Target Discrimination in Synthetic Aperture Radar Using Artificial Neural Networks

José C. Principe, *Senior Member, IEEE,* Munchurl Kim, and John W. Fisher, III, *Member, IEEE*

***Abstract*— This paper addresses target discrimination in synthetic aperture radar (SAR) imagery using linear and nonlinear adaptive networks. Neural networks are extensively used for pattern classification but here the goal is discrimination. We will show that the two applications require different cost functions. We start by analyzing with a pattern recognition perspective the two-parameter constant false alarm rate (CFAR) detector which is widely utilized as a target detector in SAR. Then we generalize its principle to construct the quadratic gamma discriminator (QGD), a nonparametrically trained classifier based on local image intensity. The linear processing element of the QGD is further extended with nonlinearities yielding a multilayer perceptron (MLP) which we call the NL-QGD (nonlinear QGD).**

MLP's are normally trained based on the L_2 norm. We experimentally show that the L_2 norm is not recommended to train MLP's for discriminating targets in SAR. Inspired by the Neyman–Pearson criterion, we create a cost function based on a mixed norm to weight the false alarms and the missed detections differently. Mixed norms can easily be incorporated into the backpropagation algorithm, and lead to better performance. Several other norms (L_8, cross-entropy) are applied to train the NL-QGD and all outperformed the L_2 norm when validated by receiver operating characteristics (ROC) curves. The data sets are constructed from TABILS 24 ISAR targets embedded in 7 km$_2$ of SAR imagery (MIT/LL mission 90).

***Index Terms*— Gamma kernels, mixed norm training, neural networks, synthetic aperture radar, target discrimination.**

I. Introduction

AUTOMATIC target recognition (ATR) seeks to classify specific targets in natural environments (background clutter). The architecture of ATR systems is normally divided into a focus of attention block followed by a classifier for practical reasons [21]. The computational bandwidth required to directly classify each image pixel is prohibitive and the method is wasteful since most of the imagery is background clutter. Hence, the focus of attention stage selects areas with a large probability of containing targets. Only these areas are further scrutinized by the classifier. The focus of attention must therefore be a computationally simple system that provides a high probability of detection creating at the same time as few false alarms as possible. A false alarm is a system detection that corresponds to clutter. The price/performance ratio of the full ATR system is therefore very much dependent upon the figure of merit of the focus of attention.

The focus of attention is built from a detector (prescreener) enhanced by a discriminator to reduce further the number of false alarms. The most popular prescreener utilized in synthetic aperture radar (SAR) is the two parameter constant false alarm (CFAR) detector [17], which is based on a normalized test of the pixel intensity versus its local neighborhood. Its popularity in millimeter SAR is due to an excellent compromise simplicity/performance. However, the discriminating power of the two parameter CFAR is not enough to reduce the false alarms to acceptable levels. The focus of attention subsystems described in the literature incorporate complementary features and discriminants such as fractal dimension, weighted rank fill ratio, size, polarimetric purity, etc. [17], [21]. Acceptable performance has been reported [19], but the systems can be further improved.

In this paper, we briefly describe a focus of attention based on adaptive systems concepts. Instead of extending the CFAR with unrelated features and other discriminators, we optimize the utilization of local pixel intensity features extracted by a modified CFAR detector, the γCFAR [14]. We start by analyzing the two-parameter CFAR detector in a pattern recognition perspective and generalize its structure to the quadratic gamma discriminator (QGD), which is able to create a quadratic discriminant function in the pixel intensity space. The QGD was implemented as a linear machine [24], i.e. as a weighted combination of all the quadratic terms of the intensity of the pixel under analysis and the intensity of a neighboring region. Least squares was used to adapt the weights of the QGD. The combination CFAR/QGD has been tested by MIT/LL on real target data and showed a substantial improvement with respect to the CFAR performance [25].

The purpose of this paper is to report on further enhancements to the QGD. Extending the linear processing element (PE) of the QGD with a saturating nonlinearity provides the basis for a multilayer perceptron (MLP) topology, which we called the nonlinear QGD (NL-QGD). Training the NL-QGD with backpropagation and the mean square error criterion (L_2 norm) yielded disappointing performance. An analysis of the Neyman–Pearson criterion, the most appropriate hypothesis test for real world detection problems, led us to develop a mixed norm approach to train the NL-QGD, which improves the performance further. Several other norms were also implemented and tested. We present receiver operating

Manuscript received November 21, 1996; revised December 10, 1997. This work was supported in part by DARPA under Contracts N60921-93-C-A335 and AFOSR F-33615-97-1-1019.

J. C. Principe is with the Computational NeuroEngineering Laboratory, Electrical and Computer Engineering Department, University of Florida, Gainesville, FL 32611 USA (e-mail: principe@cnel.ufl.edu).

M. Kim is with the Electronics and Telecommunication Research Institute, Taejon, Korea.

J. W. Fisher, III, is with the Massachusetts Institute of Technology Artificial Intelligence Laboratory, Cambridge, MA 02139 USA.

Publisher Item Identifier S 1057-7149(98)05310-X.

characteristics (ROC's) for each of the discriminators utilizing a large data set of SAR imagery (about 7 km^2) provided by MIT/LL (mission 90 pass 5) with embedded targets. Unsupervised neural networks have been applied to SAR image recognition for segmentation and surface representation [2], [8] and as feature extractors [16]. In fact, the gamma kernels resemble the Mexican hat function but they have the advantage of independently setting the "inhibition region" [27]. Here, neural networks will be trained in a supervised framework, very much like the work of [5], [22], and [26]. In the context of our work, neural networks appear as a natural extension to improve performance for the focus of attention as will be described in the paper.

II. Detection and the Neyman-Pearson Criterion

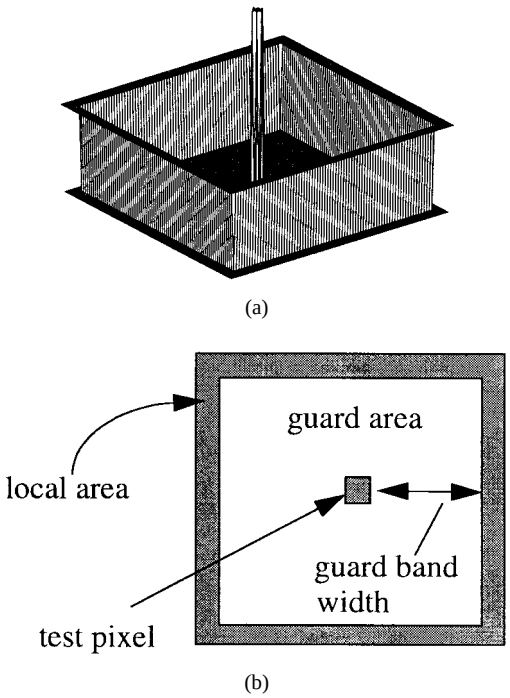
Signal detection in background noise can be formulated through the Bayes' theory of hypothesis testing. The hypothesis here are two: H_0 that corresponds to no signal present (only noise), and H_1 where the signal and noise are both present. The optimum detector (decision) according to Bayes theory minimizes the average risk, which implies knowledge of the costs of decisions, the prior probabilities of each hypothesis as well as the probability density function of both the signal $(p_1(x))$ and noise $(p_0(x))$. In realistic detection environments the knowledge of the decision costs and the *a priori* probabilities are not available. Under these circumstances, Neyman and Pearson [18] proposed a criterion that maximizes H_1, the probability of detection, given a constraint imposed on the probability of false alarms, as follows:

$$\max P_D(\delta) \quad \text{subject to} \quad P_F(\delta) = \alpha \tag{1}$$

where P_D is the probability of detection for the decision rule δ, and P_F stands for the probability of false alarms, which we constrain to be equal to α. Notice the different role played by the missed detections $(1 - P_D)$ and the false alarms.

The Neyman–Pearson criterion can be easily implemented by computing the likelihood ratio $\Gamma(x) = p_1(x)/p_0(x)$ as a function of x and comparing the outcome to a fixed threshold Γ_{th}. Hypothesis H_0 is chosen if $\Gamma(x) < \Gamma_{\mathrm{th}}$, otherwise H_1 is selected. Γ_{th} is chosen such that the false alarm rate α is obtained. The simplicity of the Neyman–Pearson criterion and its relationship with Bayes criterion makes it the preferred criterion in signal detection [13].

In many practical situations, knowledge of $p_1(x)$ and $p_0(x)$ is also not available or difficult to model, which leads to the construction of nonparametric detectors. In this case one should try to achieve only a specified false alarm rate, relaxing the requirement of maximum probability of detection. A detector whose false alarm rate is unchanged for all noise distributions in a given class is called constant false alarm rate (CFAR). A rather important case for detection is obtained when the signal amplitude is positive but unknown and the noise is assumed Gaussian distributed with zero mean and unknown variance. Under this condition and using the principle of maximum likelihood, the likelihood ratio test is equivalent to a *Student t statistics* [10] which leads to the following decision rule



(a)

(b)

Fig. 1. The CFAR detector. The amplitude of the test pixel is compared with the mean and the standard deviation of the local area. The guard area ensures that no target pixels are included in the measurement of the local statistics.

$$\frac{\overline{x}}{\hat{\sigma}_x} \underset{H_0}{\overset{H_1}{\gtrless}} Th \tag{2}$$

where $\overline{x}$ is the sampled signal mean and σ_x is the sampled standard deviation. For positive amplitude signals, hypothesis H_1 is chosen if the ratio is above a threshold Th which sets the false alarm rate. The asymptotic relative efficiency of this test is the same as the Neyman–Pearson criterion when the number of samples used to estimate the mean and variance of the noise tend to infinity.

These ideas were applied to radar signal detection by Finn [7], who showed that a simple comparison of the cell under test with the power estimated in surrounding cells produced a CFAR detector for Gaussian clutter (background noise). Goldstein [10] showed that the following test:

$$\frac{X_t - \overline{X}_c}{\hat{\sigma}_c} \underset{\text{clutter}}{\overset{\text{target}}{\gtrless}} T_{\mathrm{CFAR}} \tag{3}$$

where X_t is a pixel under test and $\overline{X}_c$ and $\hat{\sigma}_c$ are the estimates of the local clutter mean and standard deviation, produced a CFAR detector for log-normal and Weibull distributed clutter. Novak adapted this test for synthetic aperture radar (SAR) [3] and called it the two parameter CFAR detector. The area where the clutter statistics are estimated is defined by a stencil as shown in Fig. 1.

The CFAR compares the intensity of a pixel under test with the normalized intensity of a surrounding area. Since man-made objects are normally bright in millimeter SAR imagery, this is a very effective test, which can be efficiently implemented in digital hardware. The shape of the stencil

ensures that when the center pixel is on target, the neighborhood falls in the background such that its local statistics can be reasonably well estimated. The shape of the stencil (in particular the guardband) is governed by the target size [21]. In SAR imagery, the reflectivity of the object is only weakly coupled to its geometric shape, so *a priori* stencil dimensions based solely on target size cause suboptimal performance [25].

A. Statistical Pattern Recognition Interpretation of the CFAR Detector

In terms of statistical pattern recognition, we can interpret the two parameter CFAR detector as a feature extractor followed by a discriminant function. One can model the CFAR stencil as a local *intensity feature extractor* in the neighborhood of the pixel under test. The CFAR detection rule is combining the features to produce a decision between the classes of targets and clutter, i.e., it is effectively creating a *discriminant function for a two class problem*. The CFAR equation (3) can be rewritten as

$$X_t^2 - 2X_t\overline{X}_c + \overline{X}_c^2 - T_{\mathrm{CFAR}}^2\overline{X_c^2} + T_{\mathrm{CFAR}}^2\overline{X}_c^2 \gtrless 0 \quad (4)$$

where X_t is a pixel under test, T_{CFAR} is the threshold for the CFAR detector, and $\overline{X}_c, \overline{X_c^2}, \overline{X}_c^2$ are the estimates for the mean, power and mean square intensity measured in a neighboring area defined in the CFAR stencil. Hence, we can interpret the two parameter CFAR as implementing a *restricted* and *anchored* quadratic discriminant function of the local image intensity. It is restricted because some terms of the quadratic classifier are missing, and anchored because the discriminant function has fixed parameters so it can not be moved in pattern space for optimal performance.

If the goal is detection, (4) provides an answer with the interesting properties of constant false alarm rate for certain probability density functions (pdf's). But can we use the same intensity features to decrease further the number of false detections, i.e. use the same features to build a discriminator? From the perspective of an optimal discriminator that is utilizing local intensity features, we can immediately enumerate shortcomings in the two parameter CFAR detector: i) it uses only some of the quadratic terms of the pixel intensity and its surroundings; ii) it implements a fixed parametric combination of these features; iii) there is little flexibility in the feature extraction because the stencil size is chosen in an *ad-hoc* manner, particularly for SAR. These three aspects can be greatly improved if more mathematically oriented local projection operators are chosen and if trainable classifiers are built. We will briefly review the gamma kernels as alternatives to the CFAR stencil to improve iii), and will introduce the QGD (quadratic gamma discriminator) as an improvement to i) and ii). A more detailed analysis of the QGD can be found in [24] and [25].

III. Gamma Kernels

The gamma delay operator was originally developed for time series analysis [6]. The goal was to create a signal processing structure that would have a variable memory depth for a fixed number of stages (taps), i.e., which would be able to implement a linear warping of the time axis for performance improvements. This was accomplished by introducing a local feedback loop around the delay operator, creating a generalized feedforward filter structure [23].

The concept of a time warping parameter extrapolates to the spatial domain as a scale parameter that controls the region of support of the two-dimensional (2-D) gamma stencil. So, we define the 2-D gamma kernels $g_{k,\mu}(n_1, n_2)$ of order k as

$$g_{k,\mu}(n_1, n_2) = C g_{k,\mu}(t)|_{t=\sqrt{n_1^2+n_2^2}} \quad (5)$$

where the constant C is a normalization factor, μ is the scale parameter, and n_1 and n_2 the two space variables. The resulting 2-D gamma kernels have circularly symmetric shapes given by

$$g_{k,\mu}(n_1, n_2) = \frac{\mu^{k+1}}{2\pi k!}(\sqrt{n_1^2+n_2^2})^{k-1} e^{-\mu\sqrt{n_1^2+n_2^2}}$$
$$\Omega = \{(n_1, n_2); -N \le n_1, n_2 \ge N\} \quad (6)$$

where Ω is the region of support of the kernel, k the kernel order, and μ the parameter that controls the shape and scale of the kernel. Fig. 2 depicts the characteristics of 2-D gamma kernels in the spatial domain. The 1st order kernel $(k = 1)$ has its peak at the pivot point (0, 0) with an exponentially decaying amplitude. The gamma kernels with a higher order $(k > 1)$ have peaks at the radius k/μ, creating concentric smooth rings around the pivot point. For a fixed kernel order, the radial distances where the kernels peak are still dependent upon the parameter μ, as in the one-dimensional (1-D) case [6].

By analogy to the CFAR stencil, any combination of the first gamma kernel with one of the higher order kernels produces a CFAR-like stencil, although the shapes with the 2-D gamma kernels are smoother (Fig. 3). We call the CFAR detector with this stencil the γCFAR. With the γCFAR, we have a better handle on the shape of the stencil due to the analytic formalism that depends on a single parameter per kernel. In fact, after fixing the order of a gamma kernel, we have a single parameter that controls its spatial extent, and we can derive equations that will adapt the parameter μ to minimize the output error in the training set.

IV. Quadratic Gamma Discriminator (QGD)

Quadratic discriminant functions implement the optimal classifier for Gaussian distributed classes [6]. A quadratic discriminant function $q(X)$ in d-dimensional space is

$$q(X) = \sum_{j=1}^{d} w_{jj}x_j^2 + \sum_{j=1}^{d-1}\sum_{k=j+1}^{d} w_{jk}x_j x_k + \sum_{j=1}^{d} w_j x_j + w_{d+1} \quad (7)$$

where w_{ij} is a set of adjustable parameters. One well-established construction of a quadratic classifier is to create a quadratic preprocessor that creates all the terms of $q(X)$, followed by a linear machine which simply weights each one of these terms [9].

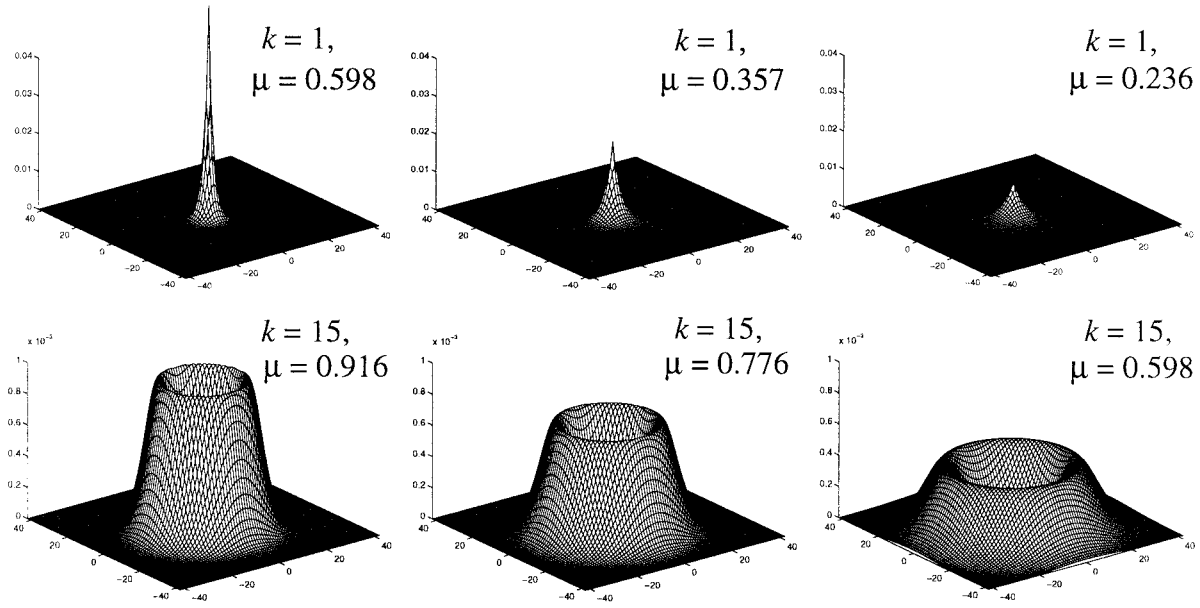
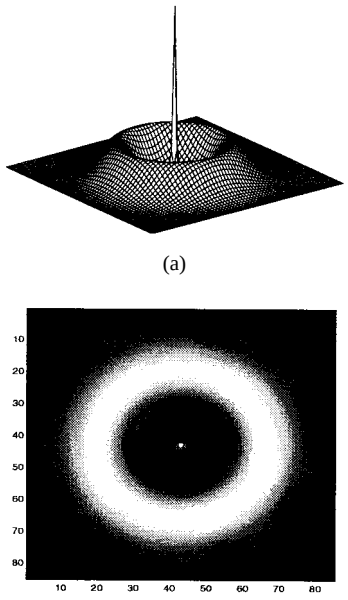


Fig. 2. Two-dimensional gamma kernels $(k = 1, k = 15)$ for different values of the scale parameter μ.



(a)

(b)

Fig. 3. γCFAR detector. The center kernel has an order of one, and the rounding kernel is of an order 15. The rounding kernel defines a local area where the local statistics of mean and standard deviation are measured. The peaky kernel averages a pixel under test and the very closely neighbored pixels around a pixel under test.

We compute the four basic features $X_t, X_t^2, \overline{X}_c, X_c^2$: (4) with two gamma kernels such that

$$g_{k,\mu} \cdot X^p = \sum_k \sum_l g_{k,\mu}(n_1, n_2) x^p(n_1, n_2) \qquad p = 1 \text{ or } 2 \tag{8}$$

where $\cdot$ stands for the convolution operator. The two input features are the intensities at the pixel under test $g_1 \cdot X$ and the intensity in the ring neighborhood $g_k \cdot X$ with $k > 1$. From these quantities the traditional quadratic discriminator creates $(g_1 \cdot X)^2, (g_k \cdot X)^2, g_1 \cdot X, g_k \cdot X, (g_1 \cdot X)(g_k \cdot X)$, and the bias. For a direct comparison with the CFAR we will add $g_1 \cdot X^2$ and $g_k \cdot X^2$ for a total of seven features and a bias. The complete feature vector reads

$$\begin{aligned} \boldsymbol{F}_{\mu_1,\mu_n} = [&g_{1,\mu_m} \cdot X \quad g_{k,\mu_n} \cdot X \quad g_{1,\mu_m} \cdot X^2 \quad g_{k,\mu} \cdot X^2 \\ &(g_{1,\mu_m} \cdot X)^2 \quad (g_{k,\mu_n} \cdot X)^2 \\ &(g_{1,\mu_m} \cdot X)(g_{k,\mu_n} \cdot X) \quad 1]^T \end{aligned} \tag{9}$$

and the quadratic discriminator becomes

$$y = \boldsymbol{W}^T \boldsymbol{F}_{\mu_1,\mu_n} \tag{10}$$

where

$$\boldsymbol{W} = [w_1 \quad w_2 \quad w_3 \quad w_4 \quad w_5 \quad w_6 \quad w_7 \quad w_8]^T \tag{11}$$

is the parameter vector. Fig. 4 depicts the implementation of the quadratic gamma discriminator (QGD).

With this formulation, we can better understand what was said previously regarding the restricted nature of the two parameter CFAR detector when seen from a pattern recognition stand point. The parameter vector for the γCFAR (or equivalently for the two parameter CFAR) is

$$\boldsymbol{W} = [0 \quad 0 \quad 0 \quad -T^2 \quad 1 \quad 1+T^2 \quad -2 \quad 0]^T \tag{12}$$

where some of the parameters are set to zero and others are fixed. Since the increase of the number of free parameters of a system is coupled with more flexibility, and adapting the parameters achieves optimal performance, we can improve the γCFAR performance by creating more parameters and adapting them with representative data.

A. Training the QGD

In an adaptive pattern recognition framework, the free parameters of the classifiers which define the positioning of the discriminant function for maximum performance are

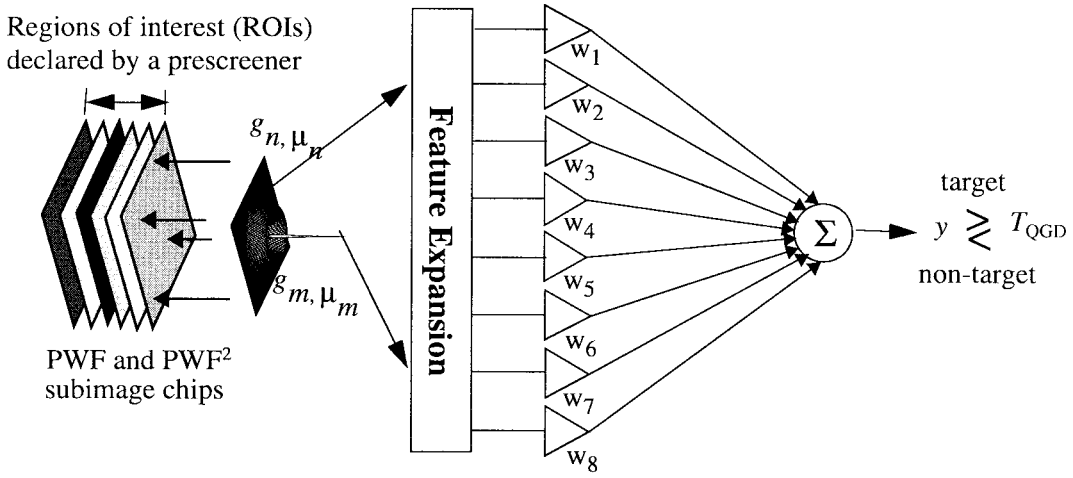


Fig. 4. Quadratic gamma discriminator.

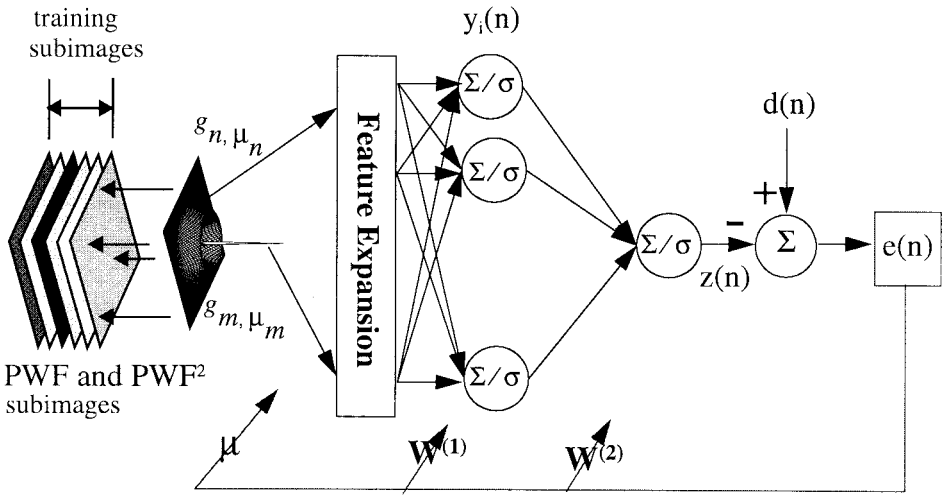


Fig. 5. Implementation of the NL-QGD and its adaptation.

learned from a set of representative data. This step is called the *training of the classifier*. Given a set of training image chips $\{X_1, X_2, \cdots, X_N\}$ centered around points of a known class, we compute the corresponding feature vector F. The corresponding desired values of the image chips are one's for target class and zero's for clutter, which construct a desired vector $d = \{d_1, d_2, \cdots, d_N\}^T$.

If the mean square error between the system output and the desired response is selected as the cost function, there is an analytical solution to the problem [12]. The method solves in the least square sense an overdetermined (assuming $N > 8$) system of linear equations for the unknown coefficient vector W.

$$\min_W ||\boldsymbol{d} - \boldsymbol{FW}||_2 \tag{13}$$

yielding

$$\boldsymbol{W} = (\boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T\boldsymbol{d}. \tag{14}$$

Since the feature vector $\boldsymbol{F}$ is a function of the parameters μ the problem we are facing is in fact a parametric least square, which does not have a closed form solution due to the nonlinear dependence on the parameter. There are two possibilities of solving this problem. Either we first determine the best value of μ_1 and μ_n for prescreening and use here the same values, or we have to use an iterative approach to find both the weight vector and μ_1 and μ_n. The results we will show later utilize the first method. The second method will be utilized in the following section to train the nonlinear QGD.

V. Nonlinear Extension of the QGD (NL-QGD)

In the context of our work the MLP appears as a natural extension of the processing power of the QGD. In fact, an MLP can be thought of as a nonlinear combination of QGD's (Fig. 5). Since the MLP is capable of creating arbitrary discriminant functions [12] this extension has the potential to improve performance. Note that the QGD creates a quadratic discriminant function of the image intensity, which is optimal only for Gaussian pdf's [9]. Fig. 5 displays the block diagram of the NL-QGD. We can think of the QGD as the linear part of one of the hidden layer processing elements. The three basic issues that need to be addressed when using a neural network are the training rules, the cost function and the network topology.

A. Training the NL-QGD

In order to fully develop the neural network approach, an iterative learning scheme is required to adapt the weights and the parameter μ. Availability of on-line methods of adapting μ effectively means that the system output error can be used to search for the optimal local area and guard band of the training set data.

The NL-QGD is trained with a desired signal d (one's for the target class and zeroes for the nontarget class) using the backpropagation (BP) algorithm [12]. The sum of squared errors is initially utilized here, i.e.,

$$\min_{W}\left(E(n)=\frac{1}{2N}\sum_{n=1}^{N}||d(n)-z(n)||^2\right). \tag{15}$$

Moreover, in order to adapt the parameter m which controls the scale of the γCFAR stencil, the error generated at the detector output is backpropagated up to the input layer. The decision boundary of the NL-QGD is therefore formed including the parameter μ. The weights W are adapted applying directly the BP algorithm. The parameter μ is recursive, so we present here its learning equations. The correction $\Delta\mu(n)$ at each iteration is proportional to the instantaneous gradient $\partial E(n)/\partial\mu(n)$. According to the chain rule, this gradient is expressed as follows:

$$\frac{\partial E(n)}{\partial\mu(n)}=\sum_{p}\frac{\partial E(n)}{\partial y_p(n)}\frac{\partial y_p(n)}{\partial\mu(n)}=\Delta_P^T(n)\frac{\partial \boldsymbol{F}(n)}{\partial\mu(n)} \tag{16}$$

where the input to the network is the feature expansion F given by (9) at each iteration n $F(n)=Y_P(n)=[y_1(n),y_2(n),\cdots,y_P(n)]^T$ and $y_i(n)$ is the input to each nonlinearity (with $P=8$)

$$y_i(n)=\sum_{p=1}w_{ip}(n)\boldsymbol{F}(n). \tag{17}$$

The local gradient vector at the input layer is $\Delta_P(n)=[\delta_1(n),\delta_2(n),\cdots,\delta_P(n)]^T$. The local gradient at the pth PE in the input layer of the MLP is obtained by

$$\delta_p(n)=-\frac{\partial E(n)}{\partial y_p(n)}=\sum_i\delta_i(n)w_{ip}(n) \tag{18}$$

where $\delta_i(n)$ is the local gradient at the ith PE in the first hidden layer and $w_{ip}(n)$ is a weight between the pth PE in the input layer and the ith PE in the first hidden layer.

The gradient $\partial\boldsymbol{F}(n)/\partial\mu(n)$ is given by

$$\begin{aligned}\frac{\partial F_{\mu_1\mu_{15}}}{\partial\mu_1}=\Bigg[&\frac{\partial g_1}{\partial\mu_1}\cdot X\quad 0\quad \frac{\partial g_1}{\partial\mu_1}\cdot X^2\quad 0\\ &2(g_1\cdot X)\frac{\partial}{\partial\mu_1}(g_1\cdot X)\quad 0\\ &2(g_{15}\cdot X)\frac{\partial}{\partial\mu_1}(g_1\cdot X)\quad 0\Bigg]^T\\ \frac{\partial F_{\mu_1\mu_{15}}}{\partial\mu_{15}}=\Bigg[&\frac{\partial g_{15}}{\partial\mu_{15}}\cdot X\quad 0\quad \frac{\partial g_{15}}{\partial\mu_{15}}\cdot X^2\quad 0\\ &2(g_{15}\cdot X)\frac{\partial}{\partial\mu_{15}}(g_{15}\cdot X)\quad 0\\ &2(g_1\cdot X)\frac{\partial}{\partial\mu_{15}}(g_{15}\cdot X)\quad 0\Bigg]^T.\end{aligned} \tag{19}$$

The derivative of $g(\cdot)$ with respect to μ is

$$\begin{aligned}&\frac{\partial}{\partial\mu}g_{k,\mu}(n_1,n_2)\\ &\quad=\frac{(k+1)}{\mu}\left(g_{k,\mu}(n_1,n_2)-g_{k+1,\mu}(n_1,n_2)\right).\end{aligned} \tag{20}$$

The sample by sample adaptation of the parameter μ is therefore given by

$$\begin{aligned}\mu_j(n+1)&=\mu_j(n)-\beta\frac{\partial E(n)}{\partial\mu_j(n)}\\ &=\mu_j(n)+\beta\Delta_P^T(n)\frac{\partial}{\partial\mu_j(n)}F_{\mu_1,\mu_{15}}(n)\end{aligned} \tag{21}$$

where β is the step size, and $j=1,15$ (one equation for each kernel).

VI. The Need for a Different Performance Criterion to Train the NL-QGD

Preliminary comparison between the QGD and the NL-QGD trained with the L_2 norm demonstrated that the QGD was a better discriminator, although its final mean square error (MSE) was much higher than that of the NL-QGD (see Table IV). This was an unexpected result.

We analyzed the NL-QGD outputs to understand the source of the problem. Fig. 8(b) shows the test set responses of the QGD (dotted line) and of the NL-QGD (solid line) in a test set built from 3905 clutter chips and 345 target chips. For display purposes, the clutter chips were presented sequentially first followed by the target chips. So the first 3905 values in Fig. 8(b) correspond to clutter responses (a desired response of zero) followed by target responses (desired response of one). The first observation is that the output of the NL-QGD is in the average much closer to the target values of zero and one than the QGD. This is to be expected due to the inclusion of the nonlinearity in the input–output map. What was unexpected was the behavior of the nonlinear system when the input was misclassified. Notice that when the NL-QGD makes a mistake, the error can be very large, in fact much larger than that of the QGD.

For a discriminator this characteristic is unacceptable for the following reason: According to the Neyman–Pearson criterion, we set a threshold value that provides a given detection performance, let us say 100% detection ($\mathrm{Pd}=1$). The goal is then to minimize the number of false alarms. To implement $\mathrm{Pd}=1$, we have to choose a threshold that is lower than the smallest response in the target class. Therefore, if the discriminator yields a weak response to one of the targets, many false alarms in the clutter will result. Observing the plots for the NL-QGD we can see that the system response to targets has values as low as 0.2, so many false alarms (any output within the clutter chips that is above this threshold) will occur. In fact many more than for the linear counterpart, since the QGD never made as "gross" mistakes in the target class as the NL-QGD. This explains the reason why the QGD

performed better than the NL-QGD for discrimination (see the results for a complete comparison).

Ultimately, the performance of any adaptive system is controlled by the number, position and shape of the discriminant functions. The shape and number of the discriminant functions is dictated by the network topology, while their position in the input space depends upon the final weight values, which are a function of the performance criterion (also called the cost) which here was the L_2 norm.

There are two main problems in the use of the L_2 norm for discrimination. Notice that the L_2 norm (any norm for that matter) is blind to the error type, i.e., it treats in the same way the false alarms and the missed detections. The Neyman–Pearson criterion (1) and the construction of the ROC curve clearly demonstrate that these two types of errors affect differently the detection performance, so their contribution to the positioning of the discriminant functions should be controlled independently. This lead us to propose a mixed norm as an implementation of the Neyman–Pearson criterion. Moreover, the specific norm for the false alarms and missed detections should also be carefully selected, i.e., it is not clear that the L2 norm is the natural choice. Hampshire [11] also studied the norm selection and concluded that the L_2 norm is not appropriate for detection. We treat these two aspects below.

A. Mixed Norm Approximation to the Neyman–Pearson Criterion

Since the core idea of the Neyman–Pearson criterion is to treat differently the two types of errors in detection, this can only be achieved with a mixed norm formulation by defining the error as

$$E(n) = \left[\frac{1}{N_c}\sum_{x(n)\in\omega_1}|d - y(x(n),w)|^{p_c}\right] + \left[\frac{1}{N_t}\sum_{x(n)\omega_2}|d - y(x(n),w)|^{p_t}\right] \tag{22}$$

where N_c is the number of clutter samples, N_t is the number of target samples, p_c is the norm for the clutter class, p_t is the norm for the target class, and $\omega_{1,2}$ are the clutter and target classes, respectively.

In order to mimic the Neyman–Pearson criterion with norms, one wishes to minimize the largest deviation to the desired response in the target class such that the threshold can be set as high as possible; while in clutter one wishes to keep most of the errors small (deemphasize large errors) such that as few as possible of clutter responses cross the target threshold. This can be achieved by using the $L_{\inf}$ norm for target exemplars and the L_1 norm (or even fractional norms) for clutter exemplars. Due to the difficulty of training with these two extreme norms, we created a mixed norm $L_{1.1}$ for clutter and L_8 for targets. Next, we explain how the mixed norm criterion can be included in the BP algorithm.

B. L_p Norms and the BP Algorithm

The L_p norm is defined as

$$E = \sum_{n=1}^{N}\sum_{k=1}^{C}|d_k(n) - y_k(x(n);w)|^p \tag{23}$$

where N is the number of training samples, C the number of PE's in the output layer, and p the order of the norm $(p>0)$. When $p=1$ the contribution of large errors is equal to that of the small errors. When p increases, the contribution of large errors to E increases. Since the weights are changed with the error information and they define the location and shape of the decision boundary, different norms will move the decision boundaries in the input space and affect the performance of the classifier.

The training equations for the MLP can be easily modified to work with L_p norms [4]. In fact, the norm of the error only enters in the *computation of the injected error* used as input to the dual system. The BP algorithm can be utilized without modification. In equations we have

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_l}\frac{\partial y_l}{\partial w_{lj}} \tag{24}$$

where y_l is one of the output PE's and w_{lj} is an internal weight. By changing the norm we only affect the first term in the equation, which computes the injected error in the BP procedure. The local gradients in the output layer are computed as

$$\begin{aligned}\partial_l(n) &= \frac{\partial}{\partial y_l(n)}E(n)\\ &= \varphi'(y_l(n))\sum_{l=1}^{L}\operatorname{sign}(d_l(n)-y_l(n))|d_l(n)-y_l(n)|^{p-1}\end{aligned} \tag{25}$$

where l and L are a PE index and the number of nodes in the output layer, respectively. $E(n)$ is the instantaneous error defined as

$$E(n) = \frac{1}{p}\sum_{l=1}^{L}|d_l(n) - y_l(n)|^p. \tag{26}$$

Larger norms slow down the training. This effect can be seen by rewriting (25) as follows:

$$\begin{aligned}\delta_l(n) &= |e(n)|^{p-2}\varphi'(y_l(n))\\ &\quad\cdot\sum_{l=1}^{L}\operatorname{sign}(d_l(n)-y_l(n))|e(n)|.\end{aligned} \tag{27}$$

Comparing this equation with the L_2 norm equivalent, we conclude that when $|e(n)<1$, δ_l becomes very small for large norms $(p>2)$ so that the convergence slows down. In the L_2 norm, all the errors contribute equally to the computation of the gradient. For $p<2$, the gradient is more sensitive to smaller errors, while for $p>2$, larger errors affect more the weight updates.

Returning to the mixed norm criterion (22), in order to prevent the clutter error power from dominating the target error power when errors are less than one (the ones of interest due

to the nonlinearity) the terms are normalized by the number of class samples and the power of the inverses of p_c and p_t, respectively. The weight adaptation is batch training. The local gradient for the mixed norm at the output can be written as

$$\begin{aligned}\delta_l = &\left[\frac{1}{N}\sum_{x\in\omega_1}|d-y(x,w)|^{p_c}\right]^{(1/p_c)-1}\\ &\cdot\frac{1}{N}\sum_{x\in\omega_1}|d-y(x,w)|^{p_c-1}\varphi'(\nu_l)\\ &+\left[\frac{1}{N_t}\sum_{x\in\omega_2}|d-y(x,w)|^{p_t}\right]^{(1/p_t)-1}\\ &\cdot\frac{1}{N_t}\sum_{x\in\omega_2}|d-y(x,w)|^{p_t-1}\varphi'(\nu_l). \qquad (28)\end{aligned}$$

For convenience of notation, sample index n was dropped for δ_l, d, and x. For comparison purposes with the proposed mixed norm, the L8 norm was also implemented, as well as the cross-entropy cost (KL) function [1]. The cross entropy has been suggested as an appropriate criterion in cases where the distribution of the error is binomial as it may happen when one trains with two target values of 1/0.

VII. Experiments

A. Description of the Data

The data used to test the two-parameter CFAR and our proposed discriminators is high-resolution SAR imagery (mission 90 pass 5) collected at 23° depression angle near Stockbridge, NY, with the Advanced Detection Technology Sensor (ADTS). The ADTS is a fully polarimetric, coherent (HH, HV, VH, and VV polarizations), 33-GHz synthetic aperture radar with 1 ft $\times$ 1 ft resolution. The SAR image data was further processed by the polarimetric whitening filter (PWF) for speckle reduction [20]. The mission 90 pass 5 SAR data contains only natural and man-made clutter, lacking tactical targets. On the other hand, the TABILS 24 ISAR (inverse SAR) data set supplies tactical targets at various depression angles and many aspect angles.

Since at the time of this work there was no public source for target data in clutter, we had to embed targets to proceed with the testing of our algorithm. Target embedding is a compromise and should be avoided for classification. However, to test a focus of attention based on intensity features, a careful embedding method will provide realistic figures of merit. Moreover, here the emphasis is in relative performance of several different norms, not in absolute performance.

We chose ISAR target data that was taken at approximately the same depression angle as the mission 90 pass 5 SAR imagery (within 3° difference). Both data sets were calibrated. The target embedding was done coherently in the fully polarimetric clutter data by substituting clutter by target pixels. Target shadows were not included. Targets were only embedded in open field locations. A full description of the procedure is omitted here due to length limitations, but can be found in [15]. We can add that this embedding procedure was indirectly validated for the QGD when our measure of performance improvements with respect to the CFAR was later corroborated by Novak using targets in clutter [25].

B. Two-Parameter CFAR Processing

The two-parameter CFAR detector was run over the 127 frames (about 7 km^2) of the mission 90 pass 5 data. Three hundred forty-five targets from the TABILS 24 ISAR data base were embedded for testing purposes. The size of the CFAR stencil was 85 $\times$ 85 pixels as suggested by Novak *et al*. [21]. The local mean and standard deviation are computed in the outmost four-pixel wide boundary of the stencil. The intensity of the test pixel is computed by averaging 3 $\times$ 3 pixels in the center of the stencil. We could have used our γCFAR detector as described in [14], but we preferred to use an established detector to better compare the improvements of the discriminator.

After the CFAR processing, multiple detection points occur in targets and other regions because targets and man-made clutter (and tree tops) normally consist of many high reflectivity pixels that trigger the prescreener repeatedly (raw detections). Clustering of the multiple detections is performed for a more representative count of detections and false alarms. The size of the clustering region was determined from the size of targets (in this case, clustering radius is 22 pixels long, which encompasses the largest size of embedded targets) [14]. The QGD or NL-QGD only operate on the clustered locations, i.e., regions that triggered the CFAR detector. The number of false alarms is counted from the clustered detections.

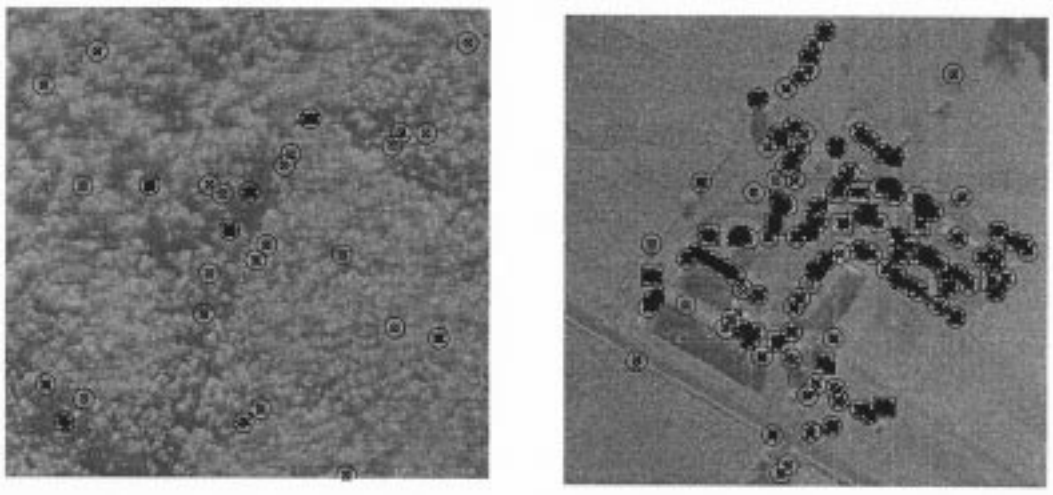
After clustering, the two-parameter CFAR detector yielded 4455 false alarms over the 127 frames of the mission 90 pass 5 SAR data set when the detection threshold was set at 100% target detection (all 345 targets). Fig. 6 shows an example of detection performance of the two-parameter CFAR detector in natural and cultural clutter areas. The objects in the square boxes in Fig. 6(b) indicate embedded targets.

C. Training the QGD

Training of the QGD finds the set of optimal scale parameters (μ_1, μ_{15}) and the weight vector W in (10). From the false alarms created by the two-parameter CFAR detector, 550 were randomly selected as clutter chips, and integrated with the set of 275 embedded target chips to form the training set. For the results presented here, the (μ_1, μ_{15}) space was exhaustively searched for the best possible stencil size in terms of minimum number of false positives for 100% detection $(\mathrm{Pd} = 1)$. Since there are only two parameters, this is realistic. Alternatively, the optimal scale could be determined by using a gradient descent procedure to minimize the output MSE as explained in [25]. Least squares was utilized to determined the optimal weight vector (14).

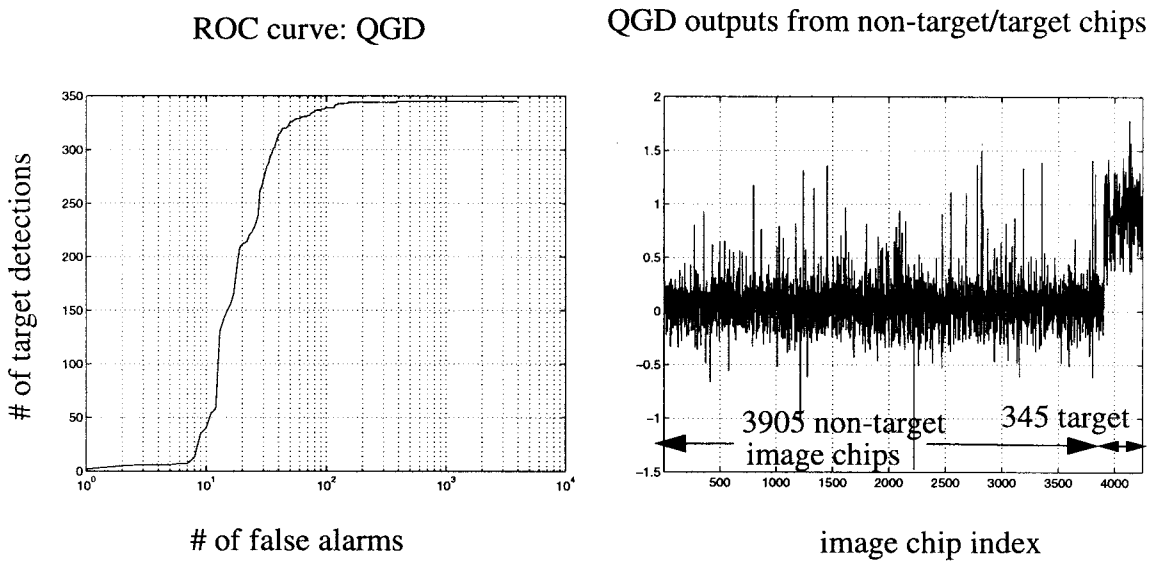
D. Testing the QGD

With the optimal set of parameters and optimal weights, the QGD was tested on different sets of clutter and target image chip embeddings (See Fig. 7). Table I shows the performance of the QGD in the testing phase. The table shows



(a) (b)

Fig. 6. Detection and clustering: The detections enclosed by the rectangular boxes indicate the location of targets embedded. 54 locations were false-detected by the two-parameter CFAR detector. (a) Natural clutter area. (b) Cultural clutter area.



(a) (b)

Fig. 7. Discriminating performance of the QGD in the test set. (a) ROC curves. (b) QGD outputs for the inputs of clutter and target image chips after training.

TABLE I
QGD DETECTION IN TRAINING AND TESTING

Target detection rates	Pd=1	Pd=.99	Pd=.98	Pd=.95	Pd=.92	MSE
Training	14	14	8	4	3	0.0273
Testing (# false alarms)	385	118	97	53	42	0.0202

the MSE and false alarms for selected detection rates close to 100% detection, the desirable operating point for the focus of attention.

The QGD reduces the false alarms from 3905 to 385 for $\mathrm{Pd} = 1$. An increase of discrimination power of about $1:10$ (385/3905) was obtained with the QGD over the two parameter CFAR detector at $\mathrm{Pd} = 1$. The number of false alarms decreases to 118, 97, 53, and 42 for $\mathrm{Pd} = 0.99, 0.98, 0.95,$ and 0.92, respectively. A small decrease in detection accuracy reduces greatly the number of false alarms. The testing of the NL-QGD is performed with 3905 clutter image chips and 345 target image chips.

E. Training and Testing NL-QGDs

In order to compare the performance of the QGD and NL-QGD based on the same feature values, the same optimal

TABLE II
PERFORMANCE OF THE NL-QGD WITH L_2

	Network Topologies	False detections					MSE
		Pd=1	Pd=.99	Pd=.98	Pd=.95	Pd=.92	
training	NL-QGD731	80	6	3	3	2	0.005039
	NL-QGD751	80	6	3	3	2	0.005038
	NL-QGD771	80	6	3	3	2	0.005048
testing	NL-QGD731	1456	96	67	35	30	0.006263
	NL-QGD751	1453	96	67	35	30	0.006261
	NL-QGD771	1445	97	67	35	30	0.006262

TABLE III
PERFORMANCE OF THE NL-QGD WITH MIXED NORM

	Network Topologies	False detections					MSE
		Pd=1	Pd=.99	Pd=.98	Pd=.95	Pd=.92	
training	NL-QGD731	12	12	10	5	4	0.00682
	NL-QGD751	12	12	10	5	4	0.00662
	NL-QGD771	12	12	10	5	4	0.00664
testing	NL-QGD731	232	97	76	40	36	0.00830
	NL-QGD751	252	96	76	48	36	0.00833
	NL-QGD771	251	96	76	48	36	0.00833

parameters for the QGD, μ_1 and μ_{15}, were used for the NL-QGD. To train the NL-QGD an iterative training algorithm such BP must be utilized. Since the image features extracted with the gamma stencil are highly correlated, it is desirable to whiten the training data set for faster convergence [15]. A cross validation set of 550 clutter image chips and another set of 275 embedded target image chips was created to stop the training at the point of maximum generalization.

The size of the NL-QGD hidden layer was changed from 3, 5, and 7 to quantify the effect of topology size on performance. In the following tables, NL-QGD731 indicates that the network has seven input nodes, three nodes in the hidden layer, and one output node. Note that our goal is to have small topologies to guarantee good generalization. Backpropagation with a momentum term was used to train the NL-QGD weights with the whitened training data set. For NL-QGD training, the learning rate (h) and momentum (a) were 0.1 and 0.8, respectively, for all the networks. The cross validation set was also whitened by using the estimate of the covariance matrix of the training data set. Training was stopped when the error in the cross-validation set started to increase. This happened at different iteration counts depending on the size of the networks and the norm utilized.

The performance of the NL-QGD's trained with L_2 norm is measured in terms of MSE and the number of false alarms for different network sizes. The results are tabulated in Table II for training and testing. The NL-QGD is able to provide a smaller final MSE than the QGD (an improvement of an order of magnitude). However, the performance of a discriminator is not measured in terms of MSE but number of false alarms for a given detection accuracy. Hence, the ROC curve of the two systems must be compared [Fig. 8(a)]. The NL-QGD trained with the L_2 norm did not perform as well as the QGD at $\mathrm{Pd} = 1$ (1880 false alarms versus only 422 for the QGD). However, at $\mathrm{Pd} = 0.99$, 0.98, and 0.95, the NL-QGD outperformed the QGD. Note that changing the number of hidden nodes from three to seven did not affect the performance of the NL-QGD much.

The performance with mixed norm is shown in Table III. A small norm $(p = 1.1)$ was imposed on the nontarget class and a large norm $(p = 8)$ on the target class. The BP algorithm was modified according to (28). The most obvious result of Table III versus Table II is a drastic reduction of the false alarms at $\mathrm{Pd} = 1$, and an increase in the MSE.

In fact, this mixed norm was the best performer for $\mathrm{Pd} = 1$ among all the experiments that we conducted involving the L_2, L_8 norm, the cross-entropy criterion and the QGD [15]. The ROC curves and the detections per class of all the norm experiments are shown in Fig. 8.

It is interesting to analyze this figure to gain an intuition for how different norms treat the errors. Notice that the L_2 norm produces large errors in the target class that are very costly for

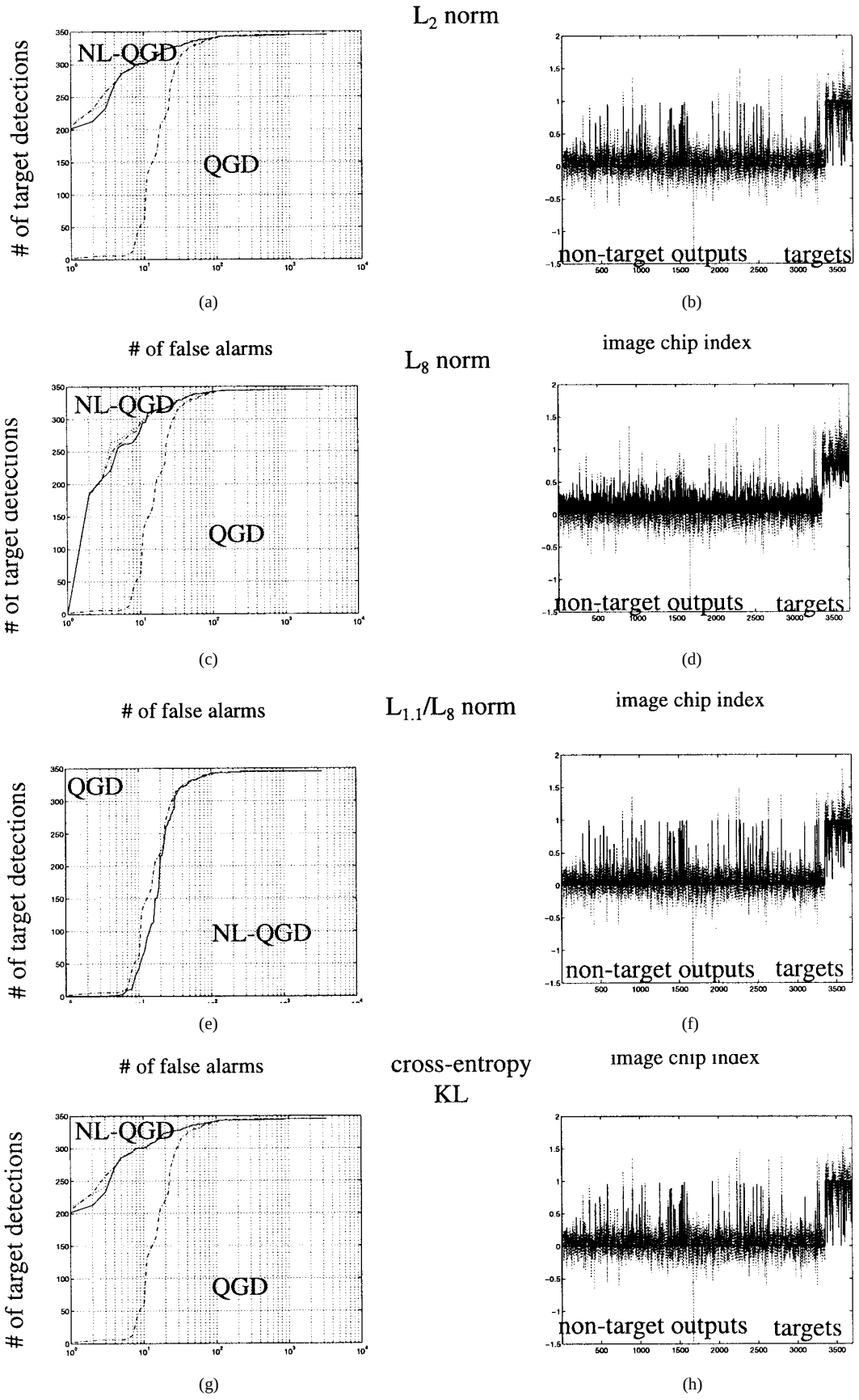


Fig. 8. ROC curves and detections for the NL-QGD trained with four different criteria.

TABLE IV
FINAL RANKING OF ALL THE DISCRIMINATORS

Rank	Detection rates				
	100%	99%	98%	95%	92%
1	NL-QGD731$_{L1.1/L8}$ (292)	NL-QGD771$_{L8}$ (114)	NL-QGD751$_{L8}$ (70)	NL-QGD771$_{KL}$ (36)	NL-QGD771$_{KL}$ (20)
2	NL-QGD771$_{L1.1/L8}$ (315)	NL-QGD751$_{L8}$ (115)	NL-QGD771$_{L8}$ (73)	NL-QGD751$_{KL}$ (36)	NL-QGD751$_{KL}$ (20)
3	NL-QGD751$_{L1.1/L8}$ (316)	NL-QGD731$_{L8}$ (119)	NL-QGD731$_{L8}$ (73)	NL-QGD731$_{KL}$ (39)	NL-QGD731$_{KL}$ (20)
4	QGD$_{LS}$ (422)	NL-QGD771$_{L2}$ (120)	NL-QGD771$_{L2}$ (83)	NL-QGD771$_{L8}$ (42)	NL-QGD751$_{L8}$ (29)
5	NL-QGD751$_{L8}$ (632)	NL-QGD751$_{L2}$ (120)	NL-QGD751$_{L2}$ (83)	NL-QGD731$_{L8}$ (42)	NL-QGD771$_{L8}$ (30)
6	NL-QGD771$_{L8}$ (640)	NL-QGD731$_{L2}$ (120)	NL-QGD731$_{L2}$ (83)	NL-QGD751$_{L8}$ (43)	NL-QGD731$_{L8}$ (31)
7	NL-QGD731$_{L8}$ (652)	NL-QGD771$_{L1.1/L8}$ (123)	NL-QGD751$_{KL}$ (90)	NL-QGD771$_{L2}$ (46)	NL-QGD771$_{L2}$ (37)
8	NL-QGD731$_{KL}$ (910)	NL-QGD751$_{L1.1/L8}$ (123)	NL-QGD771$_{KL}$ (91)	NL-QGD751$_{L2}$ (46)	NL-QGD751$_{L2}$ (37)
9	NL-QGD751$_{KL}$ (1064)	NL-QGD731$_{L1.1/L8}$ (123)	NL-QGD731$_{KL}$ (97)	NL-QGD731$_{L2}$ (46)	NL-QGD731$_{L2}$ (37)
10	NL-QGD771$_{KL}$ (1152)	NL-QGD771$_{KL}$ (125)	NL-QGD771$_{L1.1/L8}$ (101)	QGD$_{LS}$ (57)	QGD$_{LS}$ (44)
11	NL-QGD771$_{L2}$ (1880)	NL-QGD751$_{KL}$ (129)	NL-QGD751$_{L1.1/L8}$ (101)	NL-QGD771$_{L1.1/L8}$ (63)	NL-QGD771$_{L1.1/L8}$ (45
12	NL-QGD751$_{L2}$ (1889)	NL-QGD731$_{KL}$ (130)	NL-QGD731$_{L1.1/L8}$ (101)	NL-QGD751$_{L1.1/L8}$ (63)	NL-QGD751$_{L1.1/L8}$ (45)
13	NL-QGD731$_{L2}$ (1892)	QGD (162)	QGD$_{LS}$ (109)	NL-QGD731$_{L1.1/L8}$ (63)	NL-QGD731$_{L1.1/L8}$ (45)

a discriminator. However, the minimization produces a very good match "in the mean" for the targets values of zero and one as seen by the MSE.

The L_8 norm produces a much more irregular response throughout the classes, and reduces the distance between the mean value of the target and clutter responses. However, notice that in fact there are fewer large errors in both classes. The $L_{1.1/8}$norm reduces the size of the errors for targets and keeps most of the clutter detects near zero, as the Neyman–Pearson criterion dictates. So we think that the mixed norm is a good approximation of the Neyman–Pearson criterion for $\mathrm{Pd} = 1$. The cross entropy criterion resembles the performance of the L_2 norm but produces fewer large errors.

Table IV shows the final ranking of all discriminators for the probability of detection of $\mathrm{Pd} = 1, 0.99, 0.98, 0.95, 0.92$. It is also very instructive to analyze this table. The first obvious result is that the discriminators trained with the same criterion are normally grouped together. This means that the biggest factor affecting performance is the norm chosen for the criterion, while the topology is a second-order effect. So for our problem, the time spent in the conventional struggle to find the best topology can be best utilized in finding the best norm to represent the data.

A large discrepancy in the discriminators performance for $\mathrm{Pd} = 1$ detection is also observed. The best performer created only 292 false alarms, and the worst performer detected 1892 clutter chips, so the definition of the best norm is critical for $\mathrm{Pd} = 1$. It is remarkable that for $\mathrm{Pd} = 1$ detection, the QGD is the second best discriminator.

The best norm for training depends upon the operating point chosen in the ROC curve (the probability of detection). This can be seen in the alternation of the discriminator rankings in Table IV. The mixed norm jumps from first rank for $\mathrm{Pd} = 1$ to last for the operating points $\mathrm{Pd} = 0.95$ and $\mathrm{Pd} = 0.92$, which means that the decision boundary is being placed with an enormous emphasis to minimize the largest errors on targets. This is not a good strategy for the other operating points where the discriminator is allowed to trade precision for sensitivity. For $\mathrm{Pd} = 0.99$ and 0.98, the L_8 norm is the best performer, but for the other two operating points ($\mathrm{Pd} = 0.95$ and 0.92), the best performer is the cross-entropy measure. Notice however, that the number of false alarms for all these detection rates are rather similar, which means that the choice of the norm is not as critical as for $\mathrm{Pd} = 1$. Fig. 9 illustrates this fact by displaying the number of false alarms as a function of the probability of detection for the best performer in each of the norms.

It is interesting to note that the L_2 norm never produced the best results which indicates that for detection there are better alternatives than the MSE criterion. Finally, the QGD

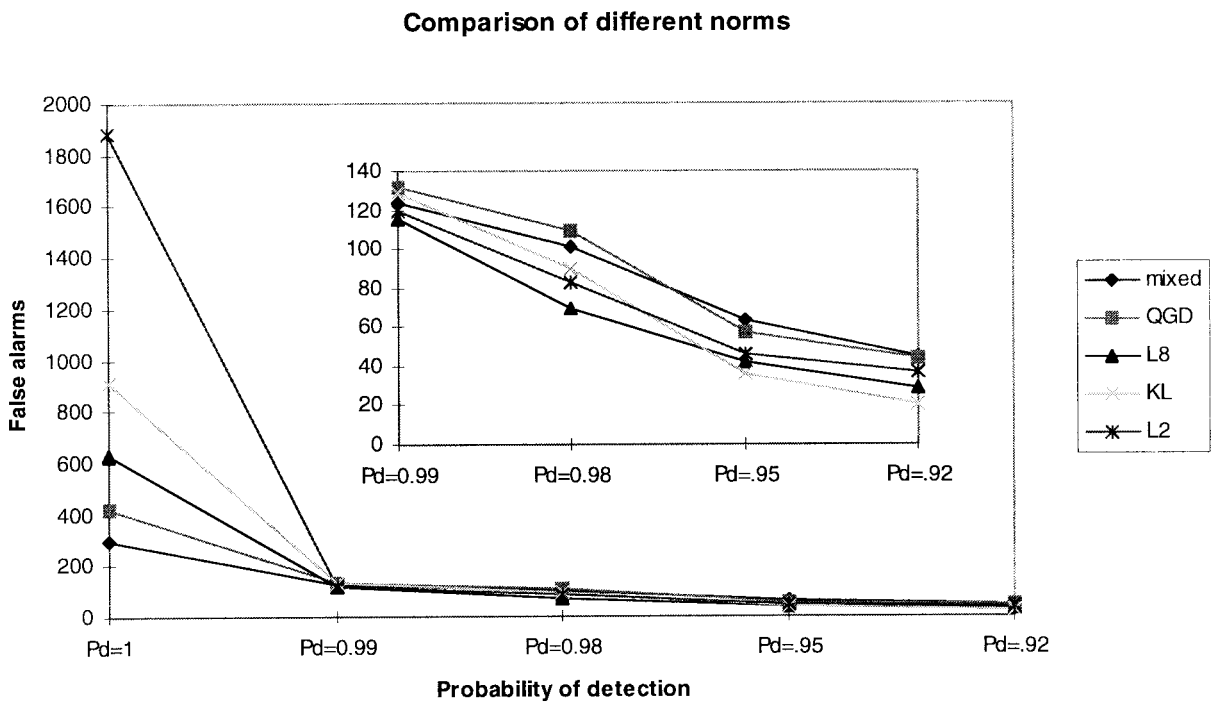


Fig. 9. Comparison of the ROC's for high probability of detection.

goes from fourth in the ranking for $\mathrm{Pd} = 1$ to basically last for all the other detection rates. So, we conclude that there is an advantage of using nonlinear systems to design focus of attention blocks for SAR.

VIII. Conclusion

There are two main set of conclusions resulting from this work. We first address the area of ATR for SAR. The combination of the two parameter CFAR with the QGD was shown to be a superior implementation of the focus of attention, producing approximately ten false alarms per square km of imagery [15]. The QGD can be used in conjunction with other discriminators (such as fractal dimension, weighted rank fill ratio, size, polarimetric purity, etc.). In fact, its response tends to be uncorrelated with the response of the other discriminators [Novak, personal communications] helping reduce the number of false alarms in conventional focus of attentions for SAR.

In this paper, we extend further the system performance by substituting the QGD with a MLP, which we called the NL-QGD. The NL-QGD outperforms the QGD at all the detection rates tested ($\mathrm{Pd} = 1$, 0.99, 0.98, 0.95, 0.92), reducing the number of false alarms by and average of 40%. The NL-QGD is still a system with small number of parameters (24 weights), so it is fast to adapt and should generalize well. This work shows a practical advantage of the MLP in a very difficult and relevant problem.

The other set of conclusions addresses neural network issues. We found that training the MLP for discrimination with the commonly used MSE criterion produced disappointing performance. In fact, it produces a discriminator that was worse than its linear counterpart, the QGD, for $\mathrm{Pd} = 1$. The neural network trained with the L_2 norm never was the best performer for any of the operating points tested. So we conclude that while the MSE is extensively used in classification, it should not be used to train MLP's for discrimination of targets in SAR.

We investigated the reasons, and concluded that the problem is related to the different goals of detection and classification. Classification is still a special case of function approximation with indicator functions. In this class of problems MSE has been shown optimal for Gaussian distributed errors [1]. However, detection is a very different problem. For detection, the missed detections and the false alarms should be treated very differently as suggested by the Neyman–Pearson criterion. The criterion suggests that the target norm should be $L_{\inf}$, and the clutter norm should be L_1 (or even fractional norms). We found a way to construct a mixed norm cost function and successfully applied it to train the NL-QGD using the backpropagation algorithm. The NL-QGD trained with the mixed norm was the best performer for $\mathrm{Pd} = 1$, the suggested operating point for the focus of attention.

The experimental results also show that the best error norm depends upon the set point chosen for the probability of detection. This can be understood in the light of the Neyman–Pearson approach, but a systematic way to translate the choice of the operating point to the error criterion is lacking at this time. The other error norms tested such as the L_8 or the cross-entropy (KL) outperformed the L_2 norm. However, we also found that the operating point where performance is more dependent upon the norm is the region around $\mathrm{Pd} = 1$. More work should be conducted with larger number of targets and clutter chips to quantify better the performance near $\mathrm{Pd} = 1$ (which is the point with the largest estimator variance).

Finally, the performance was much more dependent upon the norm than the size of the MLP topology. This suggests

that modeling the pdf of the error to construct the best norm should be a design goal. Unfortunately, this is not common practice in neural network applications.

REFERENCES

[1] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford Univ. Press, 1995.

[2] A. Bernardon and J. Carrick, "A neural system for automatic target learning and recognition applied to bare and camouflaged SAR targets," *Neural Networks, Special Issue Target Recognit.*, vol. 8, pp. 1103–1108, 1995.

[3] M. C. Burl, G. J. Owirka, and L. M. Novak, "Texture discrimination in synthetic aperture radar imagery," in *Proc. 23rd Asilomar Conf. Signals, Systems, Computers*, Pacific Grove, CA, Oct. 30–Nov. 1, 1989, pp. 399–404.

[4] P. Burrascano, "A norm selection criterion for the generalized delta rule," *IEEE Trans. Neural Networks*, vol. 2, 1991.

[5] D. Casasent and L. Neiberg, "Classifier and shift invariant automatic target recognition neural network," *Neural Networks, Special Issue Target Recognit.*, vol. 8, pp. 1053–1080, 1995.

[6] B. deVries and J. Principe, "The gamma neural model: A new model for temporal processing," *Neural Networks*, vol. 5, pp. 565–576, 1992.

[7] H. Finn, "Adaptive detection in clutter," in *Proc. Nat. Electronics Conf.*, vol. 22, p. 562, 1966.

[8] S. Grossberg, E. Mingolla E., and Williamson, "Synthetic aperture radar processing by a multiple scale neural system for boundary and surface representation," *Neural Networks, Special Issue Target Recognit.*, vol. 8, pp. 1005–1028, 1995.

[9] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York: Academic, 1990.

[10] G. B. Goldstein, "False-alarm regulation in log-normal and Weibull clutter," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-9, pp. 84–92, Jan. 1973.

[11] J. Hampshire and A. Waibel, "A novel objective function for improved phoneme recognition using time delay neural networks," *IEEE Trans. Neural Networks*, vol. 1, pp. 216–228, 1990.

[12] S. Haykin, *Neural Networks: A Comprehensive Foundation*. New York: Macmillan, 1994.

[13] C. Helstrom, *Statistical Theory of Signal Detection*. New York: Pergamon, 1968.

[14] M. Kim, J. Fisher, and J. Principe, "A new CFAR stencil for target detections in synthetic aperture radar (SAR) imagery," *Proc. SPIE*, vol. 2757, pp. 432–442, 1996.

[15] M. Kim, "Focus of attention based on gamma kernels for automatic target recognition," Ph.D. dissertation, Univ. Florida, Gainesville, FL, 1996.

[16] M. Koch, M. Moya, L. Hostetler, and R. Fogler, "Cueing, feature discovery, and one class learning for synthetic aperture radar automatic target recognition," *Neural Networks, Special Issue Target Recognit.*, vol. 8, pp. 1081–1102, 1995.

[17] D. E. Kreithen, S. D. Halversen, and G. J. Owirka, "Discriminating targets from clutter," *Lincoln Lab. J.*, vol. 6, pp. 25–52, 1993.

[18] J. Neyman and E. Pearson, "On the problem of the most efficient tests of statistical hypothesis," *Philos. Trans. R. Soc. A*, vol. 231, pp. 289–337, 1933.

[19] L. Novak, S. Halversen, G. Owirka, M. Hiett, "Effects of polarization and resolution on the performance of a SAR automatic target recognition system," *Lincoln Lab. J.*, vol. 8, pp. 49–68, 1995.

[20] L. M. Novak and M. C. Burl, "Optimal speckle reduction in polarimetric SAR imagery," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 26, pp. 293–305, Mar. 1990.

[21] L. M. Novak, G. J. Owirka, and C. M. Netishen, "Performance of a high-resolution polarimetric SAR automatic target recognition system," *Lincoln Lab. J.*, vol. 6, no. 1, pp. 11–24, 1993.

[22] L. Perlovski, J. Chernick, and W. Schoendorf, "Multi-sensor ATR and identification of friend or foe using MLANS," *Neural Networks, Special Issue Target Recognit.*, vol. 8, no. 7/8, pp. 1185–1120, 1995.

[23] J. C. Principe, B. deVries, and P. G. de Oliveira, "The gamma filter—A new class of adaptive IIR filters with restrictive feedback," *IEEE Trans. Signal Processing*, vol. 41, pp. 649–656, 1993.

[24] J. C. Principe *et al.*, "Target prescreening based on 2D gamma kernels," *Proc. SPIE—Int. Soc. Opt. Eng., Algorithms Synthet. Apert. Radar Imag. II*, vol. 2487, pp. 251–258, Apr. 1995.

[25] J. Principe *et al.*, "Target prescreening based on a quadratic gamma discriminator," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 34, 1998.

[26] S. Rogers *et al.*, "Neural networks for automatic target recognition," *Neural Networks, Special Issue Target Recognit.*, vol. 8, pp. 1153–1184, 1995.

[27] A. Waxman, M. Seibert, A. Bernardon, and D. Fay, "Neural systems for automatic target learning and recognition," *Lincoln Lab. J.*, vol. 6, pp. 77–116, 1993.

José C. Principe (M'83–SM'90) is a Professor of Electrical and Computer Engineering at the University of Florida where he teaches signal processing and artificial neural networks (ANN's). He is the Founder and Director of the University of Florida Computational NeuroEngineering Laboratory (CNEL). His primary area of interest is processing of nonstationary signals with adaptive neural models. He proposed a new, biological plausible, model for the classification of time varying patterns (the gamma model). The CNEL Lab has been studying the use of ANN's for dynamical modeling and target detection and recognition.

Dr. Principe is a member of the advisory board of the University of Florida Brain Institute.

Munchurl Kim was born in Korea in 1966. He received the Bachelor of Engineering degree in electronics from KyungPook National University, Korea, in 1989, and the Master's and Ph.D. degrees in electrical engineering from the University of Florida, Gainesville, in 1992 and 1996, respectively.

He is currently with the Electronics and Telecommunication Research Institute, Taejon, Korea. His interests are in automatic target recognition, image processing, and artificial neural networks.

John W. Fisher, III (M'91) received the Ph.D. degree in electrical and computer engineering from the University of Florida, Gainesville, in 1997.

He is currently a Research Scientist in the Laboratory for Information and Decision Systems and a Post-doctoral Associate in the Artificial Intelligence Laboratory, Massachusetts Institute of Technology (MIT), Cambridge. Prior to joining MIT, he was affiliated with the University of Florida, as both a faculty member and graduate student since 1987, during which time he conducted research in the areas of ultrawideband radar for ground penetration and foliage penetration applications, radar signal processing, and automatic target recognition algorithms. His current area of research, begun as a Ph.D. candidate in the Computational NeuroEngineering Laboratory at the University of Florida, is in information theoretic approaches to signal processing, machine learning, and computer vision.

Dr. Fisher is a member of the SPIE.