Recent Advances to Nonlinear MACE Filters

John W. Fisher III and Jose C. Principe
Computational NeuroEngineering Laboratory
University of Florida
Gainesville, FL 32611
fisher@cnel.ufl.edu, principe@cnel.ufl.edu

ABSTRACT

We present recent advances in the development of nonlinear extensions to the minimum average correlation energy (MACE) filter. The MACE filter and its variations have been applied to the area of automatic target detection and recognition (ATD/R). Nonlinear extensions (Fisher and Principe, 1994) have been presented based on a statistical formulation of the optimization criterion, of which the linear MACE filter is a special case. The method by which nonlinear topologies can be incorporated into the filter design is reviewed. We present recent advances to this nonlinear method as well as new experimental results applying the technique to inverse synthetic aperture radar (ISAR) data. The methods described result in faster convergence times and significantly better classification performance.

Keywords: correlation filters, neural networks, pattern recognition, ISAR

1. INTRODUCTION

We have presented a methodology by which minimum average correlation energy (MACE) filtering techniques can be extended to nonlinear signal processing.[1] Extension of linear methods to nonlinear processing comes with added complexity. As a practical matter, closed form solutions are difficult and such filters must be computed iteratively via an adaptive method such as gradient search. Consequently, the nonlinear approach was based upon an adaptive framework. Emphasis was placed on a statistical perspective of the family of optimized correlators which includes the MACE filter. From this perspective the optimization criteria of the MACE filter can be interpreted as a description of an implicit rejection class by its second order statistics while the recognition class is characterized by exemplars. In contrast to the linear MACE filter, however, the output plane variance for a nonlinear mapping cannot, in general, be determined with anything akin to Parseval's theorem. Therefore, the challenge for nonlinear extensions is to transfer the characterization of the recognition and rejection classes to an adaptive framework.

We discuss several issues in this paper with regards to nonlinear extensions. We begin by examining common measures of generalization for distortion invariant filtering. Our analysis shows that with regards to classification these measures are not particularly good indicators of performance for the inverse synthetic aperture (ISAR) data that we are working with. The measures discussed relate to how well a filter generalizes to between aspect exemplars of the training images which have not been used in the filter computation. The challenge for a classifier, however, is to recognize images from the same class of vehicle. The classification is measured, therefore, with regards to recognition of different vehicles in the same class. This analysis is useful for establishing a proper basis upon which to compare our nonlinear extensions to their linear counterparts.

It is also shown in this analysis that emphasis on the MACE filter criterion leads to superior classification performance and so motivates our choice of pre-processor within a nonlinear architecture.

In subsequent sections we review the method by which we extend the MACE filter to nonlinear adaptive systems. This is followed by experimental results which illustrate potential pitfalls mentioned in our previous work as well as present recent advances to this technique. It is worth stating that brute force to train the nonlinear system as a black box approach yield dissatisfying results. Hence, the advantage of the nonlinear system hinges on efficient methods for representing the rejection class during training. Two new training methods are also presented. The first results in faster training times while the second results in significantly better classification performance.

Throughout the experiments we will be using ISAR imagery of vehicles rotated through a range of aspect angle to test our nonlinear extensions. Independent vehicles, with slightly different configurations and radar depression angles than the training vehicle, will be used for testing. This, we feel, results in a more realistic scenario in which to compare nonlinear classification performance.

1.1 *Nonlinear Architecture*

It well known that the MACE filter can be decomposed as a pre-whitening filter followed by a synthetic discriminant function (SDF),[2,3] which can also be viewed as a special case of Kohonen's linear associative memory (LAM).[4,5] This decomposition is shown at the top of figure 1. The nonlinear filter architecture with which we are experimenting is shown in the middle of figure 1. In this architecture we replace the LAM with a nonlinear associative memory, specifically a feed-forward multi-layer perceptron (MLP), shown in more detail at the bottom of figure 1. We have shown among other desirable qualities that this modification maintains the shift invariance property of the MACE filter.[1]

Another reason for choosing the multi-layer perceptron (MLP) is that it is capable of achieving a much wider range of discriminant functions. It is well known that an MLP with a single hidden layer can approximate any smooth discriminant function.[6] One of the shortcomings of optimized correlators such as the MACE filter is that we are trying to fit a hyper-plane to our training exemplars as the discriminant function. Using an MLP in place of the LAM relaxes this constraint. In general MLPs do not allow for closed form solutions, therefore, we are forced to go to an iterative framework.

2. Classifier performance and measures of generalization

One of the issues for any adaptive method which relies on exemplars is the number of training exemplars to use in the computation of the discriminant function. In addition, for iterative methods, there is the issue of when to stop the adaptation process. In the case of distortion invariant filters, such as the MACE filter, some common heuristics are used to determine the number of training exemplars. Typically samples are drawn from the training set and used in the closed form analytic solution until the minimum peak response over the remaining samples exceeds some threshold.[7] A similar heuristic is to continue to draw samples from the training set until the mean square error of the peak response over the remaining samples drops below some preset threshold. These measures are then used as indicators of how well the filter generalizes to between aspect exemplars from the training set which have not been used for the computation of the filter coefficients.

The ultimate goal, however, is classification. Generalization in the context of classification must be related to the ability to classify a previously unseen input.[8] We show by example that the measures of generalization mentioned above may be misleading as predictors of classifier performance. In fact the result of the experiments will show that the way in which the data is pre-processed is more indicative of classifier performance than these other indirect measures.

We illustrate this point with an example using ISAR image data. Figure 2 shows ISAR images of size 64×64 taken from five different vehicles and two different vehicle types. The images are all taken with the same radar. Data taken from vehicles in the same class vary in the vehicle configuration and radar depression angle (15 or 20 degrees depression). Images have been formed from each vehicle at aspect variations of 0.125 degrees from 5 to 85 degrees aspect for a total of 641 images for each vehicle. Figure 2 shows each of the vehicles at 5, 45, and 85 degrees aspect.

We will use vehicle type 1 as the recognition class and vehicle type 2 as a confusion vehicle. Images of vehicle 1a will be used as the set from which to draw training exemplars. Classification performance will then be measured as the ability to recognize vehicles 1b and 1c while rejecting vehicles 2a and 2b. The filter we will use is a form of the optimal trade-off synthetic discriminant function[9] (OTSDF) which is computed in the spectral domain as

$$H = [\alpha P_x + (1-\alpha)\overline{P_x}]^{-1} X [X^{\dagger} [\alpha P_x + (1-\alpha)\overline{P_x}]^{-1} X] d\,, \tag{1}$$

where the columns of the data matrix $X \in C^{N_1 N_2 \times N_t}$ the Fourier coefficients of N_t exemplar images of dimension $N_1 \times N_2$ of vehicle 1a reordered into column vectors. The diagonal matrix $P_x \in \Re^{N_1 N_2 \times N_1 N_2}$ contains the coefficients of the average power spectrum measured over the N_t exemplars of vehicle 1a, while $\overline{P_x} \in \Re^{N_1 N_2 \times N_1 N_2}$ is the identity matrix scaled by the average of the diagonal terms of P_x. Finally, $d \in \Re^{N_t \times 1}$ is a column vector of desired outputs, one for each exemplar. The elements of d are typically set to unity. When α is set to unity equation (1) yields exactly the MACE filter, when it is set to zero the result is the SDF. The filter we are using is therefore trading off the MACE filter criterion with the SDF criterion. The SDF criterion can also be viewed as the MVSDF[10] criterion when the noise class is represented by a white noise random process. This filter can also be decomposed as in figure 1.

In these experiments we would like to examine the relationship between the two commonly used measures of generalization and two measures of classification performance such that conclusions can be drawn about the appropriateness in a classification framework. The generalization measures are the minimum peak response and the mean square error taken over the aspect range of the images of the training vehicle (excluding the aspects used for computing the filter). The classification measures are taken from the receiver operating characteristic (ROC) curve measuring the probability of detecting, P_d, a testing vehicle in the recognition class (vehicles 1b and 1c) versus the probability of false alarm, P_{fa}, on a testing vehicle in the confusion class (vehicles 2a and 2b) based on peak detection. The specific measures are the area under the ROC curve, a general measure of the test being used, while the second measure is the probability of false alarm when the probability of detection equals 80%, which measures a single point of interest on the ROC curve.

Two filters are used, one with $\alpha = 0.5$ and the other with $\alpha = 0.95$, or one in which both criterion are weighted equally and one which is close to the MACE filter criterion. The number of exemplars drawn from the training vehicle (1a) is varied from 21 to 81 sampled uniformly in aspect (1 to 4 degrees aspect separation between exemplars).

Examination of figures 3 and 4 show that for both cases (α equal to 0.5 and 0.95) no clear relationship emerges in which the generalization measures are indicators of good classification performance. Table I compares the classifier performance when the generalization measures as described above are used to choose the filter versus the best ROC performance achieved throughout the range of aspect separation. In one regard, the generalization measures were consistent in that the same aspect separation was predicted by both measures for both settings of α. In figure 5 we compare the ROC curves for four cases the filter chosen using the generalization measures and the best achieved ROC curve for both settings of α. We would expect that for each α the filter using the generalization measure would be near the best ROC performance. As can be seen in the figure this is not the case.

Table I. Classifier performance measures when filter is determined by either of the common measures of generalization as compared to best classifier performance for two values of α.

		Generalization Measure		
		y_{min}	y_{mse}	**Best**
$\alpha = 0.50$	**P_{fa}@P_d=0.8**	0.24	0.24	0.16
	ROC area	0.83	0.83	0.90
$\alpha = 0.95$	**P_{fa}@P_d=0.8**	0.16	0.16	0.07
	ROC area	0.94	0.94	0.95

It is obvious from figures 3 and 4 that the generalization measures are not significantly correlated with the ROC performance. In fact, as summarized in table II, the generalization measures are negatively, albeit weakly, correlated with ROC performance. One feature of figures 3 and 4 is that although ROC performance varies independent of minimum peak response or MSE, there does appear to be dependency on α. This leads to a second experiment.

Table II. Correlation of generalization measures to classifier performance. In both cases (α equal to 0.5 or 0.95) the classifier performance as measured by the area of the ROC curve or P_{fa} at P_d equal 0.8, has an opposite correlation as to what would be expected of a useful measure for predicting performance. Furthermore, as seen in figures 3 and 4 and by the computation below, these measures are only weakly correlated to the classifier's performance.

		Performance Measures			
		ROC area	**P_{fa}(@P_d=0.8)**	**ROC area**	**$P_{fa(}$@P_d=0.8)**
		$\alpha = 0.50$		$\alpha = 0.95$	
Generalization Measures	y_{min}	-0.39	0.21	-0.40	0.41
	y_{mse}	0.32	-0.11	0.31	-0.35

In the second experiment we examine the relationship between the parameter α and the ROC performance. The aspect separation between training exemplars is set to 2, 4, and 8 degrees. The value of α, the emphasis on the MACE criterion, is varied in the range zero to unity. Figure 6 shows the relationship between

ROC performance and the value of α. It is clear from the plots that there is a positive relationship between the emphasis on the MACE criteria and the ROC performance. However, the peak in ROC performance is not achieved at α equal to unity. In all three cases, the ROC performance peaks just prior to unity with the performance drop-off increasing with aspect separation at α equal to unity.

The difference between the SDF and MACE filter is the pre-processor. What is shown by this analysis is that, in general, the pre-processor from the MACE filter criterion leads to better classification, but too much emphasis on the MACE filter criterion, as measured by α equal to unity, leads to a filter which is too specific to the training samples. The problems described above are well known. Alterations to the MACE criterion have been the subject of many researchers.[7,11,12] There is still, as yet, no principled method found in the literature by which to set the parameter α.

There are two conclusions from this analysis that are pertinent to the nonlinear extension we are using. First the results show that pre-whitening over the recognition class leads to better classification performance. For this reason we choose to use the pre-processor of the MACE filter in our nonlinear filter architecture. The issue of extending the MACE filter to nonlinear systems can in this way be formulated as a search for a more robust nonlinear discriminant function in the pre-whitened image space.

The second conclusion is that comparisons of the nonlinear filter to its linear counterpart must be made in terms of classification performance only. There are simple nonlinear systems, such as a soft threshold at the output of a linear system for example, that will outperform the MACE filter or its variations in terms of maximizing the minimum peak response over the training vehicle or reducing the variance in the output image plane. These measures are not, however, sufficient to describe classification performance. We have also used these measures in the past but feel that they are not the most appropriate for classification.[1,13]

3. NONLINEAR ADAPTIVE METHODS

The MACE filter is the best linear system that minimizes the energy in the output correlation plane subject to a peak constraint at the origin. An advantage of linear systems is that we have the mathematical tools to use them in optimal operating conditions. Such optimality conditions, however, should not be confused with the best possible performance. In the case of ISAR data, for example, there is little reason to believe that a linear discriminant function will give the best possible discrimination performance.

Our goal is to extend the optimality condition of MACE filters to adaptive nonlinear systems. The optimality condition of the MACE filter considers the entire output plane, not just the response when the image is centered. With regards to the nonlinear filter architecture of figure 1, a brute force approach would be to present training images and all shifted versions to a neural network with a desired output of unity for the centered images and zero for the shifted versions. This would indeed emulate the optimality of the MACE filter, however, the result is a training algorithm of order $N_1 N_2 N_t$ for N_t training images of size $N_1 \times N_2$ pixels. This is clearly impractical.

Previously we have presented a statistical viewpoint of optimized correlators from which such nonlinear extensions fit naturally into an adaptive framework.[1] This treatment results in an efficient way to capture the optimality condition of the MACE filter using a training algorithm which is approximately of order N_t and which leads to better classification performance than the linear MACE. We review that development here.

3.1 *A statistical perspective*

Consider images of dimension $N_1 \times N_2$ re-ordered by column or row into vectors. Let the random vector, $X_1 \in \Re^{N_1 N_2 \times 1}$, represent a rejection class. We know the second-order statistics of this class as represented by the average power spectrum (or equivalently the autocorrelation function). The columns of a data matrix $x_2 \in \Re^{N_1 N_2 \times N_t}$ contain a set of N_t observations of the random vector, $X_2 \in \Re^{N_1 N_2 \times 1}$, similarly re-ordered, which represents the recognition class. We wish to find the parameters, ω of a mapping, $g(\omega, X):\Re^{N_1N_2 \times 1} \rightarrow \Re$ such that we may discriminate the recognition class from the rejection class. Here, it is the mapping function, g, which defines the discriminator topology.

Towards this goal, we wish to minimize the objective function

$$J = \mathrm{E}(g(\omega, X_1)^2)$$

over the mapping parameters, ω, subject to the system of constraints

$$g(\omega, x_2) = d^{\mathrm{T}}, \tag{2}$$

where $d \in \Re^{N_t \times 1}$ is a column vector of desired outputs. It is assumed that the mapping function is applied to each column of x_2, and $E(\ )$ is the expected value function.

Using the method of Lagrange multipliers, we can augment the objective function as

$$J = \mathrm{E}(g(\omega, X_1)^2) + (g(\omega, x_2) - d^{\mathrm{T}})\lambda, \tag{3}$$

where $\lambda \in \Re^{N_t \times 1}$ is a vector whose elements are the Lagrange multipliers, one for each constraint. Computing the gradient with respect to the mapping parameters yields

$$\frac{\partial J}{\partial \omega} = 2\mathrm{E}\left(g(\omega, X_1)\left(\frac{\partial g(\omega, X_1)}{\partial \omega}\right)\right) + \frac{\partial g(\omega, x_2)}{\partial \omega}\lambda. \tag{4}$$

Equation (4) along with the constraints of equation (2) can be used to solve for the optimal parameters, ω^{o}, assuming our constraints form a consistent set of equations. This is, of course dependent on the mapping topology. In our previous work we have shown that with a linear topology and a suitable estimator of the second-order statistics of the rejection class solutions to equation (4) will yield the MACE filter as well as other optimized correlators.[1] For arbitrary nonlinear mappings it will, in general, be very difficult to solve for globally optimal parameters analytically. Our purpose is instead to develop adaptive training algorithms which are practical and yield improved performance over the linear mappings. It is through the implicit description of the rejection class by its second-order statistics from which we have developed an efficient method extending the MACE filter and other related correlators to nonlinear topologies such as neural networks.

Our goal, therefore, is to find mappings, defined by a topology and a parameter set, which improve upon the performance of the MACE filter in terms of generalization while maintaining a sharp constrained peak in the center of the output plane for images in the recognition class. In prior work, we approximated the objective function of equation (3) with the following objective function,

$$J = (1-\beta)\mathrm{E}(g(\omega, X_1)^2) + \beta[g(\omega, x_2) - d^{\mathrm{T}}][g(\omega, x_2) - d^{\mathrm{T}}]^{\mathrm{T}}, \tag{5}$$

which relaxes the equality constraints, but allows for the parameters of the mapping function to be solved adaptively. Varying β in the range $[0, 1]$ controls the degree to which the average response to the rejection class is emphasized versus the variance about the desired output over the recognition class. Other researchers have proposed similar objective functions and relaxations of the equality constraints within the context of a linear topology.[9,14] Our purpose here is to allow for a gradient search method within a nonlinear topology.

As in the linear case, we can only estimate the expected variance of the output due to the rejection class. In the MACE (or SMACE[15]) filter formulation, X_1 is characterized by all 2D circular (or linear) shifts of the recognition class away from the origin. This term can be estimated with a sampled average over the exemplars, x_2, for all such shifts. As already stated this leads to a computationally intensive gradient search method which trains exhaustively over the entire output plane. Another equivalent characterizations of the rejection class, evident in the statistical formulation, is that we are minimizing the response of the nonlinear filter to images with the same second order statistics of the rejection class. If the exemplars have been preprocessed as in figure 1 then the rejection class can be represented with random white images at the input to the MLP. It is this characterization of the rejection class as random white sequences that enables us to efficiently train the MLP such that the qualities of the MACE filter are maintained (i.e. sharp constrained peak in the center, low variance elsewhere) and a more powerful nonlinear discriminant function, as measured by classification performance, is obtained. Furthermore, we need not train over the entire output plane exhaustively, but rather on the centered exemplars which represent the recognition class and random white noise images which characterize the rejection class. The result is a training algorithm which is of order $N_t + N_{ns}$, where N_t is the number of training exemplars in the recognition class and N_{ns} is the number of white noise sequences rather than $N_t N_1 N_2$. The number of white noise images can be kept relatively small by increasing the parameter β of equation (5).

4. Experimental Results

We now present experimental results which illustrate the techniques and potential pitfalls described in our earlier work[1] as well as recent advances to this method. There are two significant outcomes in the new methods. The first result, in which we restrict the rejection class to a subspace, yields a significant decrease in the convergence time. The second result, in which we borrow from the idea of using the interior of the convex hull to represent the rejection class[16], yields significantly better classification performance.

In these experiments we use the same data depicted in figure 2. As in the previous experiments images from vehicle 1a will be used as the training set. Vehicles 1b and 1c will be used as the recognition class while vehicles 2a and 2b will be used as a rejection/confusion class for testing purposes. In each case comparisons will be made to a baseline linear filter.

Specifically, in all cases the value of α for the linear filter is set to 0.99. The aspect separation between training images is 2.0 degrees. This results in 41 training exemplars from vehicle 1a. These settings of α and aspect separation were found to give the best classifier performance for the linear filter with this data set.

The nonlinear filter will use the same pre-processor as the linear filter (i.e. $\alpha = 0.99$). The MLP structure is shown at the bottom of figure 1. It accepts an $N_1 N_2$ input vector (a preprocessed image reordered into a column vector), followed by two hidden layers, and a single output node. The output as a function of the input vector can be written

$$y = \sigma(W_3^{\mathrm{T}} \sigma(W_2^{\mathrm{T}} \sigma(W_1^{\mathrm{T}} x) + \varphi)), \qquad (6)$$

where $\sigma(\ )$ is the hyperbolic tangent function and the parameters

$$W_1 \in \Re^{N_1 N_2 \times 2} \qquad W_2 \in \Re^{2 \times 3} \qquad W_3 \in \Re^{3 \times 1} \qquad \varphi \in \Re^{3 \times 1}$$

are to be determined through gradient search. The gradient search technique used in all cases will be the well known backpropagation algorithm.

Another aspect of this architecture to note is that the mapping is linear to the points u_1 and u_2 in figure 1. Since the pre-processor is also linear, these operations can be combined after training is complete so that the trained system operates on the original image space.

4.1 *Experiment I - noise training*

As stated, the rejection class is characterized by white noise sequences at the input to the MLP. The recognition class is characterized by the exemplars. It is from these white noise sequences that the MLP, through the learning algorithm, captures information about the rejection class. So it would seem a simple matter, during the training stage, to present random white noise sequences as the rejection class exemplars. In our earlier reporting we observed that with this method of training the linear solution was a strong attractor. The first experiment is designed to illustrate this point.

Figure 7 shows the peak output response taken over all images of vehicle 1a for both the linear (top) and nonlinear (bottom) filters. In the figure we see that for the linear filter the peak constraint (unity) is met exactly for the training exemplars with degradation for the between aspect exemplars. It should be noted that were the pure MACE filter criterion used (α equal to unity), the peak in the output plane is guaranteed to be at the constraint location.[3] It turns out that for this data set the peak output also occurs the constraint location for the training images, however, with $\alpha = 0.99$ it was not guaranteed. Examination of the peak output response for the nonlinear filter the constraints are met very closely (but not exactly) for the training exemplars also with degradation in the peak output response at between aspect locations. The degradation for the nonlinear filter is noticeably less than in the linear case and so in this regard it has outperformed the linear filter.

Figure 8 shows the output plane response for a single image of vehicle 1a (not one used for computing the filter coefficients) for the linear filter (top) and the nonlinear filter (bottom). Again in this figure we see that both filters result in a noticeable peak when the image is centered on the filter and a reduced response when the image is shifted. The reduction in response to the shifted image is again noticeably better in the nonlinear filter than in the linear filter. Such would be found to be true for all images of vehicle 1a and so in this regard we can again say that the nonlinear filter had outperformed the linear filter.

However, as we have already illustrated, these measures are not sufficient to predict classifier performance alone and are certainly not sufficient to compare linear systems to nonlinear systems. This point is made clear in table III which summarizes the classifier performance at two probabilities of detection for all of the experiments reported here when vehicles 1b and 1c are used as the recognition class and vehicles 2a and 2b are used for the rejection class. At this point we are only interested in the results pertaining to the linear filter (our baseline) and nonlinear filter results for experiment I. This table shows that the classifier performance for the linear filter and nonlinear filters are nominally the same, despite what may be perceived to be better performance in the nonlinear filter with regards to peak response over the training vehicle and reduced output plane response to shifts of the image. Furthermore, if we examine figure 9, which shows the ROC curve for both filters we see that they overlay each other. From a classification standpoint the two filters are equivalent.

The cause of this result is best explained by figure 10. Recall the points u_1 and u_2 labeled in figure 1. We can view these outputs as a feature space, that is, the MLP discriminant function can be superimposed on the projection of the input image onto this space. Mathematically this can be written

$$W_1^{\mathrm{T}} x = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \mathbf{u} \qquad y = \sigma(W_3^{\mathrm{T}} \sigma(W_2^{\mathrm{T}} \sigma(\mathbf{u}) + \varphi)) . \tag{7}$$

Figure 10 shows this projection for the training set (top) and the testing set (bottom). What is significant in the figure is that although the discriminant as a function of the vector **u** is nonlinear, the projection of the images lie on a single curve in this feature space. Topologically this filter can put into one-to-one correspondence with a linear projection. This results from the fact that the linear solution is a strong attractor and modifications to the learning algorithm are necessary to avoid it. This is not to say that the linear solution is undesirable, but under the optimization criterion it can be computed in closed form. Furthermore, in a space as rich as the ISAR image space it is unlikely that the linear solution will give the best classification performance.

Table III. Comparison of ROC classifier performance for to values of P_d. Results are shown for the linear filter versus four different types of nonlinear training. N: white noise training, G-S: Gram-Schmidt orthogonalization, subN: PCA subspace noise, C-H: convex hull rejection class.

P_d (%)	**P_{fa} (%)**				
	linear filter	**nonlinear filter, experiments I-IV**			
		I (N)	**II (N, G-S)**	**III (subN, G-S)**	**IV (subN, G-S, C-H)**
80	4.37	4.37	3.74	2.81	2.45
99	42.43	41.87	27.15	26.52	15.33

4.2 *Experiment II - noise training with an orthogonalization constraint*

The modification we suggested previously was to impose orthogonality on the columns of W_1 through a Gram-Schmidt process.[1] Since we are working in a pre-whitened image space, this condition is sufficient to assure the outputs in the feature space will be uncorrelated over the rejection class. Mathematically this can be shown as

$$
\begin{aligned}
E\{\mathbf{uu}^{\mathrm{T}}\} &= E\left\{W_1^{\mathrm{T}} x x^{\mathrm{T}} W_1 \middle| X_1\right\} \\
&= W_1^{\mathrm{T}} E\left\{x x^{\mathrm{T}} \middle| X_1\right\} W_1 \\
&= \begin{bmatrix} w_1^{\mathrm{T}} E(x x^{\mathrm{T}} | X_1) w_1 & w_1^{\mathrm{T}} E(x x^{\mathrm{T}} | X_1) w_2 \\ w_2^{\mathrm{T}} E(x x^{\mathrm{T}} | X_1) w_1 & w_2^{\mathrm{T}} E(x x^{\mathrm{T}} | X_1) w_2 \end{bmatrix} \\
&= \begin{bmatrix} w_1^{\mathrm{T}} (\sigma^2 I) w_1 & w_1^{\mathrm{T}} (\sigma^2 I) w_2 \\ w_2^{\mathrm{T}} (\sigma^2 I) w_1 & w_2^{\mathrm{T}} (\sigma^2 I) w_2 \end{bmatrix} \\
&= \sigma^2 \begin{bmatrix} \|w_1\|^2 & 0 \\ 0 & \|w_2\|^2 \end{bmatrix}
\end{aligned}
$$

where $w_1, w_2 \in \Re^{N_1 N_2 \times 1}$ are the columns of W_1.

The results of the training with this modification are shown in figure 11 which is the resulting feature space as measured at u_1 and u_2. From this figure we can see that not only is the discriminant function a nonlinear function of u_1 and u_2, but that different information is being extracted with regards to the both rejection and recognition classes. The bottom of the figure, showing the projection of a random sampling of the test vehicles (all 1282 would be too dense for plotting) show that both features are useful for separating vehicle 1 from vehicle 2. Examination of table III (column II in the nonlinear results) shows that at the two detection probabilities of interest improved false alarm performance has been obtained. Figure 12 shows the ROC curve for the resulting filter. It is evident that the nonlinear filter is a uniformly better test for classification.

We also show in figure 13 the same image response for this filter as was shown in figure 8. A noticeable peak at the center of the output plane has been achieved. This importance of this quality of the test is that a sharp peak is needed for localized detection.

In this way the characterization of the rejection class by its second order statistics, the addition of the orthogonality constraint at the input layer to the MLP and the use of a nonlinear topology has resulted in a better classification test.

4.3 *Experiment III - subspace noise training*

The remaining experiments illustrate recent advances to this technique. One of the issues of training nonlinear systems is the convergence time. Training methods which require overly long training times are not of much practical use. We have already shown how to reduce the training complexity by recognizing that we can sufficiently describe the rejection class with white noise sequences. We now show a more compact description of the rejection class which leads to shorter convergence times, as demonstrated empirically. This description relies on the well known singular value decomposition (SVD).

We view the random white sequences as stochastic probes of the performance surface in the whitened image space. The classifier discriminant function is, of course, not determined by the rejection class alone. It is also affected by the recognition class. We have shown previously that the white noise sequences enable us to probe the input space more efficiently than examining all shifts of the recognition exemplars.[1,13] However, we are still searching a space of dimension equal to the image size, $N_1 N_2$.

One of the underlying premises to a data driven approach is that the information about a class is conveyed through exemplars. In this case the recognition class is represented by $N_t < N_1 N_2$ exemplars placed in the data matrix $x_2 \in \Re^{N_1 N_2 \times N_t}$. It is well known that x_2, if it is full rank, can be decomposed with the SVD as

$$x_2 = U \Lambda V^{\mathrm{T}}, \tag{8}$$

where the columns $U \in \Re^{N_1 N_2 \times N_t}$ are an ortho-normal basis that span the column space of the data matrix, Λ are the singular values, and V is an orthogonal matrix. This decomposition has many well known properties including compactness of representation for the columns of the data matrix.[17] Indeed, as has been noted by Gheen[18], the SDF can be written as a function of the SVD of the data matrix.

$$h_{SDF} = U \Lambda^{-1} V^{\mathrm{T}} d \tag{9}$$

We will use this representation to further refine our description of the rejection class for training. As we stated, the underlying assumption in a data driven method, is that the data matrix x_2 conveys information about the recognition class, any information about the recognition class outside the space of the data matrix is not attainable from this perspective. The information certainly exists, but there is no mechanism by which to include it in the determination of the discriminant function within this framework. This does however lead to a more efficient description of the rejection class. We can modify our optimization criterion to reduce the response to white sequences as they are projected into the N_t-dimensional subspace of the data matrix. Effectively this reduces the search for a discriminant function in an $N_1 N_2$-dimensional space to an N_t-dimensional subspace.

The adaptation scheme of backpropagation allows a simple mechanism to implement this constraint. The adaptation of matrix W_1 at iteration k can be written as

$$W_1(k+1) = W_1(k) + x_i(k) \varepsilon'^{\mathrm{T}}_i(k) \tag{10}$$

where ε'_i is a column vector derived from the backpropagated error and $x_i(k)$ is the current input exemplar from either class presented to network which, by design, lies in the subspace spanned by the columns of U. From equation (10) if the rejection class noise exemplars are restricted to lie in the data space of x_2 which can be achieved by projecting random vectors of size N_t onto the matrix U above, and W_1 is initialized to be a random projection from this space we will be assured that the columns of W_1 only extract information from the data space of x_2. The search for a discriminant function is now reduced from within an $N_1 N_2$-dimensional space to a search from within an N_t-dimensional space. Due to the dimensionality reduction achieved we would expect the convergence time to be reduced.

This is the method that was used for the third experiment. Rejection class noise exemplars were generated by projecting a random vector, $n \in \Re^{N_t \times 1}$, onto the basis U by $x_{rej} = Un$. In figure 14 the resulting discriminant function is shown as in the previous experiments and the result is similar to experiment II. The classifier performance as measured in table III and the ROC curve of figure 15 are also nominally the same.

There are, however, two notable differences. Examination of figure 16 shows that the output response to shifted images is even lower allowing for better localization. This condition was found to be the case throughout the data set. Of more significance is the result shown in figure 17 in which we compare the learning curves of all of the experiments presented here. In this figure the dashed and dashed-dot lines are the learning curves for experiments II and III, respectively. In this case the convergence rate was increased nominally by a factor of three, from 100 epochs to approximately 30 epochs. Here an epoch represents one pass through all of the training data.

4.4 *Experiment IV - convex hull approach*

In this experiment we present a technique which borrows from the ideas of Kumar *et al.*[16] This approach designed an SDF which rejects images which are away from the boundary of the convex hull of the training set. The convex hull of a set $\{x_1, x_2, \ldots, x_{N_t}\}$ is defined as all points which can be represented as

$$x = \sum_{i=1}^{N_t} a_i x_i$$

where the a_i ‘s are constrained to satisfy

$$a_i \geq 0 \qquad \sum_{i=1}^{N_t} a_i = 1 .$$

It was pointed out that by Kumar *et al* that when the peak constraints for the SDF (or any of the linear distortion invariant filters) are all set to unity, points in the interior of the convex hull over the training exemplars are recognized as well as the those near the extremal points. This would include, for example, an image which is the mean of the training exemplars. Examination of imagery derived from points that are closer to the interior of the convex hull, rather than near the boundary are not representative of the recognition class.

It was suggested that a way to mitigate this property was to set the desired output over the training set to be complex, unity magnitude and mean zero. The magnitude of the output was then used as the response. In this way only points near the boundary of the convex hull are recognized.

The approach taken here is similar in that exemplars from the interior of the convex hull are used as representative of the rejection class. The difference is that this description is included in the learning process without à priori determining the decision surface (e.g. magnitude of the correlator output). It is the nonlinear adaptive process which determines how to separate the recognition class exemplars from the images derived from the convex hull. The result is significantly improved classification.

In this experiment we continue to use random noise projected onto the basis defined by columns of the matrix U as in experiment III. In addition, convex hull exemplars are generated by projecting a random vector $a \in \Re^{N_t \times 1}$ onto the data matrix x_2. In keeping with the basis for this approach, that elements of the convex hull that are distant from the extremal points (the training exemplars) do not convey information about the recognition class, we imposed a further restriction on the coefficients a_i, namely

$$\frac{0.9}{N_t} \le a_i \le \frac{1.1}{N_t}.$$

This restriction assures that none of the generated convex hull exemplars lie too close to one of the recognition class training exemplars. Rejection class exemplars from within the convex hull are randomly generated throughout the training from $x_{rej} = x_2 a$. Another property of these rejection class exemplars is that they also lie in the subspace of the data matrix x_2.

Examination of table III and the ROC curve of figure 19 show that this method yields significantly improved classification performance. The discriminant function shown in figure 18 is quite different and much more nonlinear than in the previous cases. In the figure the convex hull exemplars are clustered between the subspace noise exemplars and the recognition class exemplars. If this is a general property of the type of data we are using then it may be a powerful method by which to describe the rejection class within the nonlinear framework. More analysis is needed, however, before this conclusion can be made. We do conclude that in this case this method is an effective means by which to characterize the rejection class. The advantage in this technique versus the linear method of Kumar *et al*[16] is that the training learns to separate automatically the recognition class exemplars from the convex hull exemplars as opposed to à priori assigning a complex desired output for each exemplar.

There were, however, some difficulties with this technique which are worth mentioning. Recall that the motivation for using orthogonalization in the input layer was to increase the likelihood that a nonlinear discriminant function was found. When using convex hull exemplars in the rejection class, this may seem unnecessary. In practice, however, it was found that when the orthogonalization was removed, training times became extremely long. Even with orthogonalization we can see from the learning curve (solid line) in figure 17 that convergence took over an order of magnitude longer as in experiment III.

There were also stability issues as well with this type of training. The training became unstable nearly as often as it converged. However, when the training did converge, as in the results shown, the classification performance was always superior. For this reason we believe that this method bears further study.

5. Conclusions

We have discussed in detail a methodology by which linear distortion invariant filtering can be extended to nonlinear adaptive systems. Our analysis showed that for the linear systems the emphasis on the MACE filter criterion, as measured by α in the OTSDF filter formulation, was a better indicator of classification performance than commonly used measures of generalization. This result was important because it highlighted that fact that commonly used measures of generalization should not be the sole basis upon which to compare nonlinear systems to the their linear counterparts since these measures are only weakly coupled to classification performance. The results of experiment I further emphasized this point.

Another result of this analysis was that the emphasis on the MACE criterion in the OTSDF was important to classification performance and for this reason the pre-processor in the nonlinear filter was kept the same as in linear case.

Within the context of the nonlinear filter architecture the design goal was reduced to the search for a discriminant function within a pre-processed image space. The emphasis of our presentation was how to do this efficiently and robustly as measured by convergence rates and classification performance.

We presented two new advances to our method as well as further analysis of the original methodology within a more rigorous validation framework. In the first case we reduced the search space from the dimensionality of the imagery being processed to the dimensionality of the data matrix containing the training exemplars for the recognition class. In the second case we borrowed the idea of the interior of the convex hull over the training exemplars as representative of the rejection class.

The results of the subspace approach yielded slightly better classification performance with significantly faster training times. The results of the convex hull approach yielded significantly better classification performance, however, stability of the learning process was an issue. We feel that the results of the convex hull approach merit further investigation.

In the course of the discussion we presented results with respect to ISAR data. The data chosen represents, in our opinion, a fairly difficult classification problem in the sense that the range of distortions for the ISAR data not only includes rotation in aspect but modifications in the vehicle configuration and differences in the radar depression angle. In spite of these obstacles the nonlinear system generalizes quite well.

Acknowledgments

This research was partially supported by DARPA grant F-33615-97-1-1019.

References

1. J. Fisher and J.C. Principe, “A nonlinear extension of the MACE filter,” *Neural Networks* **8** (7/8), 1131-1141 (1995).

2. C. F. Hester and D. Casasent, “Multivariant technique for multiclass pattern recognition,” *Appl. Opt.* **19**, 1758-1761 (1980).

3. A. Mahalanobis, B.V.K. Vijaya Kumar, and D. Casasent, “Minimum average correlation energy filters,” *Appl. Opt.* **26** (17), 3633-3640 1987.

4. T. Kohonen, *Self-Organization and Associative Memory* (1st ed.), Springer Series in Information Sciences, vol. 8, Springer-Verlag, 1988.

5. J. Fisher and J. C. Principe, “Formulation of the MACE Filter as a Linear Associative Memory”, *Proceedings of the IEEE International Conference on Neural Networks*, **5**, 2934-2937 (1994).

6. K. Funahashi, “On the approximate realization of continuous mappings by neural networks,” *Neural Networks* **2** (3), 183-192, (1989).

7. Casasent, D., and G. Ravichandran, “Advanced distortion-invariant minimum average correlation energy (MACE) filters”, *Appl. Opt.* **31** (8), 1109-1116, (1992).

8. C. Bishop, *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford, p. 11, (1995).

9. Ph. Réfrégier and J. Figue, “Optimal trade-off filter for pattern recognition and their comparison with Weiner approach”, *Opt. Comp. Proc.* **1**, 3-10, (1991).

10. B. V. K. Vijaya Kumar, “Minimum variance synthetic discriminant functions,” *J. Opt. Soc. Am. A* **3** (10), 1579-1584 (1986).

11. Casasent, D., G. Ravichandran, and S. Bollapragada, “Gaussian minimum average correlation energy filters”, *Appl. Opt.* **30** (35), 5176-5181, (1991).

12. Ravichandran, G., and D. Casasent, “Minimum noise and correlation energy filters”, *Appl. Opt.* **31** (11), 1823-1833, (1992).

13. J. Fisher and J.C. Principé, “Experimental results using a nonlinear extension of the MACE filter,” *Optical Pattern Recognition VI,* Proceedings of SPIE **2490**, 41-52 (1995).

14. A. Mahalanobis, B.V.K. Vijaya Kumar, Sewoong Song, S.R.F. Sims, and J.F. Epperson; “Unconstrained correlation filters”, *Appl. Opt.* **33** (33), 3751-3759, (1994).

15. S. I. Sudharsanan, A. Mahalanobis, and M. K. Sundareshan, “A unified framework for the synthesis of synthetic discriminant functions with reduced noise variance and sharp correlation structure”, *Appl. Opt.* **30** (35), 5176-5181, (1991).

16. B. V. K. Vijaya Kumar, J. D. Brasher, C. F. Hester, G. Srinivasan, and S. Bollapragada, “Synthetic discriminant functions for recognition of images on the boundary of the convex hull of the training set”, *Patt. Rec.* **27** (4), 543-548, (1994).

17. Gerbrands, J., “On the Relationships between SVD, KLT, and PCA”, *Pattern Recognition*, **14**, 375-381, (1981).

18. Gheen, G., “Design of considerations for low-clutter, distortion-invariant correlation filters”, *Optical Engineering*, **29** (9), 1029-1032, (1990).

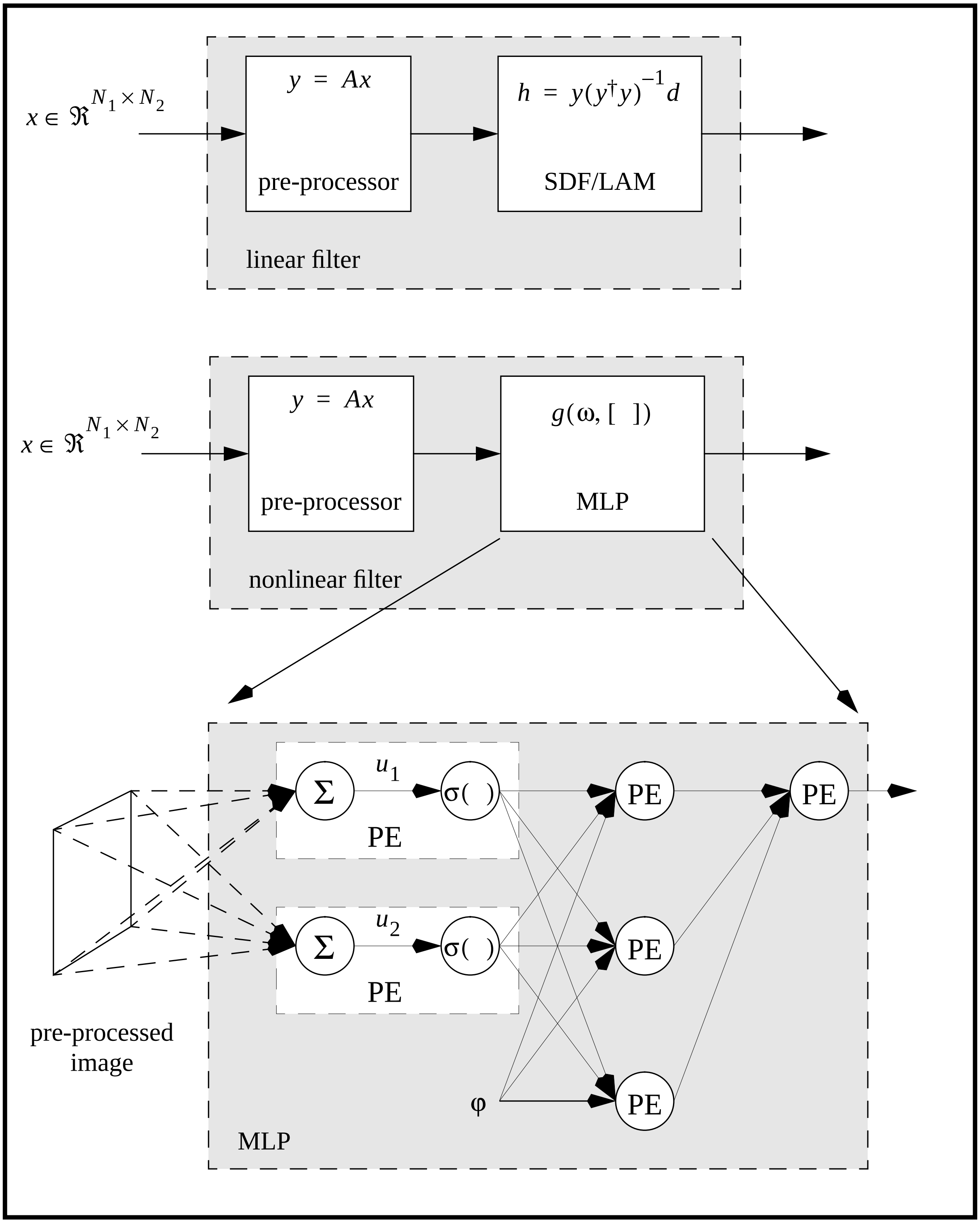

Figure 1. Decomposition of optimized correlator as a pre-processor followed by SDF/LAM (top). Nonlinear variation shown with MLP replacing SDF in signal flow (middle), detail of the MLP (bottom). The linear transformation A represents the space domain equivalent of the spectral pre-processor $(\alpha P_x + (1-\alpha)\overline{P}_x)^{-1/2}$.

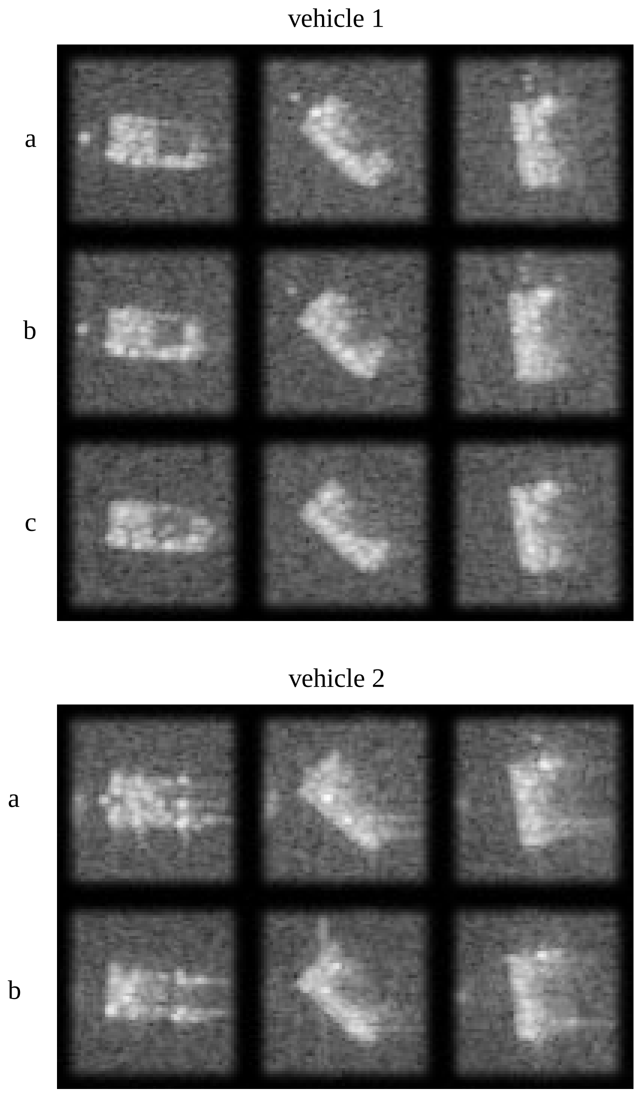

Figure 2. ISAR images of two vehicle types shown at aspect angles of 5, 45, and 85 degrees respectively. Three different vehicles of type 1 (a, b, and c) are shown, while two different vehicles of type 2 (a and b) are shown. Vehicle 1a is used as a training vehicle, while vehicles 1b and 1c are used as the testing vehicles for the recognition class. Vehicles 2a and 2b are used a s confusion vehicles.

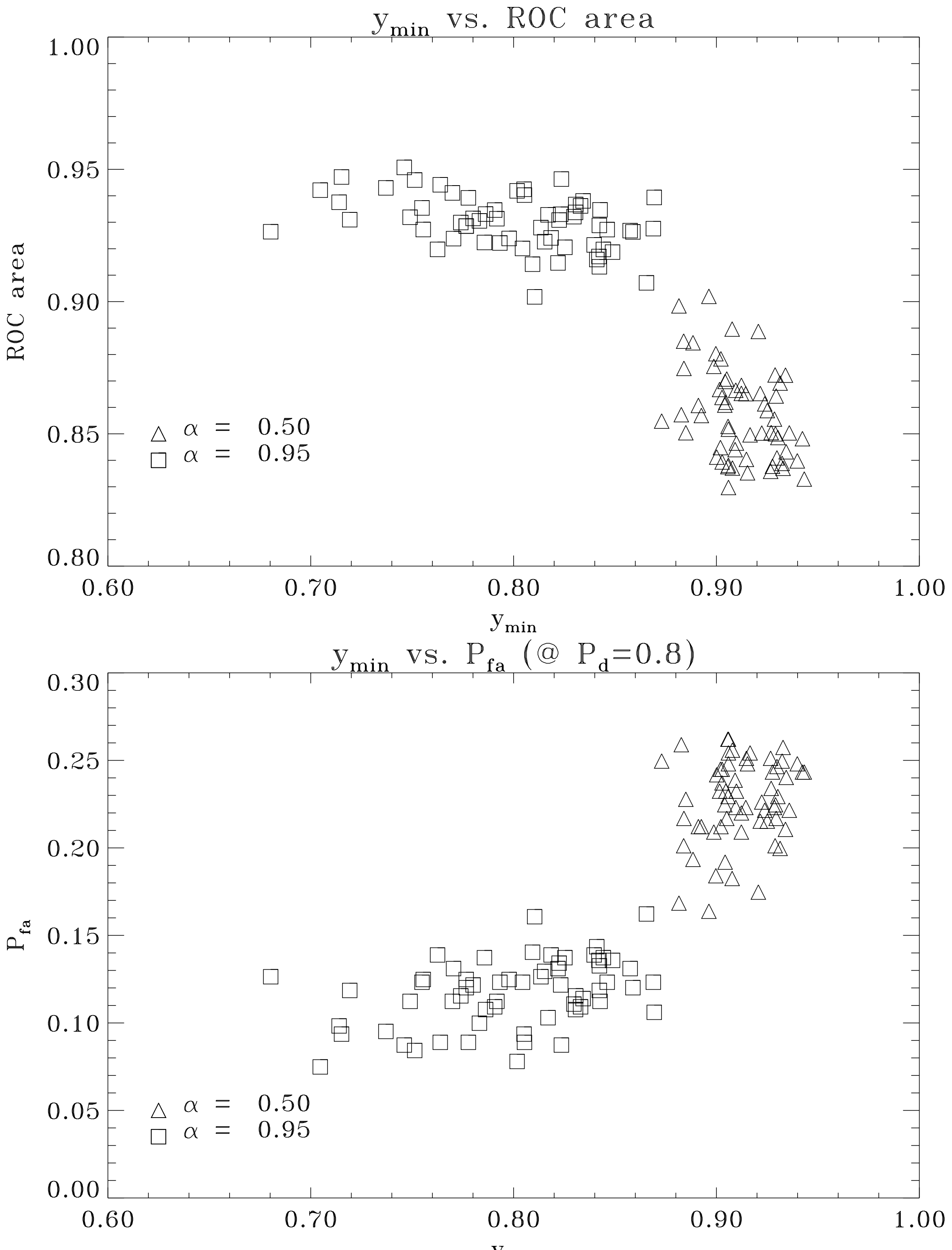

Figure 3. Generalization as measured by the minimum peak response over the training vehicle y_{min} versus classification performance measures (ROC area and P_{fa}@P_d=0.8).

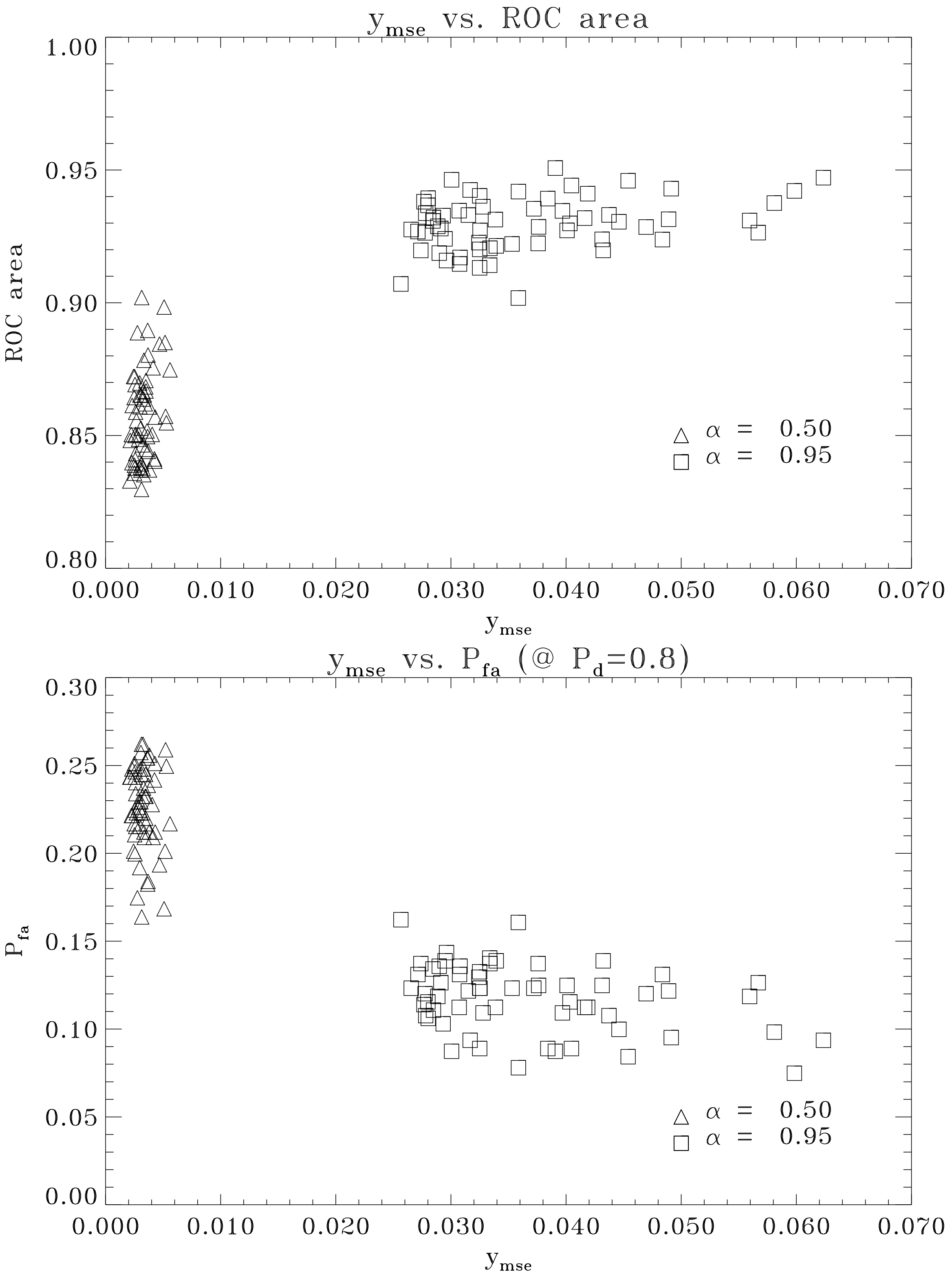

Figure 4. Generalization as measured by the MSE response of the recognition class testing vehicle y_{mse} versus classification performance measures (ROC area and P_{fa}@P_d=0.8).

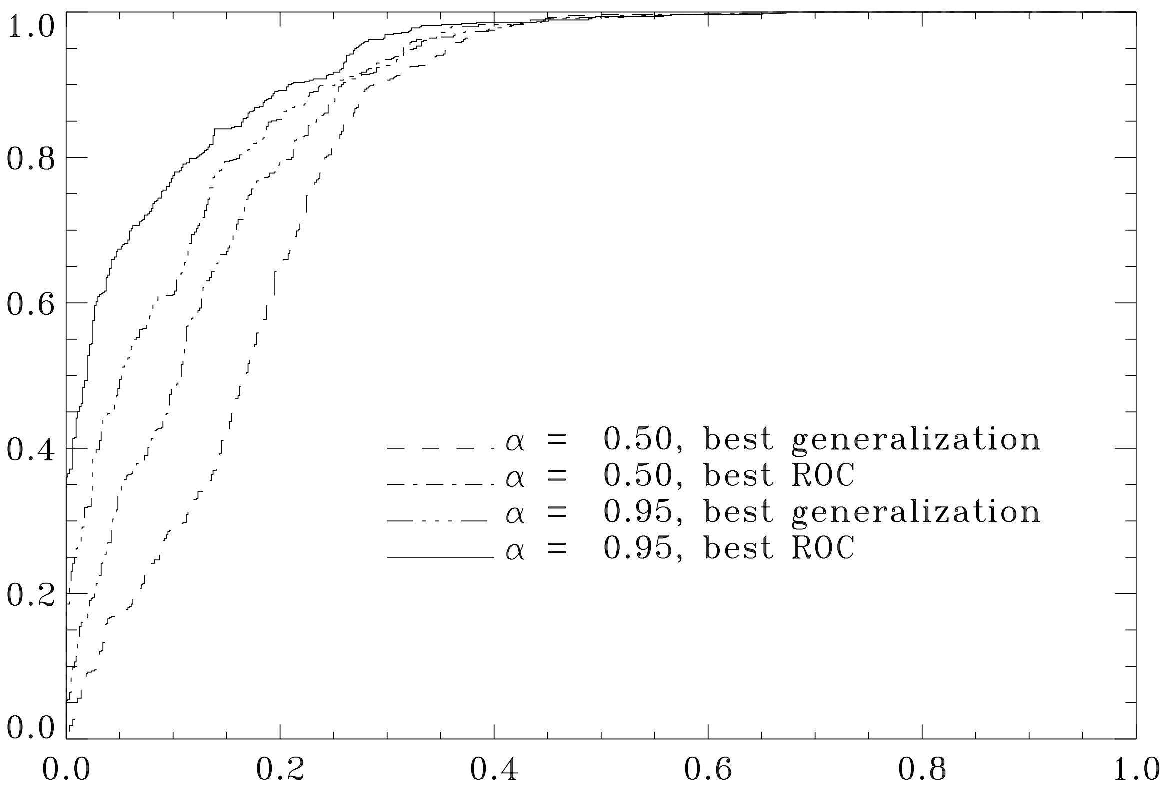

Figure 5. Comparison of ROC curves. The ROC curves for the number of training exemplars yielding the best generalization measure versus the number yielding the best ROC performance for values of α equal to 0.5 and 0.95 are shown.

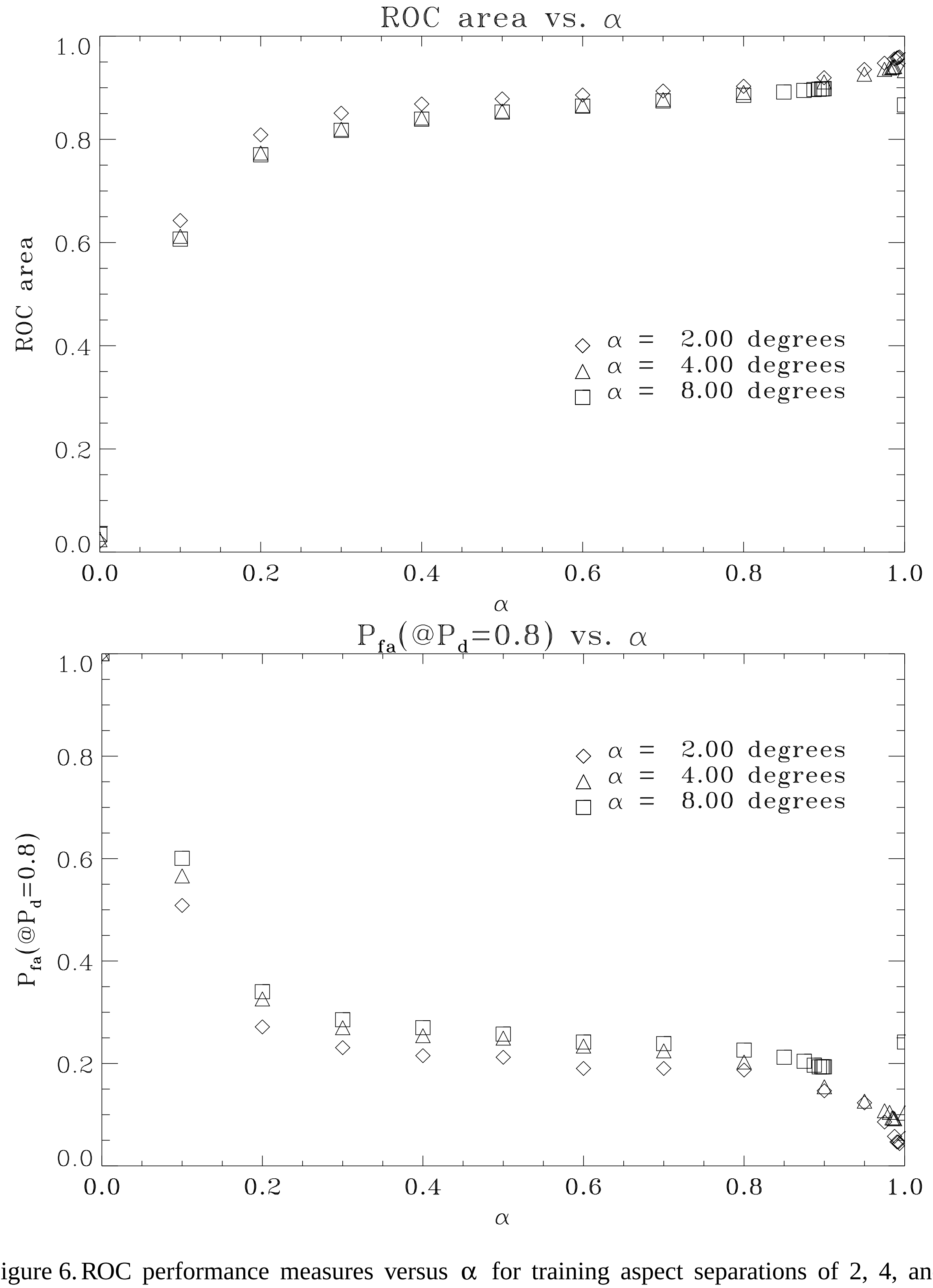

Figure 6. ROC performance measures versus α for training aspect separations of 2, 4, and 8 degrees. These plots indicate that ROC performance is positively related to α.

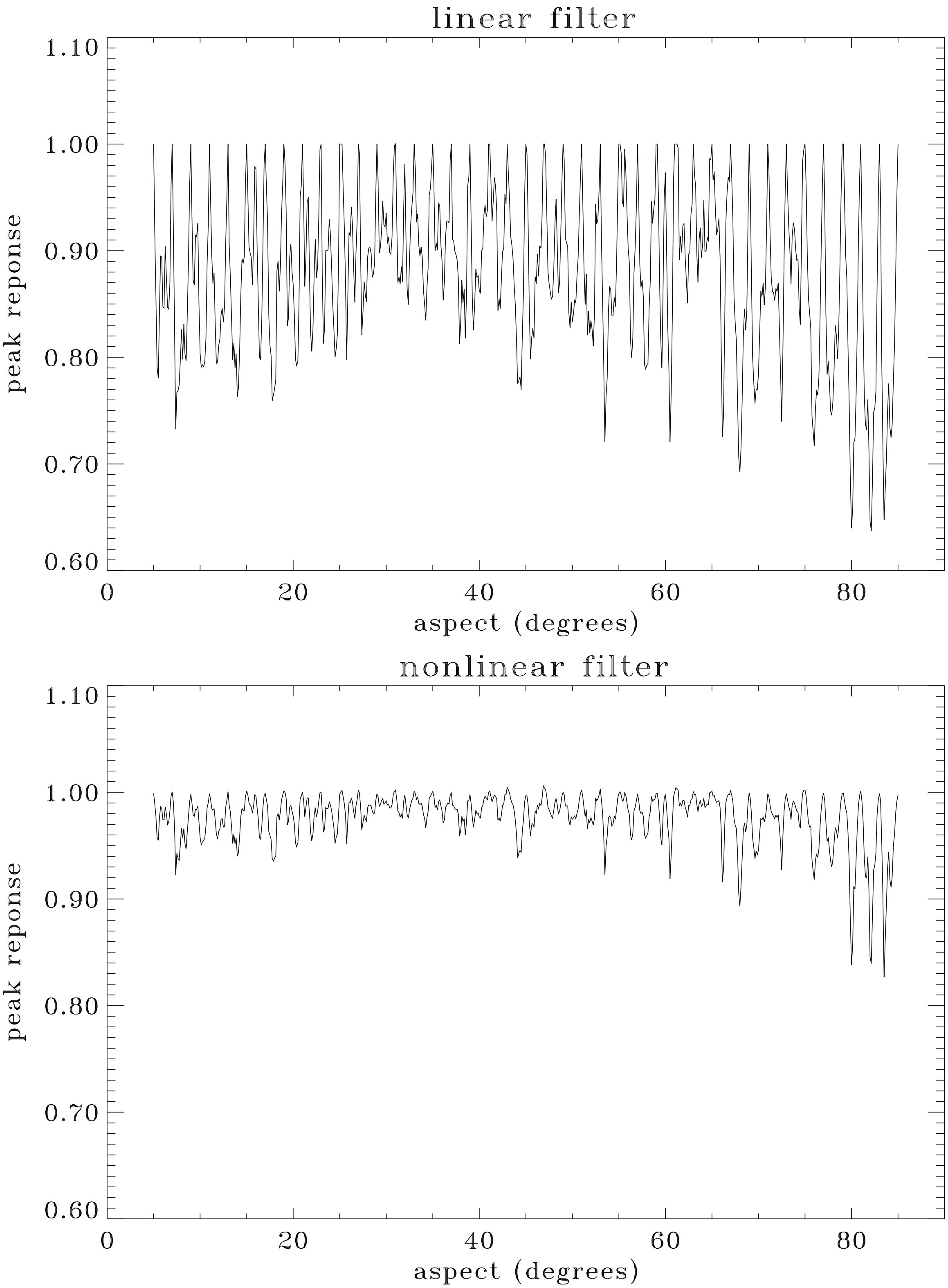

Figure 7. Peak output response of linear and nonlinear filters over the training set. The nonlinear filter clearly outperforms the linear filter by this metric alone.

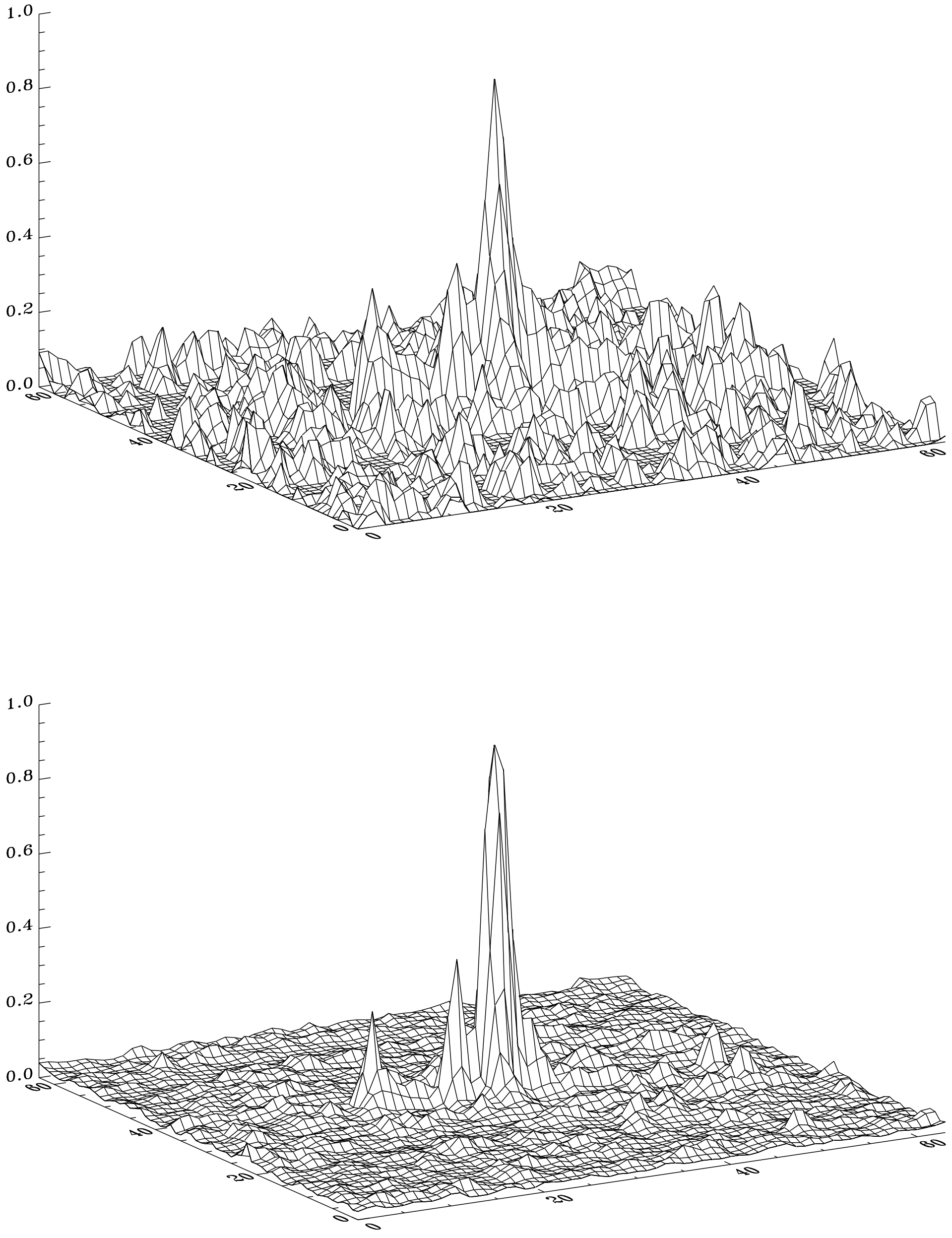

Figure 8. Output response of linear filter (top) and nonlinear filter (bottom) for a single image in the training set.

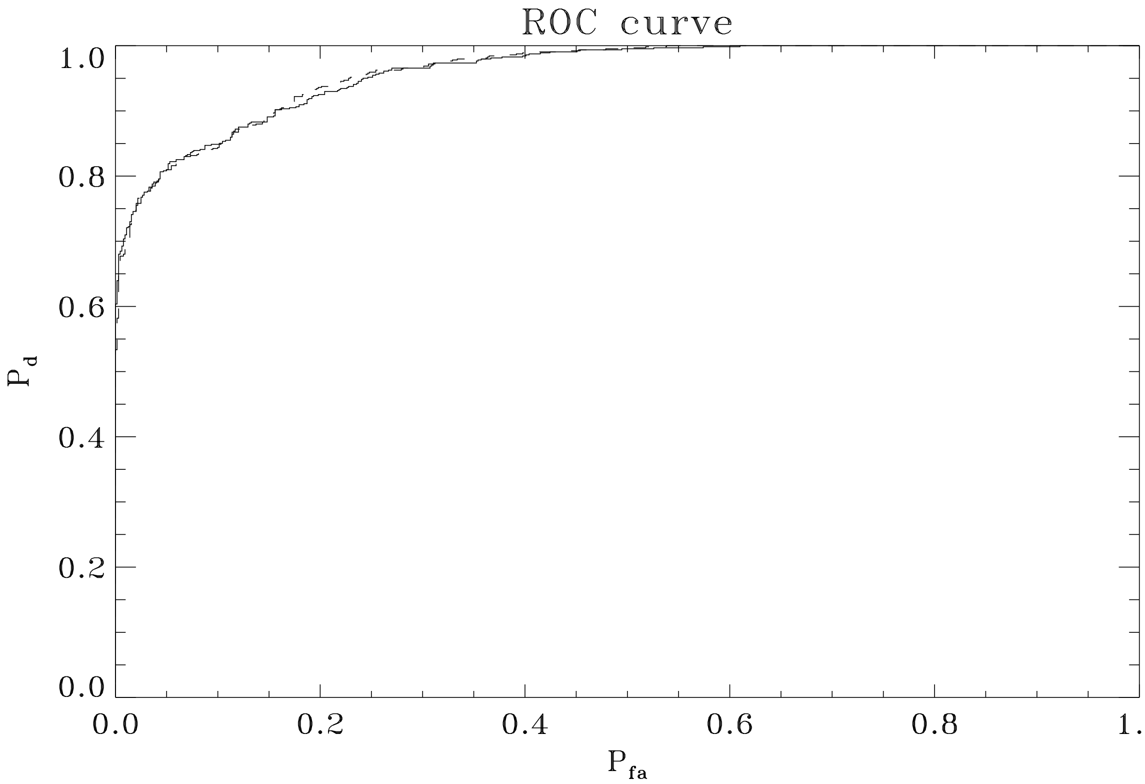

Figure 9. ROC curves for linear filter (solid line) versus nonlinear filter (dashed line). Despite improved performance of the nonlinear filter as measured by peak output response and reduced variance over the training set, the filters are equivalent with regards to classification over the testing set.

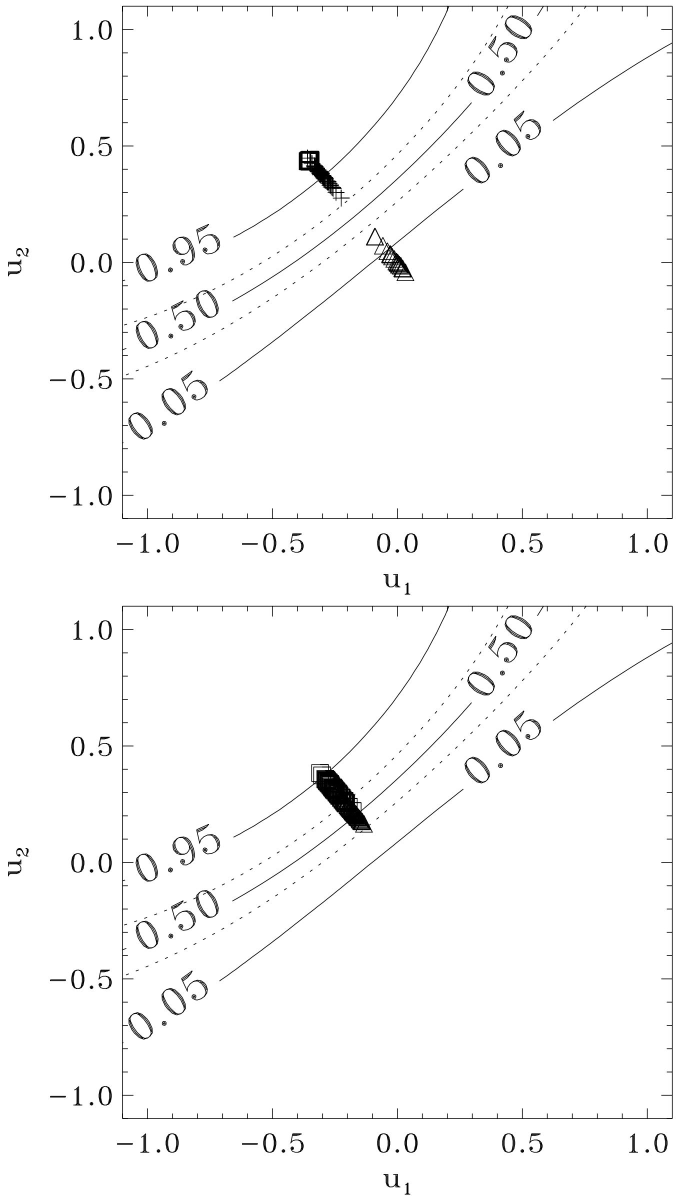

Figure 10.Experiment I: Resulting feature space from simple noise training. Note that all points are projected onto a single curve in the feature space. In the top figure squares are the recognition class training exemplars, triangles are white noise rejection class exemplars, and plus signs are the images of vehicle 1a not used for training. In the bottom figure, squares are the peak responses from vehicles 1b and 1c, triangles are the peak responses from vehicles 2a and 2b.

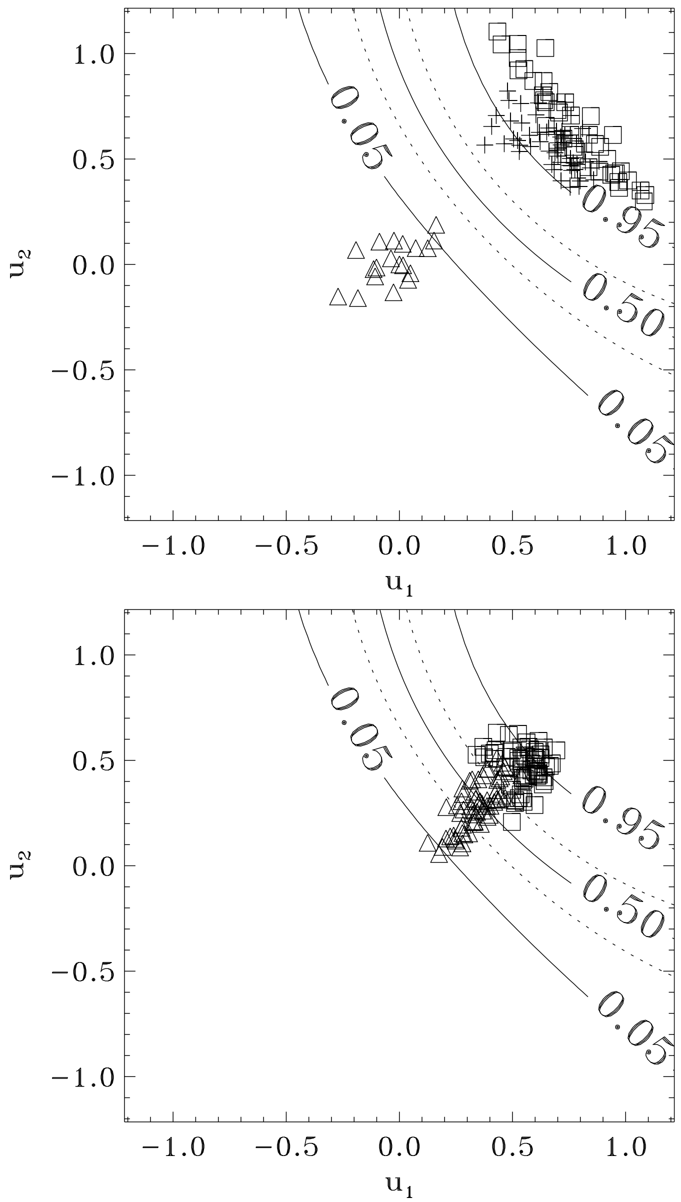

Figure 11.Experiment II: Resulting feature space when orthogonality is imposed on the input layer of the MLP. In the top figure squares are the recognition class training exemplars, triangles are white noise rejection class exemplars, and plus signs are the images of vehicle 1a not used for training. In the bottom figure, squares are the peak responses from vehicles 1b and 1c, triangles are the peak responses from vehicles 2a and 2b.

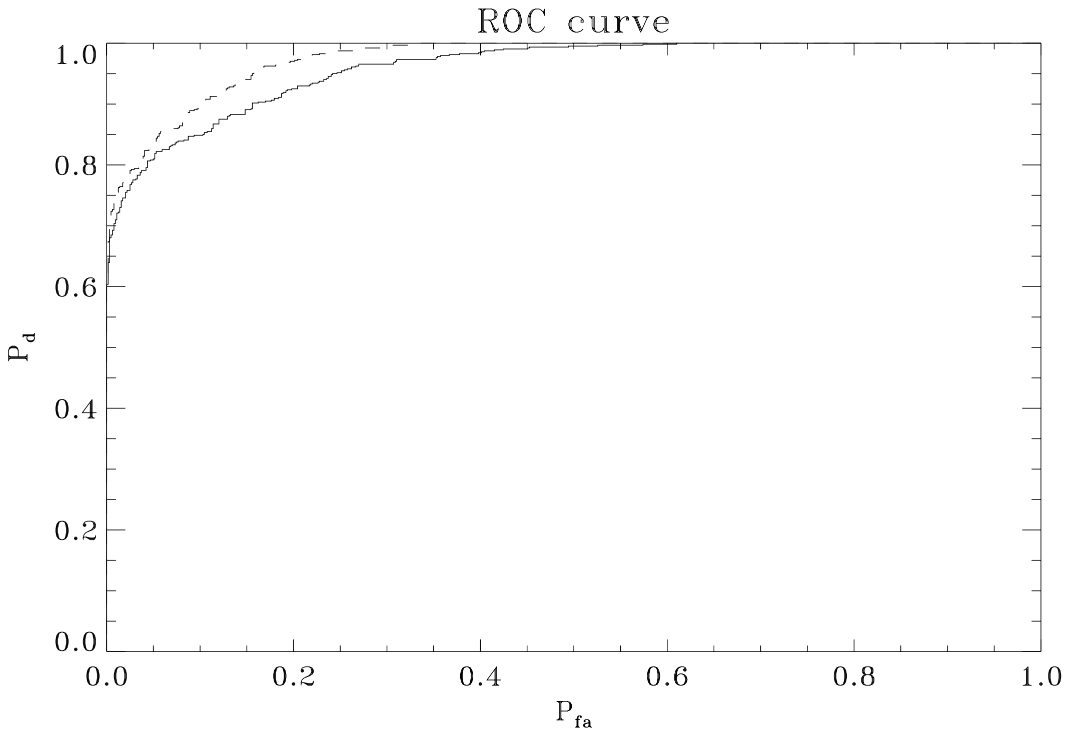

Figure 12.Experiment II: Resulting ROC curve

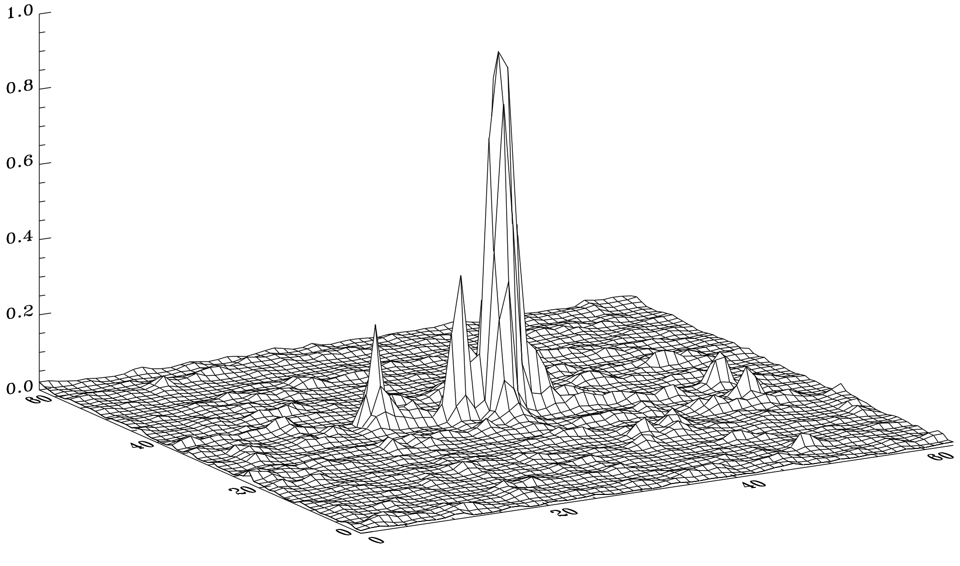

Figure 13.Experiment II: Output response to an image from the recognition class training set.

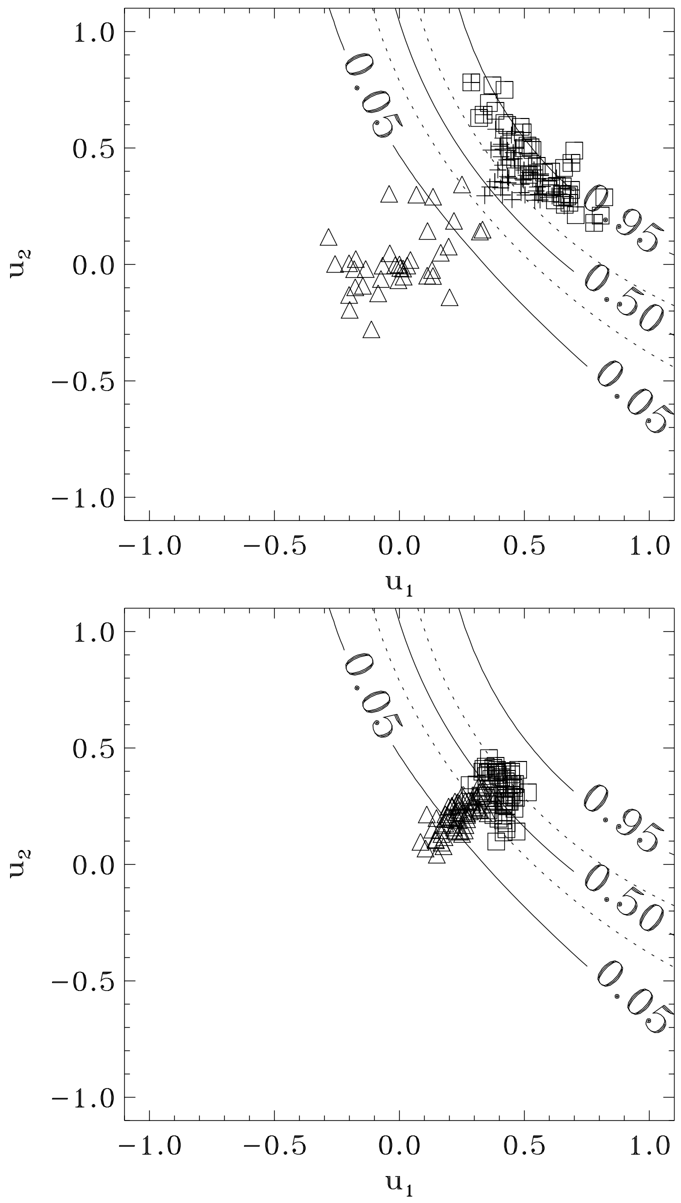

Figure 14.Experiment III: Resulting feature space when the noise rejection class is restricted to the SVD subspace of the data matrix. In the top figure squares are the recognition class training exemplars, triangles are white noise rejection class exemplars, and plus signs are the images of vehicle 1a not used for training. In the bottom figure, squares are the peak responses from vehicles 1b and 1c, triangles are the peak responses from vehicles 2a and 2b.

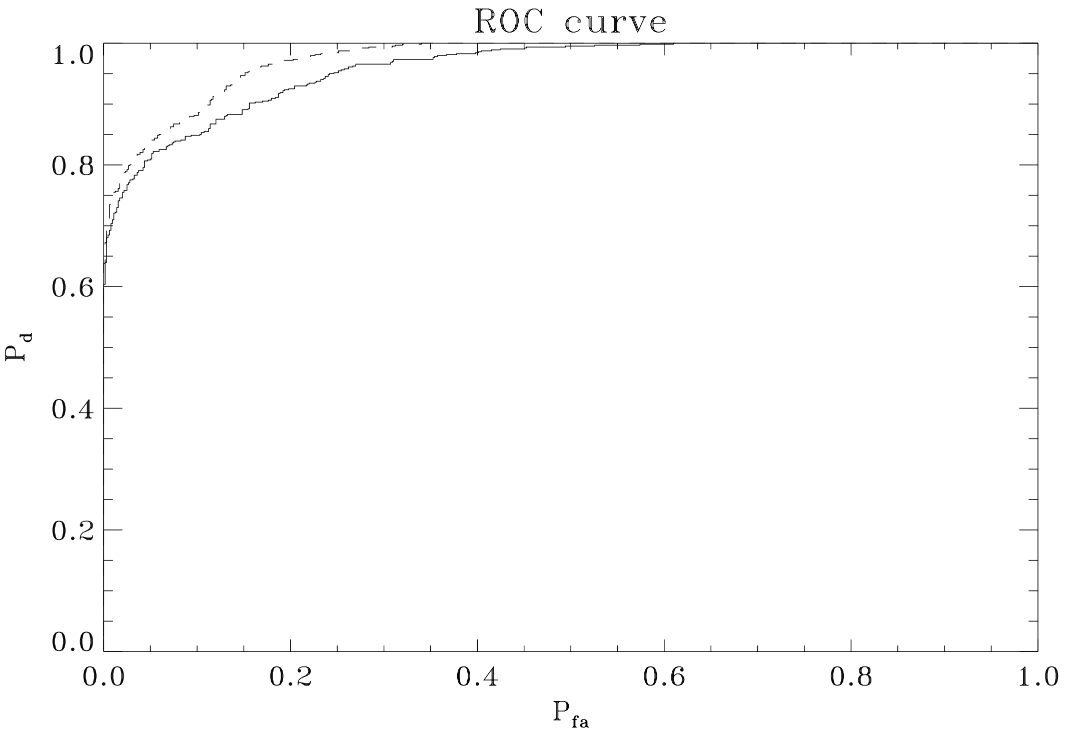

Figure 15.Experiment III: Resulting ROC curve.

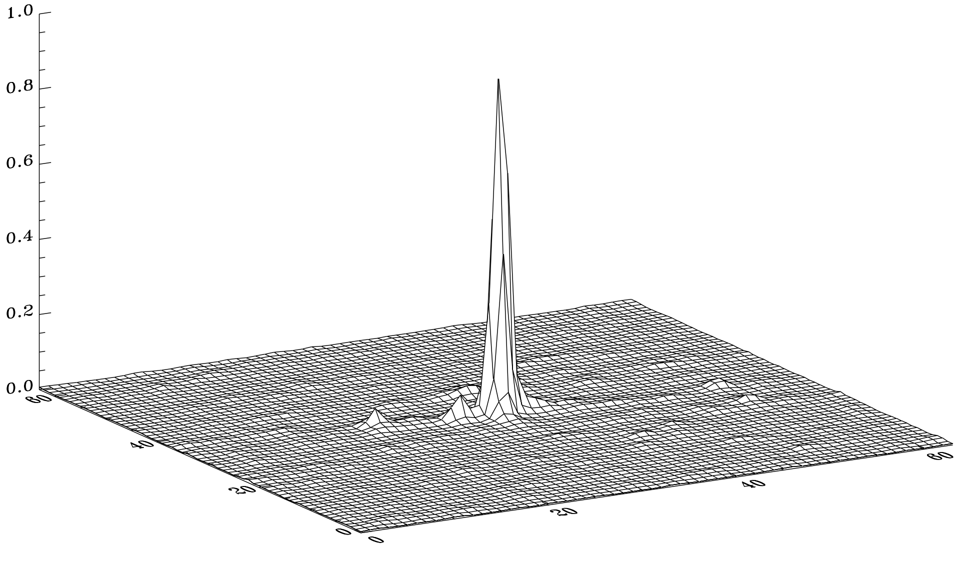

Figure 16.Output response to an image from the recognition class training set.

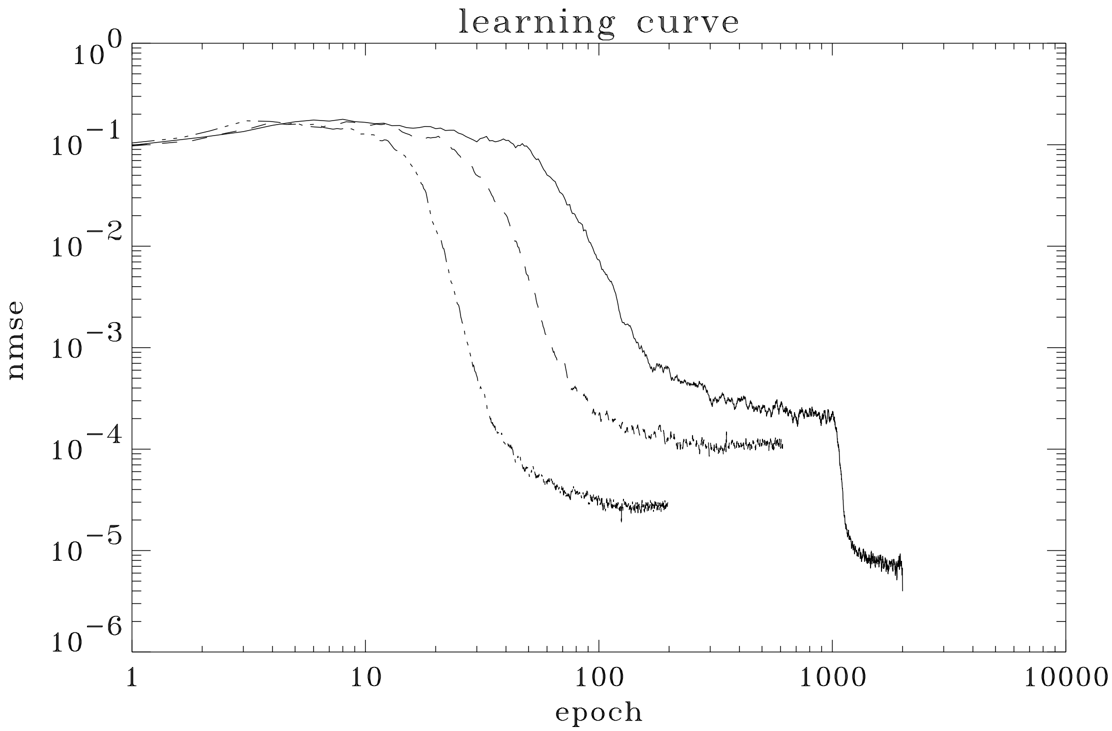

Figure 17.Learning curves for three methods. Experiment II: White noise training (dashed line). Experiment III: subspace noise (dashed-dot line). Experiment IV: subspace noise plus convex hull exemplars (solid line).

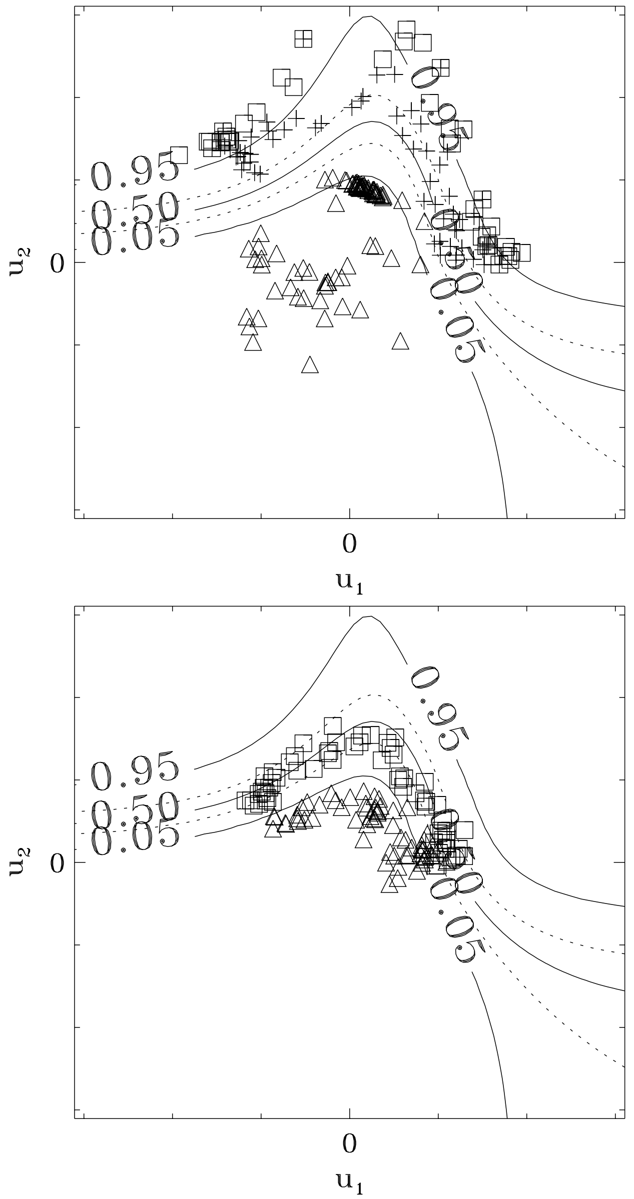

Figure 18.Experiment IV: resulting feature space when convex hull exemplars are added to the rejection class. In the top figure squares are the recognition class training exemplars, triangles are white noise rejection class exemplars, and plus signs are the images of vehicle 1a not used for training. In the bottom figure, squares are the peak responses from vehicles 1b and 1c, triangles are the peak responses from vehicles 2a and 2b.

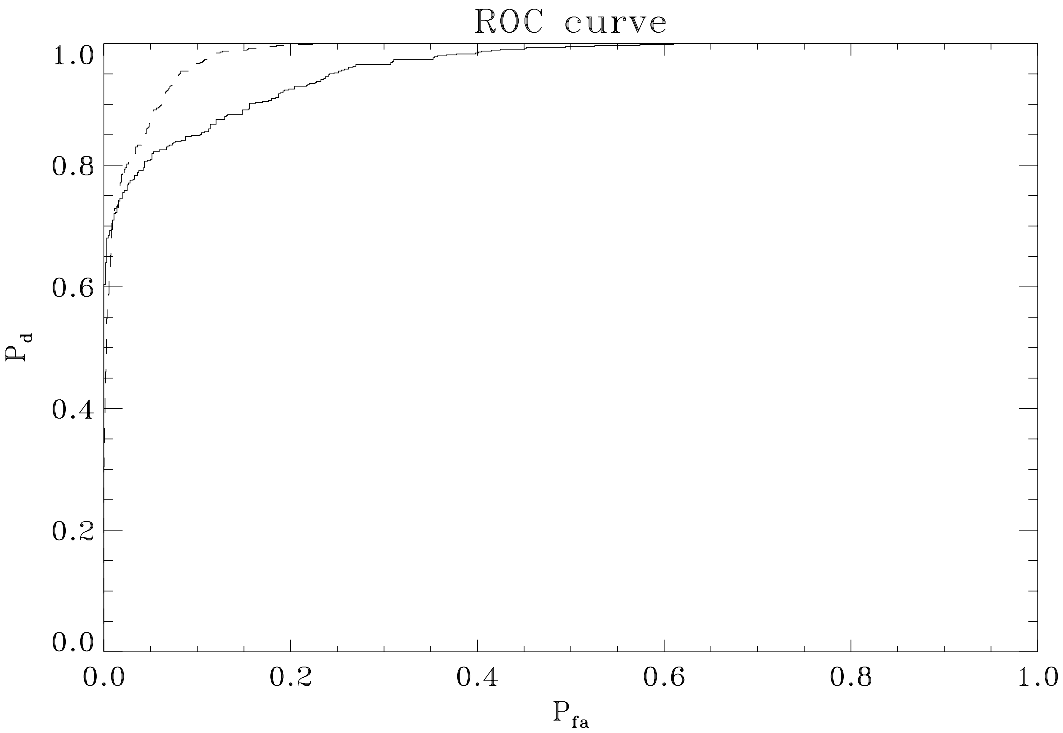

Figure 19.Experiment IV: Resulting ROC curve.