Active Learning for Detection of Mine-Like Objects in Side-Scan Sonar Imagery

Esther Dura, Yan Zhang, Xuejun Liao, Gerald J. Dobeck, and Lawrence Carin, *Fellow, IEEE*

***Abstract*—A data-adaptive algorithm is presented for the selection of the basis functions and training data used in classifier design with application to sensing mine-like targets with a side-scan sonar. Automatic detection of mine-like targets using side-scan sonar imagery is complicated by the variability of the target, clutter, and background signatures. Specifically, the strong dependence of the data on environmental conditions vitiates the assumption that one may perform *a priori* algorithm training using separate side-scan sonar data collected previously. In this paper, a novel active-learning algorithm is developed based on kernel classifiers with the goal of enhancing detection/classification of mines without requiring an *a priori* training set. It is assumed that divers and/or unmanned underwater vehicles (UUVs) may be used to determine the binary labels (target/clutter) of a small number of signatures from a given side-scan collection. These sets of signatures and associated labels are then used to train a kernel-based algorithm with which the remaining side-scan signatures are classified. Information-theoretic concepts are used to adaptively construct the form of the kernel classifier and to determine which signatures and associated labels would be most informative in the context of algorithm training. Using measured side-looking sonar data, the authors demonstrate that the number of signatures for which labels are required (via diver/UUV) is often small relative to the total number of potential targets in a given image. This procedure designs the detection/classification algorithm on the observed data itself without requiring *a priori* training data and also allows adaptation as environmental conditions change.**

***Index Terms*—Active learning, classifiaction, detection, mine-like, side-scan sonar, target, unmanned underwater vehicle (UUV).**

I. Introduction

FUTURE mine counter measure (MCM) operations will likely make use of unmanned underwater vehicles (UUVs) equipped with long- and/or short-range sensors (e.g., side-scan sonar and cameras, respectively) and employing computer-aided detection and classification (CAD/CAC) algorithms. The detection phase is defined as the process of delineating those signatures that have the possibility of being a mine. During detection, one must recognize mines with high probability, accepting a potentially large number of false alarms. In the subsequent classification stage, algorithms are designed to reject as many of the false alarms as possible while retaining actual mines. Detection algorithms are generally simple since they must be applied to all observed data while classification algorithms are typically more sophisticated, being only applied to the pruned signatures. This paper focuses primarily on the classification phase, with the goal of developing algorithms that do not require an *a priori* training set, and motivated by real-world complexities elucidated below. As discussed further below, the contribution of this paper is in the development of a technique for data-adaptive selection of the feature basis vectors and the set of training data used to design a kernel-based classifier, with performance demonstrated on measured side-scan sonar data.

The authors focus here on wide-area coverage via a side-scan sonar. Side-scan sonar affords the ability to operate at long ranges (100 to 300 m), permitting sensing over large regions. The high-frequency character of many of these sensors yields features that are similar to those found in optical imagery. For example, a paired highlight and a shadow region (from the front of the target and from acoustic blockage at the rear, respectively) are the primary features for the detection of mine-like objects. However, the detection of mines is complicated by significant variability in the appearance of the background, mine-like signatures, and clutter. The mine signature variability is caused by the large number of underwater mine types, by mine deployment variation, and by mine–background interaction over time due to water currents.

To address this problem, supervised classification techniques have been implemented [1]–[5]. With these techniques, one typically requires an *a priori* set of training data consisting of a set of signatures and associated binary labels (target/clutter). To constitute a training set, known targets (e.g., mines) must be emplaced in a given environment and side-scan data collected, with all nonemplaced scatterers assumed to be clutter. The difficulty of this procedure resides 1) in the very large number of mine types, mine deployments, and mine histories (how long they have been deployed); and 2) the significant dependence of the imagery on the properties of the environment, for example, the properties of the sea bottom. The variability of 1) and 2) makes it virtually impossible to constitute a training set that is robust to all mine deployments and environments to be encountered. In previous studies on the assessment of CAD/CAC algorithm performance, researchers have typically divided a given data set into a portion used for training and the remaining used for testing [1]. Since in this case the mine and environmental conditions of the training and testing data are often well matched, the results from such studies are an optimistic view of the performance that may be achieved in the field.

Manuscript received December 10, 2003; revised September 16, 2004; accepted January 31, 2005. Associate Editor: J. Preisig.

E. Dura, Y. Zhang, X. Liao, and L. Carin are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708-0291 USA (e-mail: edura@ee.duke.edu; yzhang@ee.duke.edu; xjliao@ee.duke.edu; lcarin@ee.duke.edu).

G. J. Dobeck is with the Dahlgren Division, Naval Surface Warfare Center (NSWC) Coastal Systems Station, Panama City, FL 32407-7001 USA (e-mail: gerald.dobeck@navy.mil).

Digital Object Identifier 10.1109/JOE.2005.850931

In this paper, the authors present a new framework applicable to the "real" MCM problem, accounting for the fact that it is unlikely that an appropriate *a priori* training set will be available for operations in general environments. It is assumed that a side-scan sonar collects data for wide-area surveillance. The authors also assume access to small/mobile UUVs and/or divers that may interrogate signatures of interest at close range (e.g., with cameras or other close-range sensors) to ascertain the associated labels (target/clutter). This yields a set of signatures and associated labels with which a classification algorithm may be designed to analyze the remaining side-scan sonar imagery. This paper addresses the problem of determining the information content accrued by a set of signatures and associated labels, and guiding the selection of those signatures for which knowledge of the associated labels would be most informative (this information content is computed without *a priori* knowledge of the labels themselves).

Stating the problem mathematically, let $\{\mathbf{x}_i\}_{i=1,N}$ represent the known measured side-scan sonar signatures of N underwater objects, with the set of all $\mathbf{x}_i$ denoted as $\mathbf{X}$. The set $\mathbf{X}$ is defined in the initial detection phase. Let $\{y_i\}_{i=1,N}$ represent the associated unknown binary labels (target/nontarget) of the signatures to be determined in the classification phase. An observed signature or feature vector $\mathbf{x}$ may be classified using a kernel-based function [6]–[8] of the form

$$f(\mathbf{x}) = \sum_{i=1}^{n} w_i K(\mathbf{x}, \mathbf{b}_i) + w_o \tag{1}$$

where $\mathbf{b}_i$ is the ith basis function, w_i are scalar weights, w_o is a scalar offset or bias, and $K(\mathbf{x}, \mathbf{b}_i)$ is a general kernel [6]–[8] defining the similarity of $\mathbf{x}$ and $\mathbf{b}_i$. Similar kernel-based approaches to the form in (1) are utilized by the support vector machine (SVM) [6], [7], the relevance vector machine (RVM) [8], as well as many other related algorithms.

For the approach presented in this paper, the set of basis functions $\mathbf{B}_n = \{\mathbf{b}_i\}_{i=1,n}$ is selected from the observed data $\mathbf{X}$, i.e., $\mathbf{B}_n \subset \mathbf{X}$. The number of required basis functions n is data dependent and is determined adaptively by the algorithm. Specifically, by using fundamental information-theoretic considerations (detailed below), the set $\mathbf{B}_n$ is defined by selecting those signatures from $\mathbf{X}$ that are most representative of the measured data. The labels (identities) of the underwater objects associated with $\mathbf{B}_n$ are not required. Once the basis set $\mathbf{B}_n$ is defined, the associated model weights $\{w_i\}_{i=0,n}$ (denoted collectively by the vector $\mathbf{w}$) are determined, and for this task labeled data are required. The authors thus define a subset of signatures $\mathbf{X}_s \subset \mathbf{X}$ for which knowledge of the associated labels $\mathbf{L}_s$ would be most informative in the context of defining the model weights. The set of signatures $\mathbf{X}_s$ is determined using information-theoretic metrics, as detailed below. Note that the sets $\mathbf{B}_n$ and $\mathbf{X}_s$ may overlap, but they are generally distinct. After the labels $\mathbf{L}_s$ associated with $\mathbf{X}_s$ have been identified (via close-range mobile UUVs and/or divers), the classification algorithm associated with (1) is trained as usual [1] and then applied to $\mathbf{x} \notin \mathbf{X}_s$. It is important to emphasize that within the framework developed here, the training set $(\mathbf{X}_s, \mathbf{L}_s)$ is determined adaptively on the observed site-dependent data via fundamental information-theoretic metrics without requiring *a priori* training data.

The remainder of the paper is organized as follows. In Section II, the authors detail the theory employed in this framework, with example results presented in Section III using measured side-scan sonar data. Conclusions are addressed in Section IV.

II. Active Classifier Design

A simple detection algorithm is employed on the side-scan imagery to define a set of signatures $\mathbf{X} = \{\mathbf{x}_i\}_{i=1,N}$ associated with the possible mines (with an anticipated large false-alarm rate). The subsequent active design of a kernel-based classifier proceeds in three steps based on these observed unlabeled signatures: 1) selection of basis functions to build the structure of the kernel classifier using unlabeled signatures from $\mathbf{X}$; 2) selection of the signatures for which knowledge of the associated labels would be most informative, this followed by the discovery of the associated labels via a near-range sensor (e.g., a camera on a mobile UUV); and 3) estimation of the kernel-classifier weights using the subset of labeled data and the basis functions determined in steps 1) and 2). Note that all of the signatures for which labels are desirable (step 2) may be determined at once, from which one may design an optimal path to guide the near-range UUV to the objects, such that the associated labels may be determined in an efficient manner. These three steps are detailed below.

A. Model Structure

The kernel-based function in (1), using n basis functions, may be expressed concisely as [6]–[8]

$$f_n(\mathbf{x}) = \sum_{i=1}^{n} w_{n,i} K(\mathbf{x}, \mathbf{b}_i) + w_{n,0} = \mathbf{w}_n^{\mathrm{T}} \phi_n(\mathbf{x}) \tag{2}$$

where

$$\phi_n(\mathbf{x}) = [1, K(\mathbf{x}, \mathbf{b}_1), K(\mathbf{x}, \mathbf{b}_2), \ldots, K(\mathbf{x}, \mathbf{b}_n)]^{\mathrm{T}} \tag{3}$$

$$\mathbf{w}_n = [w_{n,0}, w_{n,1}, w_{n,2}, \ldots, w_{n,n}]^{\mathrm{T}}. \tag{4}$$

By construction, in algorithm design, the binary label y of a given signature is set to $y = 1$ for one class and $y = -1$ for the other. In a kernel-based algorithm, the objective is for $f_n(\mathbf{x}) = 1$ if $\mathbf{x}$ is associated with the $y = 1$ class, and $f_n(\mathbf{x}) = -1$ otherwise. For the ith signature $\mathbf{x}_i$ with label y_i, the error in the kernel algorithm may be expressed

$$y_i = \mathbf{w}_n^{\mathrm{T}} \phi_n(\mathbf{x}_i) + \varepsilon_i \tag{5}$$

where $\varepsilon(\mathbf{x}_i)$ is the error term resulting from imperfections in the model. In algorithm design, one of the aims is to find the weights $\mathbf{w}$ that minimize the error observed on training data for which the data and labels are known. If the training data are well matched to the subsequent testing data, then the algorithm is likely to constitute a robust detection procedure. However, as indicated above, in many sensing problems it is impractical to have a separate training set, with this issue addressed by the information-theoretic techniques discussed below.

B. Selection of Basis Functions

Assuming that the ε_i in (5) is independent and that ε_i is modeled as zero mean with variance σ_i^2, then the Fisher information matrix associated with $\mathbf{X}$ and $\mathbf{B}_n$ is defined as [9], [10]

$$\mathbf{M}_n = \sum_{i=1}^{N} \sigma_i^{-2} \phi_n(\mathbf{x}_i) \phi_n^{\mathrm{T}}(\mathbf{x}_i) = \sum_{i=1}^{N} \sigma_i^{-2} \phi_{n,i} \phi_{n,i}^{\mathrm{T}} \quad (6)$$

where $\phi_{n,i} \equiv \phi_n(\mathbf{x}_i)$. Note that for the computation of $\mathbf{M}_n$ the labels associated with $\mathbf{B}_n$ and $\mathbf{X}$ are not required (this is a result of the fact that the model in (2) is linear in the weights $\mathbf{w}_n$). As discussed in [9], the Fisher information matrix in (6) is associated with the errors in fitting the model to all N measured $\mathbf{x}_i$ using the basis $\mathbf{B}_n$. By adding a new basis function to $\phi_n(\cdot)$, one obtains

$$\phi_{n+1}(\cdot) = \begin{bmatrix} \phi_n(\cdot) \\ \phi_{n+1}(\cdot) \end{bmatrix} \quad (7)$$

where $\phi_{n+1}(\cdot) = K(\cdot, \mathbf{b}_{n+1})$ and $\mathbf{b}_{n+1} \in \mathbf{X}$, $\mathbf{b}_{n+1} \notin \mathbf{B}_n$. Following (2), the authors may write from ϕ_{n+1} the augmented classifier f_{n+1} for which the Fisher information matrix is found to be

$$\begin{aligned} \mathbf{M}_{n+1} &= \sum_{i=1}^{N} \sigma_i^{-2} \begin{bmatrix} \phi_{n,i} \\ \phi_{n+1,i} \end{bmatrix} \begin{bmatrix} \phi_{n,i}^{\mathrm{T}} & \phi_{n+1,i} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{M}_n & \sum_{i=1}^{N} \sigma_i^{-2} \phi_{n,i} \phi_{n+1,i} \\ \sum_{i=1}^{N} \sigma_i^{-2} \phi_{n+1,i} \phi_{n,i}^{\mathrm{T}} & \sum_{i=1}^{N} \sigma_i^{-2} \phi_{n+1,i}^2 \end{bmatrix} \end{aligned} \quad (8)$$

where $\phi_{n+1,i} \equiv \phi_{n+1}(\mathbf{x}_i)$. The expression in (8) is again associated with fitting the model to the N measured $\mathbf{x}_i$, but now using an $(n+1)$-member basis set $\mathbf{B}_{n+1}$, *vis-à-vis* the n-member basis $\mathbf{B}_n$ in (6). The authors develop a metric that compares (6) and (8), thereby quantifying the information gain by adding the new basis $\mathbf{b}_{n+1}$.

Of the many ways of comparing the information content reflected by $\mathbf{M}_n$ and $\mathbf{M}_{n+1}$, the so-called D-optimal procedure [9] is employed here, defined as the determinant of the information matrix. Let the logarithm of the determinant of $\mathbf{M}_n$ be denoted as q_n, then using the matrix identity

$$\det \begin{bmatrix} \mathbf{A} & \mathbf{c} \\ \mathbf{c}^{\mathrm{T}} & d \end{bmatrix} = (d - \mathbf{c}^{\mathrm{T}} \mathbf{A}^{-1} \mathbf{c}) \det(\mathbf{A})$$

the authors have

$$q_{n+1} = q_n + \ln r(\mathbf{b}_{n+1}) \quad (9)$$

where

$$\begin{aligned} r(\mathbf{b}_{n+1}) &= \sum_{i=1}^{N} \sigma_i^{-2} \phi_{n+1,i}^2 \\ &\quad - \sum_{i=1}^{N} \sigma_i^{-2} \phi_{n+1,i} \phi_{n,i}^{\mathrm{T}} \mathbf{M}_n^{-1} \sum_{i=1}^{N} \sigma_i^{-2} \phi_{n,i} \phi_{n+1,i}. \end{aligned} \quad (10)$$

Since $N \geq n$, the matrix $\mathbf{M}_n$ is full rank and its inverse exists (assuming that n of the vectors $\{\phi_n(\mathbf{x}_i)\}_{i=1,N}$ are linearly independent). Under these conditions, it may be shown that $r > 0$, and therefore $\ln r$ in (9) is generally valid.

It is known from information theory [10] that the inverse of $\mathbf{M}_n$ gives the Cramer–Rao lower bound (CRLB) of the covariance matrix of the estimate of $\mathbf{w}_n$. A large q_n implies low variances of the components of $\mathbf{w}_n$. Given the nth order decision function f_n, q_n is fixed, and one relies on the maximization of $\ln r(\mathbf{b}_{n+1})$ to obtain a large value of q_{n+1}. This can be achieved by conducting a "greedy" search for the new $\mathbf{b}_{n+1}$ in $\mathbf{X}$ with the previously selected support data excluded as

$$\mathbf{b}_{n+1} = \arg\max_{\mathbf{b} \in \mathbf{X}, \mathbf{b} \notin \mathbf{B}_n} \ln r(\mathbf{b}). \quad (11)$$

Using the procedure outlined above, basis elements $\mathbf{b}_n$ are appended until the information gain reflected in $q_{n+1} - q_n$ is no longer deemed significant. Note from (9) and (10) that evaluation of (11) does not require knowledge of the target labels y_i, and therefore no identification (i.e., navigation of a UUV to specific locations) is required to determine the basis $\mathbf{B}_n$.

The authors have introduced a variance σ_i^2 to model the error of the regression model with respect to $\mathbf{x}_i$. By using different σ_i^2, one may weigh the relative importance the algorithm associates with $\mathbf{x}_i$. In the paper presented here, the σ_i^2 are assumed to be the same for all data samples, and therefore from (9), (10), and (11) σ_i^2 is simply a constant that does not affect which feature vectors are selected as basis functions.

Note that there are other procedures one may consider for the design of the basis set $\mathbf{B}_n$ such as vector quantization (VQ) [11], learning vector quantization (LVQ) [12], and principal component analysis (PCA) [13]. In each of these, the final sets of basis vectors $\mathbf{B}_n$ are not, in general, members of $\mathbf{X}$ (e.g., they are eigenvectors or centroids for PCA and VQ, respectively). Any of these approaches may be used for the design of $\mathbf{B}_n$; the authors have chosen the approach elucidated above because all members of $\mathbf{B}_n$ are members of $\mathbf{X}$, allowing a direct comparison to other machine-learning algorithms such as the SVM [6], [7] and the RVM [8]. In the SVM and RVM algorithms, the basis $\mathbf{B}_n$ comes from a labeled training data set, where in the procedure outlined above the basis vectors are members of the unlabeled measured data $\mathbf{X}$.

It is interesting to note that if σ_i is the same for all i, as assumed below, (6) reduces to the matrix employed in PCA [13], where in this case the eigenvectors are associated with the set of vectors $\phi_n(\mathbf{x}_i) = [1, K(\mathbf{x}_i, \mathbf{b}_1), K(\mathbf{x}_i, \mathbf{b}_2), \ldots, K(\mathbf{x}_i, \mathbf{b}_n)]^{\mathrm{T}}$ for $i = 1, \ldots, N$. In PCA, the authors use $\{\phi_n(\mathbf{x}_i)\}_{i=1,N}$ and compute the associated eigenvectors, retaining those with large eigenvalues. By maximizing the determinant of (6) with each new data as in (11), the authors are essentially defining $\mathbf{B}_n$ as those members of $\mathbf{X}$ that add new information to the associated eigenbases of $\{\phi_n(\mathbf{x}_i)\}_{i=1,N}$ [utilizing the connection between the determinant and eigenvalues of (6)]. Once additional members of $\mathbf{X}$ no longer increase (11) substantially, implicitly the additional eigenvalues associated with (6) are small, no longer yielding significant ("principal") eigenvectors. It is therefore to be noted that the procedure outlined above for the design of $\mathbf{B}_n$ is closely related to PCA, the distinction being that PCA yields an eigenbasis of $\{\phi_n(\mathbf{x}_i)\}_{i=1,N}$, while in the procedure discussed here the elements of $\mathbf{B}_n$ are members of $\mathbf{X}$, consistent as noted with the SVM, RVM, Kernel Matching Pursuit (KMP), and related kernel machines.

C. Selection of Labeled Data, for Model Training

Assume that the procedure discussed above selects n bases, defining $\mathbf{B}_n \subset \mathbf{X}$. The authors now require labeled data to optimize the associated model weights $\mathbf{w}$. In a manner analogous to the previous discussion, the authors select those $\mathbf{x}_i \in \mathbf{X}$ for which knowledge of the associated labels y_i would be most informative in the context of defining $\mathbf{w}$. Those $\mathbf{x}_i$ that are so selected define a subset of signatures $\mathbf{X}_s \subset \mathbf{X}$, and the identity of these objects is identified (e.g., by a near-range UUV) to learn the respective set of labels $\mathbf{L}_s$. The sets of signatures and labels $(\mathbf{X}_s, \mathbf{L}_s)$ are then used to define the weights $\mathbf{w}$ in a least squares sense, and the resulting model $f(\mathbf{x})$ is then used to specify which of the remaining signatures $\mathbf{x} \notin \mathbf{X}_s$ are likely targets of interest.

Assume that there are J signatures in $\mathbf{X}_s$, denoted $\mathbf{X}_{s,J}$. The authors quantify the information content in $\mathbf{X}_{s,J}$ in the context of estimating the model weights $\mathbf{w}$ and further ask which $\mathbf{x}_i \notin \mathbf{X}_{s,J}$ would be most informative if it and its label were added for the determination of $\mathbf{w}$. Analogous to (6), the authors have

$$\mathbf{M}_n(\mathbf{X}_{s,J}) = \sum_{i:\mathbf{x}_i \in \mathbf{X}_{s,J}} \sigma_i^{-2} \phi_{n,i} \phi_{n,i}^{\mathrm{T}}. \tag{12}$$

The expressions in (6) and (12) both employ an n-member basis set $\mathbf{B}_n \subset \mathbf{X}$. The distinction is that in (6) the authors are interested in defining $\mathbf{B}_n$ and sum over all observed signatures $\{\mathbf{x}_i\}_{i=1,N}$. By contrast, in (12), the basis set $\mathbf{B}_n$ is known and fixed, and the authors are only summing over those signatures $\mathbf{X}_{s,J}$ for which knowledge of the associated labels is most informative in defining the model weights $\mathbf{w}$.

After adding a new signature $\mathbf{x}_i \in \mathbf{X}$, $\mathbf{x}_i \notin \mathbf{X}_{s,J}$, the authors now have $\mathbf{X}_{s,J+1}$ and $\mathbf{M}_n$ is updated as

$$\mathbf{M}_n(\mathbf{X}_{s,J+1}) = \mathbf{M}_n(\mathbf{X}_{s,J}) + \sigma_{i_{J+1}}^{-2} \phi_{n,i_{J+1}} \phi_{n,i_{J+1}}^{\mathrm{T}} \tag{13}$$

where i_{J+1} represents the index of the new signature selected for $\mathbf{X}_{s,J+1}$. Using the matrix identity $\det(\mathbf{A}+\mathbf{F}\mathbf{F}^{\mathrm{T}}) = \det(\mathrm{I}+\mathbf{F}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{F})\det(\mathbf{A})$, where det denotes determinant, one obtains from (13)

$$q_n(\mathbf{X}_{s,J+1}) = q_n(\mathbf{X}_J) + \ln \rho(\mathbf{x}_{i_{J+1}}) \tag{14}$$

with

$$\rho(\mathbf{x}_{i_{J+1}}) = 1 + \sigma_{i_{J+1}}^{-2} \phi_{n,i_{J+1}}^{\mathrm{T}} \mathbf{M}_n^{-1}(\mathbf{X}_{s,J}) \phi_{n,i_{J+1}}. \tag{15}$$

Care is needed in evaluating the inverse of $\mathbf{M}_n$ since if $J < n$ the matrix is rank deficient. The authors have considered addressing this in either of two ways. A standard approach for inversion of such matrices is to add a small diagonal term to $\mathbf{M}_n$ such that its inverse exists. Alternatively, by construction, one may assume that the items associated with the basis $\mathbf{B}_n$ are all associated with $\mathbf{X}_{s,J}$ (and therefore $J \geq n$), assuring that the matrix is full rank. The authors have examined both procedures and yield comparable results. They use the second approach in all examples presented in Section III.

Having addressed the inverse of $\mathbf{M}_n$, one iteratively maximizes $\ln \rho(\mathbf{x}_{i_{J+1}})$ to obtain

$$\mathbf{x}_{i_{J+1}} = \arg\max_{\mathbf{x} \in \mathbf{X}, \mathbf{x} \notin \mathbf{X}_{s,J}} \ln \rho(\mathbf{x}). \tag{16}$$

Note that to define $\mathbf{x}_{i_{J+1}}$ the authors again do not require the signature labels. The elements of $\mathbf{X}_s$ are selected iteratively in a "greedy" fashion, as indicated in (16), until the information gain is below a prescribed threshold. After J iterations, the authors have defined those signatures $\mathbf{X}_{s,J}$ for which knowledge of the labels will best approximate the weights $\mathbf{w}$. These items are identified (the mobile UUV navigates to specific locations to "discover" the label), yielding the labels $\mathbf{L}_{s,J}$. Since the labels may be determined after all members of $\mathbf{X}_s$ are defined, one may design an optimal (efficient) path for the UUV to visit the associated objects.

It is important to emphasize that the procedure used to select the model basis functions (Section II-B) is myopic, in the sense that the authors select one basis function at a time from all of the unlabeled data, via the D-optimal procedure in (11). Similarly, in this section, the authors choose to label a subset of unlabeled signatures one at a time. One should note that, while practical computationally, such a procedure is not globally optimal. Specifically, an optimal algorithm would perform a combinatorial search for the set of potential basis functions that globally maximize the information content (Section II-B), with the same true for the feature vectors selected for labeling (this section). The suboptimal myopic procedure employed here is motivated by the goal of realizing a computationally tractable algorithm.

D. Estimation of the Weights

For the assumptions underlying the linear model in (5) and assuming that $\varepsilon(\mathbf{x}_i)$ is independent identically distributed (i.i.d.) over the set of i, with knowledge of $\mathbf{B}_n$ and $(\mathbf{X}_{s,J}, \mathbf{L}_{s,J})$ the optimal estimate for the weights $\mathbf{w}$ is expressed as [9]

$$\mathbf{w} = [\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}]^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{y} \tag{17}$$

where $\mathbf{y}$ represents the set of labels determined as discussed in the previous section as

$$\mathbf{y} = \{y_{i_1}, y_{i_2}, \ldots, y_{i_J}\}^{\mathrm{T}} \tag{18}$$

and the $J \times (n+1)$ matrix $\mathbf{\Phi}$ is defined as

$$\mathbf{\Phi} = \begin{bmatrix} \phi_n^{\mathrm{T}}(\mathbf{x}_{i_1}) \\ \phi_n^{\mathrm{T}}(\mathbf{x}_{i_2}) \\ \vdots \\ \phi_n^{\mathrm{T}}(\mathbf{x}_{i_J}) \end{bmatrix} \tag{19}$$

where, for example, $\mathbf{x}_{i_1}$ corresponds to y_{i_1}.

In the classification stage, $\mathbf{x} \notin \mathbf{X}_{s,J}$ are considered and $f(\mathbf{x})$ is computed. For a prescribed threshold t, $\mathbf{x}$ is deemed associated with the $+1$ class if $f(\mathbf{x}) \geq t$ and with the -1 class if $f(\mathbf{x}) < t$, and by varying the threshold t one yields the receiver operating characteristic (ROC). The key component of the model $f(\mathbf{x})$ is that it is linear in the weights $\mathbf{w}$, which yields a closed-form procedure for the selection of $\mathbf{B}_n$ and $\mathbf{X}_{s,J}$, as indicated in the previous sections.

III. Application to Underwater Mine Detection

A. Overview

The active-training methodology addressed in this paper may be applied to any classification problem for which the data labels are expensive to acquire and for which there is no appropriate training data. The authors consider the detection of underwater mines based on side-scan sonar images (see Fig. 1). The results

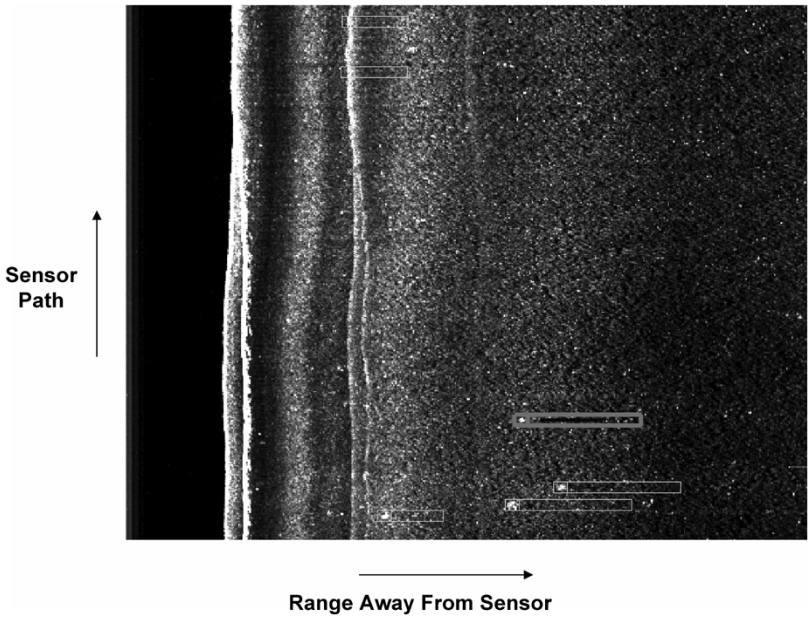

Fig. 1. Typical side-scan sonar image used in this study. Example mine signatures are identified in red, with the size of the regions used to characterize the strong (bright) response from the front of the target, as well as the longer shadow region from behind.

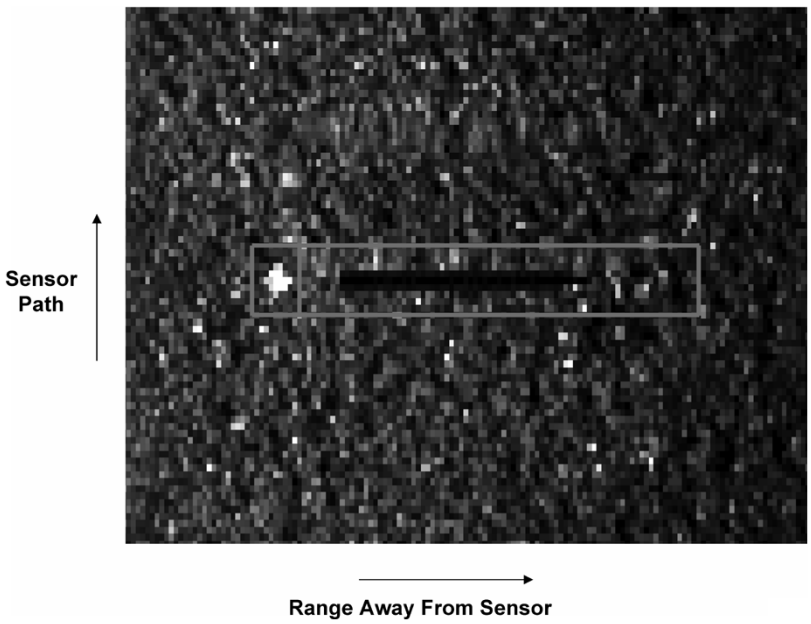

Fig. 2. Example mine signature extracted from Fig. 1. This is a good example for which the initial bright response and subsequent dark shadow are clearly present.

reported in this paper are from side-scan sonar data collected by the Naval Surface Warfare Center (NSWC) Coastal System Station (Panama City, FL). The sonar used is characterized in [1].

A total of 219 images were used for testing the performance of the algorithms. After the detection process (discussed below), $N = 22\,973$ potential targets were detected, of which 119 were mines. This set of N signatures constitutes the data set $\mathbf{X}$ described in Section II. The purpose of the algorithm presented here is to determine which n members of $\mathbf{X}$ should be used to define the basis set $\mathbf{B}_n$ and which should be used to constitute the labeled subset $(\mathbf{X}_s, \mathbf{L}_s)$. After these entities are determined, the classifier is designed as discussed in Section II-D, and classification is performed on the remaining signatures $\mathbf{x}$ satisfying $\mathbf{x} \in \mathbf{X}$, $\mathbf{x} \notin \mathbf{X}_s$.

B. Detection/Prescreening Phase

In the first stage of the algorithm, the image is preconditioned so subsequent detection and classification steps are robust to variations of background level. Simple range normalization is applied, as discussed in [1]. After this step, highlights and shadows are consistent as a function of range and stand out more clearly.

The authors next define regions of interest (ROIs) in the imagery; in this stage, the normalized image is scanned by a non-

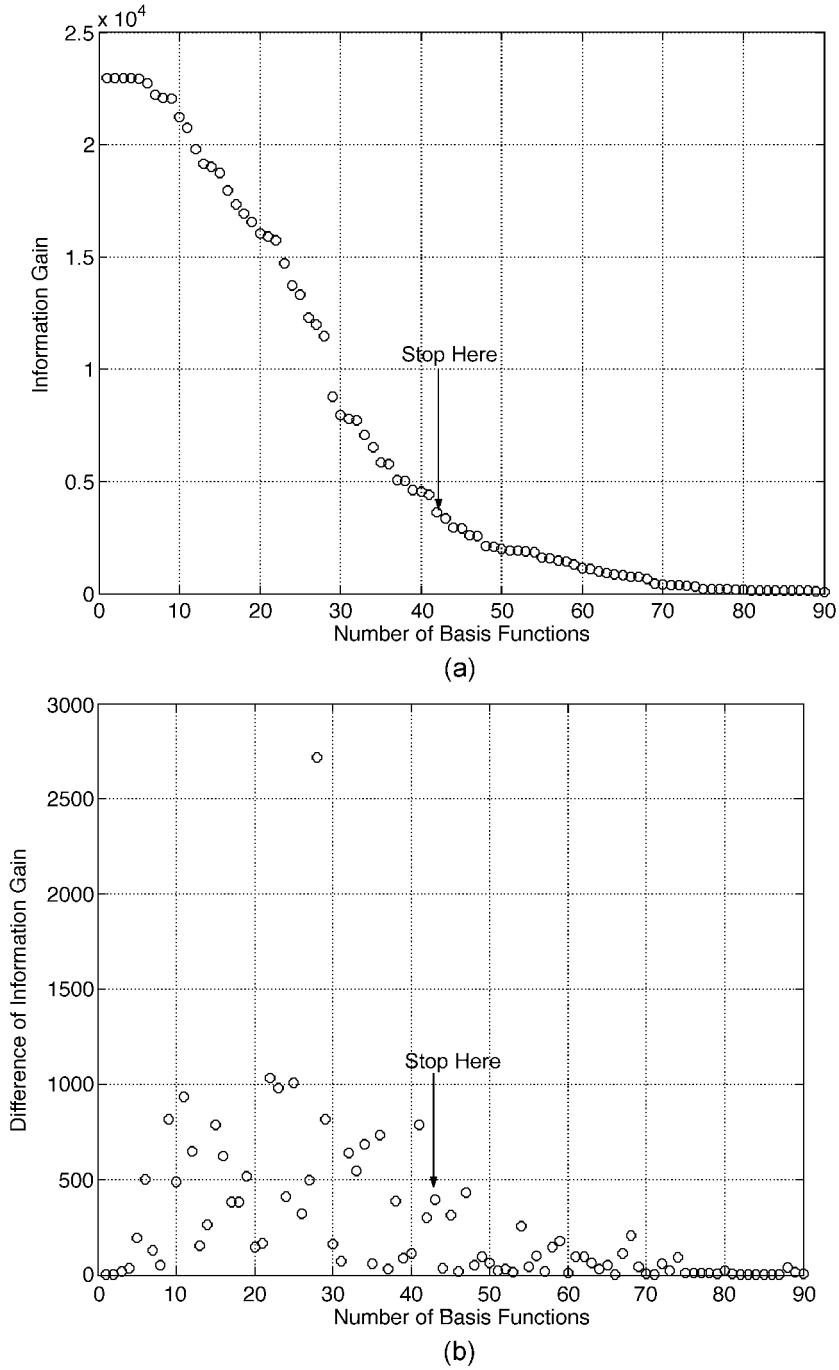

Fig. 3. Information gain as a function of the number of basis functions n selected adaptively. (a) $\ln r(\mathbf{b}_{n+1})$ in (9). (b) $\ln r(\mathbf{b}_{n+2}) - \ln r(\mathbf{b}_{n+1})$.

linear matched filter [1] to identify the mine-like candidates to be analyzed during the classification stage. The matched filter contains four distinct regions: pretarget, highlight, dead zone, and shadow (see Fig. 2) [1]. The filter was designed not only to match an expected mine signature and but also to suppress clutter areas simultaneously. Details on the detection algorithm are detailed in [1].

C. Feature Extraction

For each ROI, features are extracted from the associated side-scan imagery. The ith object has the respective feature vector $\mathbf{x}_i$, and for N ROIs this yields $\mathbf{X} = \{\mathbf{x}_i\}_{i=1,N}$. The authors briefly describe below the 106 features used to constitute the $\mathbf{x}_i$ employed in this study. Most of the features employed here are described in [1]. Additional features considered here include shape features, gray-level features, and cluster features.

Shape features are normally considered to characterize the appearance and specific geometry of an object. For our problem, the highlight and shadow cast by mine-like objects (Fig. 2), as opposed to nonmine-like objects, are characterized by regular shapes of anticipated dimensions. Hence, the following shapes features were extracted from the shadow and highlight: area; elongation; solidity; eccentricity; number of zero crossing of the curvature of the contour at small, medium, and large scales;

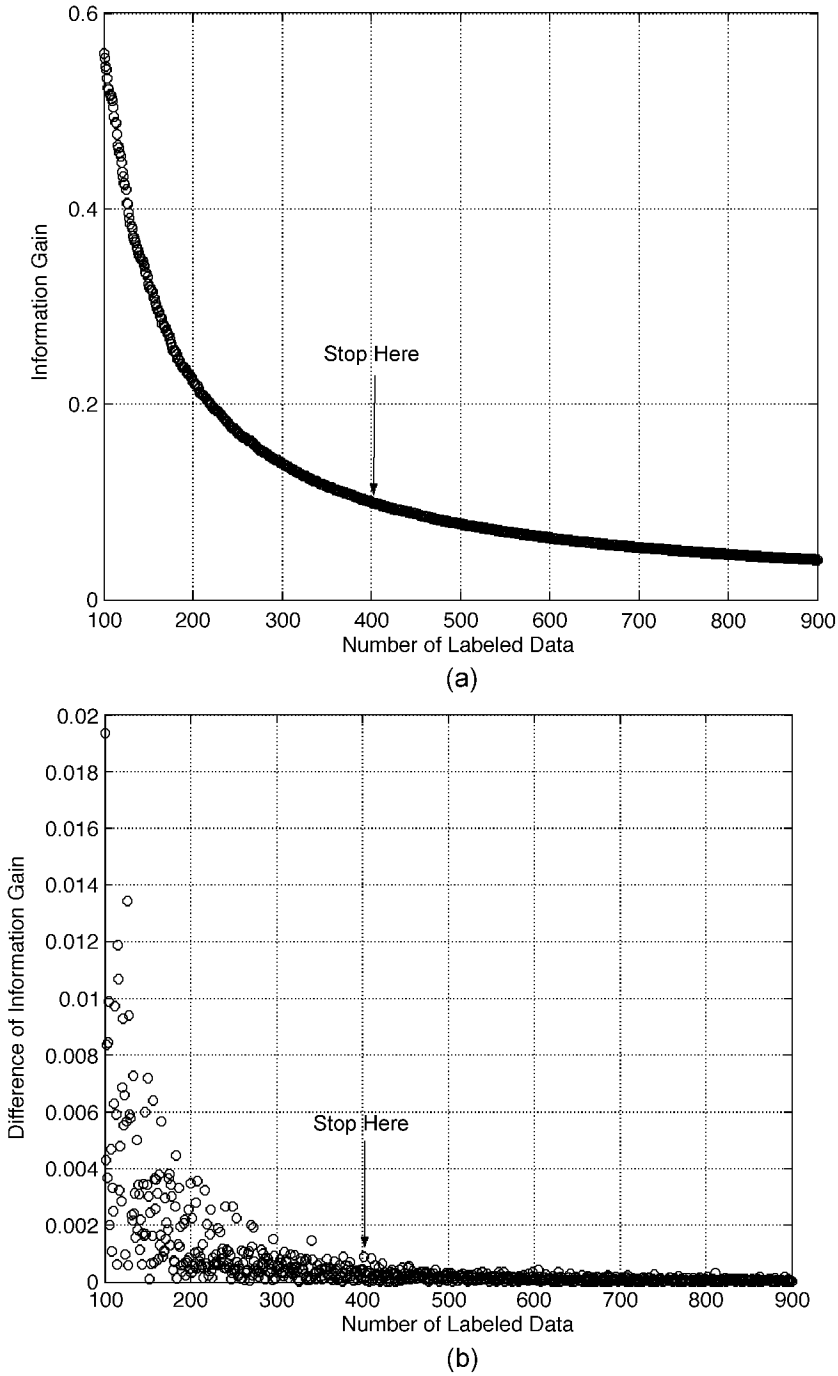

Fig. 4. Information gain as a function of the number of adaptively selected labeled signatures J. (a) $\ln \rho(\mathbf{x}_{i_{J+1}})$. (b) $\ln \rho(\mathbf{x}_{i_{J+2}}) - \ln \rho(\mathbf{x}_{i_{J+1}})$.

number of zero crossing of the radial distance of the contour; entropy of the radial distance; ratio of highlight to shadow area; ratio of highlight to shadow height; minimum distance between highlight and shadow; and horizontal alignment of shadow and highlight. In [14] and [15], one may find a detailed definition and description of these features.

When the quality and the resolution of the image are low, the side-scan sonar images may not be well characterized by the profile (shape characteristics) of the shadow and highlight. Hence, gray-level features calculated from the shadow and highlight are also used to aid in discriminating targets from clutter. The following additional features were computed: standard deviation of the highlight strength (magnitude), standard deviation of the highlight amplitude, contrast between shadow and highlight (the absolute difference of the average highlight strength and average shadow strength), contrast between shadow and background (the absolute difference of the average shadow strength and average background strength), and contrast between highlight and background (the absolute difference of average highlight strength and average background strength).

D. Active Classifier Design

The detection results are presented in the form of the ROC, quantifying the probability of detection (Pd) as a function of

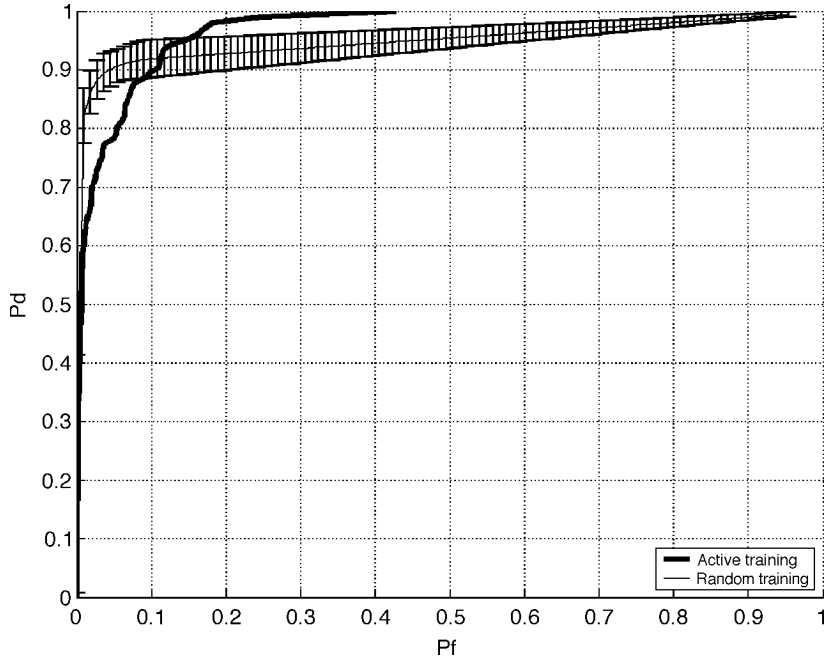

Fig. 5. ROC curves for the adaptive classifier with $J = 400$ labeled signatures and $n = 42$ basis functions. For comparison, results are shown for which an RVM [8] was employed using 50% of the data for training and 50% for testing. Results are shown for 40 random draws of the training/testing data. For the randomly generated results, the average results are presented (curve) as well as the range of variability (error bars). For the active-learning algorithm, the ROC is shown for testing on all data not selected for labeling, while for the RVM results testing is shown on 50% of the data not used for training.

the probability of false alarm (Pf), defined on $\mathbf{x} \in \mathbf{X}$ and $\mathbf{x} \notin \mathbf{X}_s$. The authors present ROC curves using the adaptive-training approach discussed in Section II, with performance compared to a conventional training algorithm. The RVM [8] is used as the comparative algorithm, it based on a form identical to (1). The distinction is that the RVM, like all conventional machine-learning algorithms, requires a distinct set of labeled training data. For the results presented here, the simplest kernel possible is considered; specifically, $K(\mathbf{x}_i, \mathbf{b}_n)$ is simply the inner product between $\mathbf{x}_i$ and $\mathbf{b}_n$.

For the experiments reported in this section and detailed in Section II-B, the authors must first select the basis functions to build the structure of the classifier. The basis functions set $\mathbf{B}_n$ are selected adaptively using the original unlabeled signatures. Specifically, the authors start with one basis element $\mathbf{b}_1$ and new basis elements $\mathbf{b}_n$ are adaptively selected until the information gain $\ln r(\mathbf{b}_{n+1}) = q_{n+1} - q_n$ is no longer significant [see (9) and (10), which are used for this computation]. The expression $\ln r(\mathbf{b}_{n+1})$ is plotted in Fig. 3 as a function of n. As shown in Fig. 3, the number of basis functions is set to $n = 42$.

Once the basis functions have been defined, the procedure in Section III-C is employed to adaptively determine the size of the desired training set $\mathbf{X}_{s,J}$ based on the information gain as J is increased. Specifically, the authors track $q_n(\mathbf{X}_{s,J})$ defined in (14) for increasing J and terminated the algorithm when the information gain is minimal. At this point, adding a new datum to the training data set did not provide significant additional information to the classifier design. The information gain $q_n(\mathbf{X}_{s,J}) - q_n(\mathbf{X}_{s,J-1})$ is plotted in Fig. 4 as a function of J. Based on the results in Fig. 4, the size of the training set is set to $J = 400$ (of which 13 are mines). Note that labels are required for $J = 400$ of the side-scan signatures, this representing 1.7% of the $N = 22\,973$ detected signatures.

E. Comparison to Traditional Approaches

As discussed above, using the active classifier design framework, $n = 42$ signatures are used as basis functions $\mathbf{B}_n$ in (1) and labels are acquired for 1.7% $(J = 400)$ of the initial detections, defining the set $(\mathbf{X}_s, \mathbf{L}_s)$. Using these parameters, the classifier is designed as described in Section II-D. It is of interest to look at the ability of the classifier to distinguish mines from nonmines (clutter) and to compare this performance to that of "conventional" approaches. For comparison, the authors train an RVM classifier, which is exactly of the form in (1). For the case of the RVM, half of the $N = 22\,973$ detection results are used for training and the other half are used for testing. The training/testing data were chosen randomly, and the curve representing RVM results in Fig. 5 represents the average performance for 40 random selections and the "error bars" associated with the RVM results represent upper- and lower-bound Pd at a given Pf. The RVM is trained as discussed in [8], with the algorithm adaptively selecting basis functions and weights from among the labeled training data.

From Fig. 5, the authors note that at low Pf (less than 0.1) the traditional RVM approach performs better than the active-learning approach, with comparable performance achieved on average for a Pf just larger than 0.1. It is also interesting to note that the active-learning approach achieves a $\mathrm{Pd} = 0.98$ (all mines correctly classified) at a Pf of approximately 0.3, with a much larger number of false alarms required of the RVM to achieve this same performance.

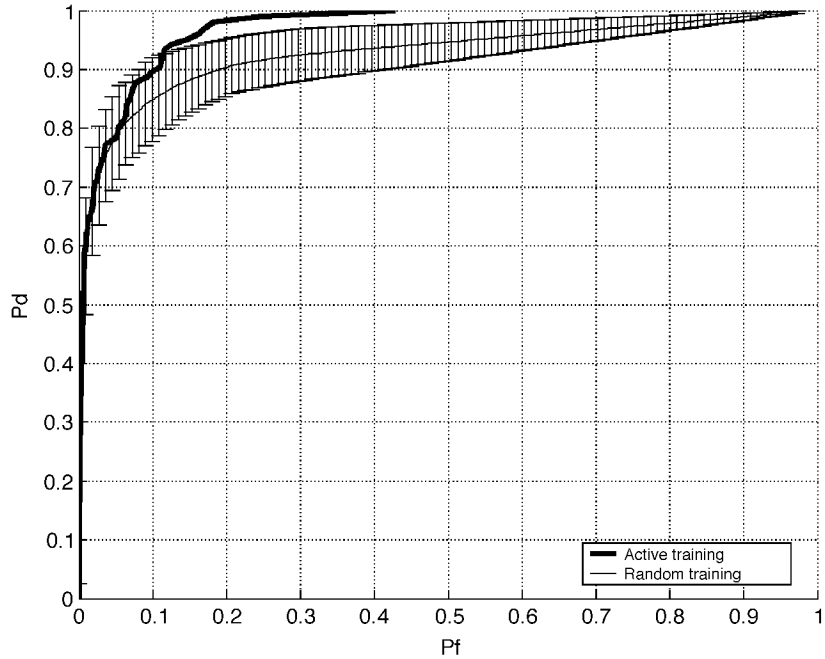

Fig. 6. As in Fig. 5, the random results are processed with 400 labeled signatures selected randomly (40 times) using the same set of basis functions as determined via the adaptive algorithm. In each of the 40 random examples, 13 of the 400 signatures corresponded to mines for direct comparison with the adaptive algorithm. For both algorithms, the ROC is shown for testing all the data not selected for labeling.

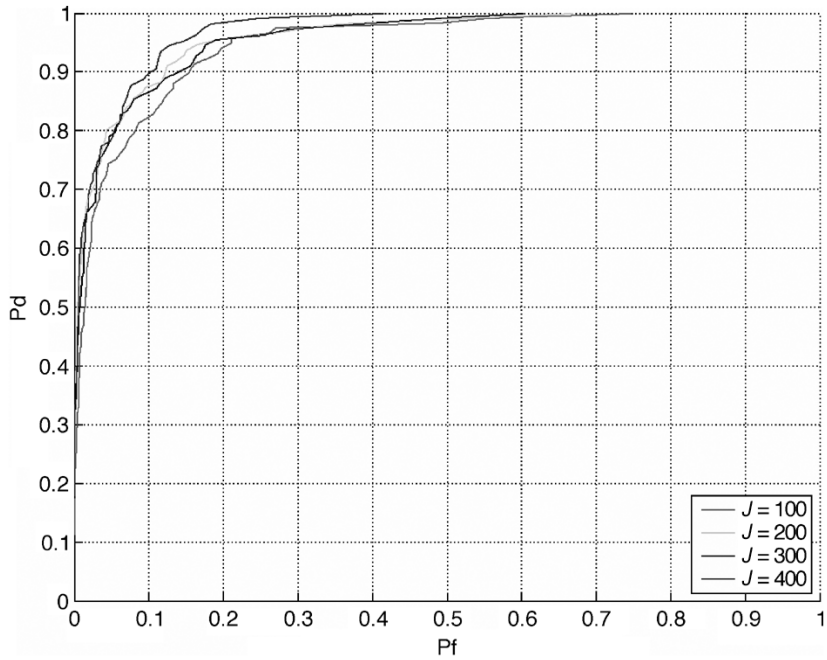

Fig. 7. Adaptive-sensing classification results as presented in Figs. 5 and 6 with $n = 42$ basis functions and now considering $J = 400$, 300, 200, and 100.

Given the basis $\mathbf{B}_n$, the authors now consider random selection of the $J = 400$ signatures for which the labels are acquired. However, the authors stipulate that 13 of these must come from the set of 119 detected mine signatures (the 13 mine signatures chosen randomly) to allow a comparison to the results of active training. Note that since $N = 22\,973$ and only 119 of the detected signatures are actually mines, if the $J = 400$ labeled signatures are defined absolutely randomly, it is unlikely that many (or any) of the signatures will correspond to mines, and therefore a classifier could not be trained. Therefore, the comparison in Fig. 6 is for illustrative purposes since the random selection of labeled signatures with $J = 400$ is unlikely in practice (due to the small number of mines relative to total detections). For the results in Fig. 6, the random selection of labeled signatures was performed 40 times, and average results are presented (curve) as well as the lower and upper bound for the Pd at a given Pf. It is clear from Fig. 6 that the active-learning results are at least as good as the average random-selection results and significantly better for $\mathrm{Pf} > 0.05$.

F. Reducing the Number of Labeled Signatures

As indicated above, requiring labels for $J = 400$ of the signatures represents less than 2% of the detected targets. Nevertheless, by reducing the size of J, one requires fewer labeled signatures, thereby improving the potential speed of mine clearance. The authors next examine ROC results as a function of J. In Fig. 7, the authors present classification results for $J = 100$, 200, 300, and 400. Note in Fig. 7 the degradation in performance with reduced J. The authors also note that performance does not necessarily improve monotonically with increasing J.

IV. Conclusion

The authors have introduced a new approach for the detection of mine targets using side-scan sonar imagery. The approach differs from most previous papers in this area [1]–[5] in two ways: i) the authors do not assume access to an *a priori* training set of labeled signatures; and ii) the authors assume access to unmanned underwater vehicles (UUVs)/divers for the determination of the binary labels (mine/no-mine) of specified signatures in the observed side-scan data. In this manner, the authors essentially build a set of labeled training data based on the observed side-scan data itself using information-theoretic concepts to quantify which signatures and associated labels would be most informative for classifier design. Using measured side-scan sonar data, the authors have demonstrated that the percentage of signatures for which labels are required is often very small ($<2\%$ in these examples), making deployment of UUVs/divers feasible. The classification performance compared favorably to "conventional" algorithm testing procedures, for which half the data are used for training and the other half for testing.

There are several directions of interest for future research. For example, in the results presented here, the classifier basis functions $\mathbf{B}_n$ and labeled signatures $(\mathbf{X}_s, \mathbf{L}_s)$ were determined "from scratch" using the observed side-scan imagery. In practice, one will likely have an available classifier of the form in (1), and it is desirable to slowly augment that classifier as new data are acquired (rather than starting classifier design over again for each new sonar scene). In this case, as new data $\mathbf{X}$ are acquired, the authors ask whether inclusion of new basis functions $\mathbf{b} \in \mathbf{X}$ adds new information and should be appended to the set of base elements in $\mathbf{B}_n$. Moreover, as new data $\mathbf{X}$ are acquired, some members of $\mathbf{B}_n$ acquired from previous data collections should be removed. Similar issues hold with respect to the labeled signatures $(\mathbf{X}_s, \mathbf{L}_s)$ because some previous labeled signatures may not be well matched to new data being observed. When acquiring labeled data, one must also account for the cost in energy and time of determining the label (e.g., with a small, near-range UUV) *vis-à-vis* the associated information gain to the classifier. These issues will be considered in future studies.

Appendix
Theoretical Justification for the Classification Algorithm

The authors have presented procedures for selecting basis functions for a kernel-based classifier based on a set of unlabeled data. After designing the basis set, the authors have also addressed selection of which signatures would be most informative for classifier training if the associated signature labels were known. In this Appendix, the authors provide theoretical justification for these design procedures.

A. Basis-Function Selection

Let the basis functions $\phi_n(\cdot)$ be evaluated for all initially unlabeled data points $\{\mathbf{x}_i\}_{i=1,N}$ and stacked to form the matrix $\tilde{\mathbf{\Phi}}_n = [\phi_n(\mathbf{x}_1), \phi_n(\mathbf{x}_2), \ldots, \phi_n(\mathbf{x}_N)]^{\mathrm{T}}$. Let the data labels be denoted $\mathbf{y} = \{\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_N\}^{\mathrm{T}}$, although these labels are not required when designing the basis functions. The difference between the true labels and those outputted by the classifier (2) for all $\{\mathbf{x}_i\}_{i=1,N}$ is expressed in vector form as

$$\begin{aligned}
&\mathbf{y} - \tilde{\mathbf{\Phi}}_n \left(\tilde{\mathbf{\Phi}}_n^{\mathrm{T}} \tilde{\mathbf{\Phi}}_n\right)^{-1} \tilde{\mathbf{\Phi}}_n^{\mathrm{T}} \mathbf{y} \\
&\quad \stackrel{1}{=} \left(\mathbf{I}_n - \tilde{\mathbf{\Phi}}_n \left(\tilde{\mathbf{\Phi}}_n^{\mathrm{T}} \tilde{\mathbf{\Phi}}_n\right)^{-1} \tilde{\mathbf{\Phi}}_n^{\mathrm{T}}\right) \mathbf{y} \\
&\quad \stackrel{2}{\approx} \left(\mathbf{I}_n - \tilde{\mathbf{\Phi}}_n \left(\lambda \mathbf{I}_{n+1} + \tilde{\mathbf{\Phi}}_n^{\mathrm{T}} \tilde{\mathbf{\Phi}}_n\right)^{-1} \tilde{\mathbf{\Phi}}_n^{\mathrm{T}}\right) \mathbf{y} \\
&\quad \stackrel{3}{=} \left(\mathbf{I}_n + \frac{1}{\lambda} \tilde{\mathbf{\Phi}}_n \tilde{\mathbf{\Phi}}_n^{\mathrm{T}}\right)^{-1} \mathbf{y}
\end{aligned} \tag{A1}$$

where $\mathbf{I}_N$ is an $N \times N$ identity matrix ($\mathbf{I}_n$ is defined similarly) and λ is a small positive number. The equality 3 in (A1) is due to the Sherman–Morrison–Woodbury formula. From (A1), the squared error between the true and estimated labels is

$$e_n^2 \approx \mathbf{y}^{\mathrm{T}} \left(\mathbf{I}_N + \frac{1}{\lambda} \tilde{\mathbf{\Phi}}_n \tilde{\mathbf{\Phi}}_n^{\mathrm{T}}\right)^{-2} \mathbf{y}. \tag{A2}$$

The expression in (A2) shows that for the given basis functions $\phi_n(\cdot)$, the authors have approximately expressed the squared error as a quadratic form of the labels $\mathbf{y}$ with a coefficient matrix $\mathbf{C}_n^{-2}$ in the form $\mathbf{C}_n = \mathbf{I}_N + \tilde{\mathbf{\Phi}}_n \tilde{\mathbf{\Phi}}_n^{\mathrm{T}}/\lambda$. The approximation can be made as accurate as desired by making λ sufficiently small. Without knowing $\mathbf{y}$, the authors prefer $\mathbf{C}_n$ to have large eigenvalues to make the error e_n^2 small. This is accomplished by making the determinant of $\mathbf{C}_n$ large. The logarithmic determinant of $\mathbf{C}_n$ is

$$\begin{aligned}
q_n^{(2)} &\stackrel{1}{=} \ln \det(\mathbf{C}_n) \\
&\stackrel{2}{=} \ln \det \left(\mathbf{I}_N + \frac{1}{\lambda} \tilde{\mathbf{\Phi}}_n \tilde{\mathbf{\Phi}}_n^{\mathrm{T}}\right) \\
&\stackrel{3}{=} \ln \frac{\det\left(\lambda \mathbf{I}_{n+1} + \tilde{\mathbf{\Phi}}_n^{\mathrm{T}} \tilde{\mathbf{\Phi}}_n\right)}{\lambda^{n+1}} \\
&\stackrel{4}{=} \ln \frac{\det(\lambda \mathbf{I}_{n+1} + \mathbf{M}_n)}{\lambda^{n+1}}
\end{aligned} \tag{A3}$$

where equality 3 is due to the property of matrix determinants and equality 4 is due to (6). Adding a new basis function to $\phi_n(\cdot)$, the authors get $\phi_{n+1}(\cdot)$ as given in (7). The logarithmic determinant of $\mathbf{C}_{n+1} = \mathbf{I}_N + \tilde{\mathbf{\Phi}}_{n+1} \tilde{\mathbf{\Phi}}_{n+1}^{\mathrm{T}}/\lambda$ is

$$q_{n+1}^{(2)} = \ln \frac{\det(\lambda \mathbf{I}_{n+2} + \mathbf{M}_{n+1})}{\lambda^{n+2}}. \tag{A4}$$

Following the method of obtaining (9) and (10), the authors can show that $q_n^{(2)}$ and $q_{n+1}^{(2)}$ are related by

$$q_{n+1}^{(2)} = \ln q_n^{(2)} + \ln \frac{r^{(2)}(\phi_{n+1})}{\lambda} \tag{A5}$$

with

$$r^{(2)}(\phi_{n+1}) = \lambda + \sum_{i=1}^{N} \phi_{n+1,i}^2 - \sum_{i=1}^{N} \phi_{n+1,i}\phi_{n,i}^{\mathrm{T}}(\lambda \mathbf{I}_{n+1} + \mathbf{M}_n)^{-1} \sum_{i=1}^{N} \phi_{n,i}\phi_{n+1,i} \tag{A6}$$

where $\phi_{n,i} \equiv \phi_n(\mathbf{x}_i)$ and $\phi_{n+1,i} \equiv \phi_{n+1}(\mathbf{x}_i)$. Since the authors wish for a $\mathbf{C}_{n+1}$ with a large determinant, they want to make $\ln(r^{(2)}(\phi_{n+1})/\lambda)$ or equivalently $\ln r^{(2)}(\phi_{n+1})$ large as λ is a constant.

Comparing (A6) to (10), the authors find that $r^{(2)}$ is approximately equal to r when λ is small. Since λ can be made as small as desired, the approximation can be made arbitrarily accurate. Therefore, the basis function obtained in (11) is the one that minimizes the determinant of $\mathbf{C}_{n+1}$ given $\mathbf{C}_n$, which in consequence will minimize the eigenvalues of $\mathbf{C}_{n+1}$, minimizing the squared error e_{n+1}^2.

B. Selection of Examples for Labeling

Assume that the basis functions $\phi_n(\cdot)$ have been selected in the manner discussed above. Moreover, assume that the authors have selected the subset $\mathbf{X}_{s,J} = \{\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \ldots, \mathbf{x}_{i_J}\}$ of J signatures for which the associated labels will be acquired. The Fisher information matrix associated with $\mathbf{X}_{s,J}$ is $\mathbf{M}_n(\mathbf{X}_{s,J}) = \sum_{k=1}^{J} \phi_{n,i_k}\phi_{n,i_k}^{\mathrm{T}}$. The Fisher information matrix for an augmented set $\mathbf{X}_{s,J} \cup \{\mathbf{x}\} = \{\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \ldots, \mathbf{x}_{i_J}, \mathbf{x}\}$ is

$$\mathbf{M}_n(\mathbf{X}_{s,J} \cup \{\mathbf{x}\}) = \mathbf{M}_n(\mathbf{X}_{s,J}) + \phi_n(\mathbf{x})\phi_n^{\mathrm{T}}(\mathbf{x}). \tag{A7}$$

Suppose the authors have the two classifiers $f_n^{\mathbf{X}_{s,J}\cup\{\mathbf{x}\}}(\cdot)$ and $f_n^{\mathbf{X}_{s,J}}(\cdot)$ that are trained using $\mathbf{X}_{s,J} \cup \{\mathbf{x}\}$ and $\mathbf{X}_{s,J}$, respectively. The authors test $f_n^{\mathbf{X}_{s,J}\cup\{\mathbf{x}\}}(\cdot)$ and $f_n^{\mathbf{X}_{s,J}}(\cdot)$ on $\mathbf{x}$ and examine how the two results are related. As given in [14, p. 121], they have

$$\left[f_n^{\mathbf{X}_{s,J}}(\mathbf{x}) - y(\mathbf{x})\right]^2 = \frac{\left[f_n^{\mathbf{X}_{s,J}\cup\{\mathbf{x}\}}(\mathbf{x}) - y(\mathbf{x})\right]^2}{1 - \phi_n^{\mathrm{T}}(\mathbf{x})\left[\mathbf{M}_n(\mathbf{X}_{s,J}) + \phi_n(\mathbf{x})\phi_n^{\mathrm{T}}(\mathbf{x})\right]^{-1}\phi_n(\mathbf{x})}. \tag{A8}$$

By using the Sherman–Morrison–Woodbury formula, they obtain

$$\left[\mathbf{M}_n(\mathbf{X}_{s,J}) + \phi_n(\mathbf{x})\phi_n^{\mathrm{T}}(\mathbf{x})\right]^{-1} = \mathbf{M}_n^{-1}(\mathbf{X}_{s,J}) - \frac{\mathbf{M}_n^{-1}(\mathbf{X}_{s,J})\phi_n(\mathbf{x})\phi_n^{\mathrm{T}}(\mathbf{x})\mathbf{M}_n^{-1}(\mathbf{X}_{s,J})}{1 + \phi_n^{\mathrm{T}}(\mathbf{x})\mathbf{M}_n^{-1}(\mathbf{X}_{s,J})\phi_n(\mathbf{x})}$$

that is used in (A8) to give

$$\left[f_n^{\mathbf{X}_{s,J}\cup\{\mathbf{x}\}}(\mathbf{x}) - y(\mathbf{x})\right]^2 = \frac{\left[f_n^{\mathbf{X}_{s,J}}(\mathbf{x}) - y(\mathbf{x})\right]^2}{1 + \phi_n^{\mathrm{T}}(\mathbf{x})\mathbf{M}_n^{-1}(\mathbf{X}_{s,J})\phi_n(\mathbf{x})}. \tag{A9}$$

Equation (A9) shows that by including $\mathbf{x}$ in the training data set, the squared test error on $\mathbf{x}$ will drop by a factor

$$\rho^{(2)}(\mathbf{x}) = 1 + \phi_n^{\mathrm{T}}(\mathbf{x})\mathbf{M}_n^{-1}(\mathbf{X}_{s,J})\phi_n(\mathbf{x}). \tag{A10}$$

If $\rho^{(2)}(\mathbf{x}) \approx 1$, the authors do not require the label for $\mathbf{x}$ as it is not important for inclusion in the training set. On the other hand, if $\rho^{(2)}(\mathbf{x}) \gg 1$, inclusion of $\mathbf{x}$ in the training set is important. Therefore, the $\mathbf{x}$ that maximizes $\rho^{(2)}(\mathbf{x})$ should be selected to seek the associated label y. Comparing (A10) to (15), the authors note that $\rho^{(2)}(\mathbf{x})$ is exactly equivalent to $\rho(\mathbf{x})$, and thus the $\mathbf{x}$ that maximizes $\rho(\mathbf{x})$ is the one that contributes maximally to make the squared test error small.

REFERENCES

[1] G. J. Dobeck, "Automated detection/classification of sea mines in sonar imagery," in *Detection and Remediation Technologies for Mines and Minelike Targets II, Proc. Int. Society Optical Engineering (SPIE)*, vol. 3079, Orlando, FL, 1997, pp. 90–110.

[2] A. R. Castellano and B. C. Gray, "Autonomous interpretation of side scan sonar returns," in *Proc. Symp. Autonomous Underwater Vehicle Technology*, Washington, DC, 1990, pp. 248–253.

[3] P. F. Schweizer and W. J. Petlevich, "Automatic target detection and cuing system for an autonomous underwater vehicle," in *Proc. Int. Symp. Unmanned Untethered Submersible Technology*, Elliott City, MD, 1989, pp. 359–371.

[4] G. J. Dobeck, "Algorithm fusion for automated sea mine detection and classification," in *Proc. IEEE OCEANS Conf.*, Honolulu, HI, 2001, pp. 5–8.

[5] J. C. Delvigne, "Shadow classification using neural networks," in *Proc. Undersea Defense Conf.*, London, U.K., 1992, pp. 214–221.

[6] B. Schölkopf and A. Smola, *Learning With Kernels, Support Vector Machines, Regularization, Optimization and Beyond*. Cambridge, MA: MIT Press, 2002.

[7] B. Schölkopf, K.-K. Sung, C. J. C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, "Comparing support vector machines with Gaussian kernels to radial basis function classifiers," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2758–2765, Nov. 1997.

[8] M. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, no. 3, pp. 211–244, 2001.

[9] V. V. Fedorov, *Theory of Optimal Experiments*. New York: Academic, 1972.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[11] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston, MA: Kluwer, 1991.

[12] T. Kohonen, "Learning vector quantization," *Neural Netw.*, vol. 1, no. 3, pp. 303–315, 1988.

[13] I. T. Jolliffe, *Principal Component Analysis*. New York: Springer-Verlag, 1986.

[14] L. Costa and M. Cesar, *Shape Analysis and Classification*. Boca Raton, FL: CRC Press, 1998.

[15] S. Theodoris and K. Koutroumbas, *Pattern Recognition*. San Diego, CA: Academic, 1999.

Esther Dura received the M.Eng. degree in computer science from the Universidad de Valencia, Spain, in 1998, the M.Sc. degree in artificial intelligence from The University of Edinburgh, U.K., in 1999, and the Ph.D. degree from Heriot-Watt University, Edinburg, Scotland, in 2002. Her thesis examined the classification and reconstruction of mine-like objects and seafloors from side-scan sonar images.

She is currently a Postdoctoral Research Associate at Duke University, Durham, NC. Her current fields of interest include underwater image processing, pattern recognition, robotics, and machine learning for remote sensing applications.

Yan Zhang received the B.S., M.S., and Ph.D. degrees in electrical engineering from Jilin University of Technology, China, in 1993, 1996, and 1998, respectively.

From January 1999 to July 2004, he was a Postdoctoral Researcher in the Department of Electrical and Computer Engineering, Duke University, Durham, NC. In 2004, he joined Innovation Center of Humana Inc., Louisville, KY, as a Research Scientist. His research interests include statistical signal processing, pattern recognition, and their applications. His current research activity focuses on subsurface target detection and predictive modeling.

Gerald J. Dobeck received the B.S. degree in physics from the University of Massachusetts, Amherst, in 1970, and the M.S. and Ph.D. degrees in electrical engineering from the University of South Florida, Tampa, in 1973 and 1976, respectively.

Since 1976, he has been with the Dahlgren Division, Naval Surface Warfare Center (NSWC) Coastal Systems Station, Panama City, FL. His current research interests include automatic detection and classification of targets in cluttered environments from synthetic/real aperture sonar imagery, the echo structure of acoustic returns, underwater electrooptic imagery, and gradiometer/magnetometer signals. Over the past 14 years, he is Project Leader for the development and application of processing technologies for 1) automated mine detection and classification and 2) data/sensor/algorithm fusion sponsored by the Office of Naval Research 6.2/6.3 Mine Countermeasures program. He pioneered the development of algorithm fusion, the fusion of multiple expert computer-aided detection and classification (CAD/CAC) algorithms. His CAD/CAC team has developed PMA2000, a mine-detection post-mission analysis software tool for the analysis of high-resolution side-looking sonar imagery. He has led the development of real-time sea mine detection/classification for Navy autonomous underwater vehicles. He has authored or coauthored more than 90 technical reports and papers.

Dr. Dobeck received the 1981 and 1996 Commanding Officer/Executive Director award for Science and Technology. In 2000, he received the National Defense Industrial Association Bronze Metal Award. He is a Reviewer for the IEEE, the American Society of Mechanical Engineers (ASME), The International Society for Optical Engineers (SPIE), and the *Journal of Underwater Acoustics*, and has been Session Chair at past IEEE and SPIE conferences.

Xuejun Liao was born in Qinghai, China. He received the B.S. and M.S. degrees in electrical engineering from Hunan University, China, in 1990 and 1993, respectively, and the Ph.D. degree in electrical engineering from Xidian University, China, in 1999.

From 1993 to 1995, he was with the Department of Electrical Engineering, Hunan University, working on electronic instruments. From 1995 to 2000, he was with the National Key Lab for Radar Signal Processing, Xidian University, working on automatic target recognition (ATR) and radar imaging. Since May 2000, he has been working as a Research Associate with the Department of Electrical and Computer Engineering, Duke University, Durham, NC. His current research interests are in Markovian techniques in ATR, blind source separation, sensor scheduling, and machine learning.

Lawrence Carin (SM'96–F'01) was born in Washington, DC, on March 25, 1963. He received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, in 1985, 1986, and 1989, respectively.

He joined the Electrical Engineering Department, Polytechnic University, Brooklyn, NY, as an Assistant Professor in 1989 and became an Associate Professor in 1994. In September 1995, he joined the Electrical Engineering Department, Duke University, Durham, NC, where he is now the William H. Younger Professor of Engineering. He was the Principal Investigator (PI) on a Multidisciplinary University Research Initiative (MURI) on demining (1996–2001) and is the current PI of a MURI dedicated to multimodal inversion. His current research interests include short-pulse scattering, subsurface sensing, and wave-based signal processing.

Dr. Carin was an Associate Editor of the IEEE TRANSACTIONS ON ANTENNAS AND PROPAGATION from 1995 to 2004. He is a Member of the Tau Beta Pi and Eta Kappa Nu honor societies.