle River, NJ: Prentice Hall.

Hodges, Julia, Susan Bridges, Bruce Wooley, Donald Karpovich, and Brannon Smith. 1997. *Knowledge Discovery in an Object-Oriented Oceanographic Database System.* October 21, 1997. Mississippi State University Technical Report #971021.

Karpovich, Donald. 1998. Choosing the optimal features and texel sizes in image categorization. In *Proceedings of the 36th ACM Southeast Conference held in Marietta, GA, April 1-3, 1998.* 104-107

Kira, K., and L. Rendell. 1992. The feature selection problem: Traditional methods and a new algorithm. In *Proceedings of the Tenth National Conference on Artificial Intelligence held in San Jose, California, July, 1992,* by AAAI, 129-34. Menlo Park, CA: AAAI Press.

Kittler, Josef, Mohamad Hatef, Robert P.W. Duin, and Jiri Matas. 1998. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 20(3). 226-239.

Klimek, L. 1998. Kohonen map package developed for Mississippi State University's NAVO project. Mississippi State University.

Klimek, L., B. Wooley, S. Bridges, J. Hodges, A. Watkins, and S. Smolensky. 1999. A comparison of the performances of a bayesian algorithm and a kohonen map for clustering texture data. In *Proceedings of the conference on artificial neural networks in engineering (ANNIE '99),* St. Louis, MO, November 7-10, 1999.

Kohonen, T., J. Hynninen, J. Kangas, and J. Laaksonen. 1996. SOM_PAK: The self-organizing map program package. Technical Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science, FIN-02150 Espoo, Finland. http://www.cis.hut.fi/research/som_lvq_pak.shtml

Narendra, P. and K. Fukunaga. 1977. A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computers* C:26(9): 917-22.

Quinlan, J.R. 1992. *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann Publishers.

Reed, Thomas Beckett IV, and Donald Hussong. 1989. Digital image processing techniques for enhancement and classification of SeaMARC II side scan sonar imagery. *Journal of Geophysical Research* 94(B6): 7469-90.

Wooley, Bruce, Yoginder Dandass, Susan Bridges, Julia Hodges, and Anthony Skjellum. 1998. Scalable knowledge discovery from oceanographic data. To appear in *Proceedings of the Artificial Neural Networks in Engineering Conference (ANNIE '98), St. Louis, MO, November 1998.*

Wooley, Bruce and George Brannon Smith. 1998. Region-growing techniques based on texture for provincing the ocean floor. In *Proceedings of the 36th ACM Southeast Conference held in Marietta, GA, April 1-3, 1998.* 99-103.

Wooley, Bruce, Susan Bridges, Julia Hodges, and Anthony Skjellum. 2000a Scaling the data mining step in knowledge discovery using oceanographic data. Accepted for presentation at and publication in the *Proceedings of the Thirteenth International Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems (IEA/AIE) June 19-22, 2000.*

Wooley, Bruce, Diane Mosser-Wooley, Anthony Skjellum, and Susan Bridges. 2000b. Machine Learning Using Clusters of Computers. In *Proceedings of MPIDC2000 Message Passing Interface and High-Performance Clusters Developer's and User's Conference, March 21-23, 2000, Cornell University.*

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	June 14, 2000	Final Technical Report 9/1/96-12/30/99

4. TITLE AND SUBTITLE
Knowledge Discovery in an Oceanographic Database System

5. FUNDING NUMBERS
Grant No.
N00014-96-1-1276
P.R. No.
96PRO7924-00

6. AUTHOR(S)
1. Julia Hodges
2. Susan Bridges

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(ES)
Mississippi State University
P.O. Box 6156
Mississippi State, MS 39762

8. PERFORMING ORGANIZATION REPORT NUMBER
P.O. Code 321SI

9. SPONSORING / MONITORING AGENCY NAMES(S) AND ADDRESS(ES)
Office of Naval Research Regional Office (Atlanta)
101 Marietta Tower Suite 2805
101 Marietta Street
Atlanta, GA 30321-0008

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
AGO Code N66020

11. SUPPLEMENTARY NOTES

a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release

12. DISTRIBUTION CODE
N68892

13. ABSTRACT *(Maximum 200 words)*

The rate at which scintific data is collected today has overwhelmed the ability of scientists to store and analyze the data. This report describes the results of a three year effort in the development of a knowledge discovery system for use by oceanographers at the Naval Oceanographic Office (NAVOCEANO) at the Stennis Space Center in the identification of provinces of interest in the ocean floor from acoustic imagery. The system is composed of a knowledge discovery component built to interact with a database system currently in use at Stennis Space Center. The knowledge discovery system applies machine learning techniques to features extracted from sonar images to identify provinces of the ocean floor based on visual texture. This requires that the images be segmented into regions of homogeneous texture using a region-growing technique, that features describing the texture of these regions be extracted, that machine learning techniques be applied to classify the regions, that classified images be constructed for visualizing the results, and that the classified images be combined and geo-referenced using a mosaic procedure. NAVOCEANO is currently supporting efforts to integrate the software developed from this project with their image analysis system.

14. SUBJECT TERMS
Knowledge Discovery, Data mining, Acoustic Imagery, and Object-Oriented Database

15. NUMBER OF PAGES
25

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
unclassified	unclassified	unclassified	