An initial assessment of discriminant surface complexity for power law features

Jeffrey L. Solka, Carey E. Priebe and George W. Rogers

SIMULATION 1992 58: 311
DOI: 10.1177/003754979205800503

An initial assessment of discriminant surface complexity for power law features

Jeffrey L. Solka, Carey E. Priebe, and George W. Rogers
Dahlgren Division
Naval Surface Warfare Center
Dahlgren, Virginia

The detection of man-made objects in natural terrain is important in both the targeting and terminal homing phase of modern warfare. The presence of man-made objects in gray-scale images has been successfully detected using a new class of density estimation neural networks to analyze power law signatures. The complex nature of the discriminant surface relating these features has been elucidated using these adaptive mixture networks.

Keywords: terrain, warfare, man-made objects, neural networks, power law signatures

Introduction

The purpose of our discussion a gray-scale image is a relation that assigns an integer value between 0 and 255 inclusive to each point in pixel space. Therefore each point in pixel space can be specified by an ordered triple (i, j, k) where $i \in [0,511]$, $j \in [0,479]$, and $k \in [0,255]$. Given a gray-scale image one is often interested in focusing attention on the man-made objects in the image. For example the capability to detect the presence of the camouflaged tank among the bushes in Figure 1 is highly desirable. The ability to detect the presence of man-made objects in natural terrain is important in the mission planning/targeting problem, and in the terminal homing phase of autonomous weapons. Techniques which rapidly identify these areas can play a role as part of an automatic image processing system or as an adjunct to wetware components in a traditional system.

Much work has been done in recent years using fractal geometry to model natural objects. This work has included artificial generation of synthetic terrain[1], the use of iterated function systems to compress images,[2] and the use of recursive update rules to generate synthetic plant images.[3] The problem of interest here can be stated as follows: given an image, what subsets of it can be identified as possessing fractal structure and can this fractal structure be used to classify portions of the image.

Previously Mandelbrot defined a fractal as any object whose Hausdorff-Besicovitch dimension exceeds it topological dimension.[4] Two properties that characterize fractals are self similarity and the prescribed variation of a measured quantity of a fractal with scale.

Figure 1. Gray-scale image of tank.

Self similarity can be thought of as the propagation of geometric features across scale space, wherein an image looks the same when viewed from different size perspectives (see Figure 2). The manner in which measured properties vary across scale space is known as Richardson's law:

$$M(\varepsilon) = K\varepsilon^{d-D} \qquad (1)$$

In equation (1) $M(\varepsilon)$ is the measured property at scale

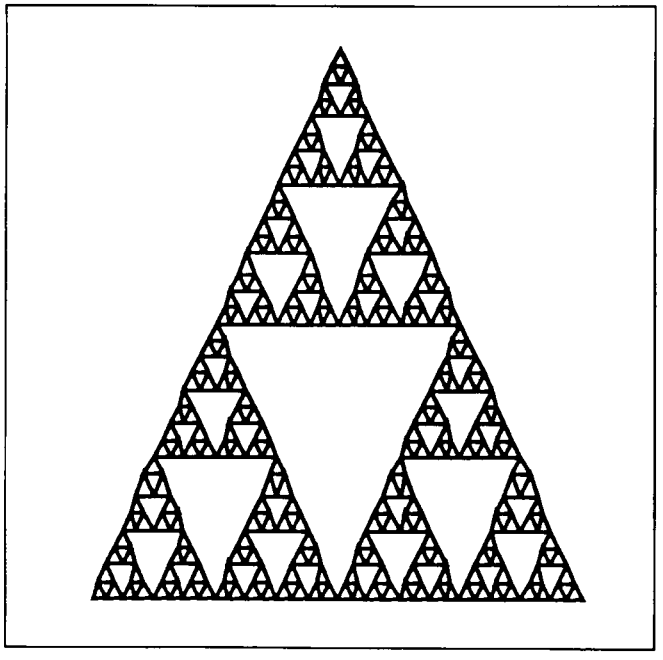


Figure 2. The propagation of geometric features across scale space as illustrated by the Sierpinski Triangle.

ε, K is a constant, d is the topological dimension, and D is the Hausdorff-Besicovitch dimension (i.e. fractal dimension). The natural regions of a gray-scale image are not true fractals, and hence should only obey Richardson's law in distribution.[5] As will be seen, this equation provides us with a method to estimate the fractal dimension D. The fractal dimension is thought to be one of the best measures of image texture currently available.[6]

Our problem, then, is one of discrimination between different object classes based on features extracted from power-law theory. In general, we consider the observations to be independent, identically distributed random variables, with the probability density of the overall (N-class) distribution being

$$D(x) = \sum_{i=1}^{N} \pi_1 \, D_i(x) \qquad (2)$$

where the π_i are the prior probabilities for the individual classes and the D_i are the probability density functions for the individual classes. It suffices for analysis to consider N = 2. A formal motivation for using density estimation in discriminant analysis is obtained via consideration of the asymptotics of the discrimant procedure. Let

$$D_0(x) = \pi_1 D_1(x) - \pi_2 D_2(x)$$

and consider our decision procedure for discrimination, given an observation ζ drawn from an unknown class, as

Decide Class 1 iff $D_0(z) \geq 0$;
Decide Class 2 iff $D_0(z) < 0$.

This procedure can be seen to be equivalent to the Bayes discriminant function.[7]

Let $P_d(e)$ be the probability of misclassification when using discriminant $d(\bullet)$. It follows from Bayes theory that $D_0(\bullet)$ is the optimal discriminant function $d^{OPT}(\bullet)$, and hence $P^{OPT}(e) = P_{D_0}(e)$ is the minimum probability of misclassification that one can expect. Furthermore, let $\{D_n(x)\}$ be a sequence of estimates of $D_0(x)$. Under appropriate conditions[8] we have $P_{D_n}(e) \rightarrow P_{D_0}(e)$ almost surely. That is, convergence of the density estimates to the true (though unknown) densities implies convergence of the discriminant procedure to the optimal in the minimum probability of misclassification sense. Thus, such a sequence of density estimates $\{D_n(x)\}$ is desirable for applications involving discriminant analysis.

The most enviable situation is that in which the true form of the class densities D_i is known. For example, in the Linear Case each Di is normally distributed with independent mean μ_i and common variance σ^2: $D_i = \Phi(\mu_i, \sigma)$, where $\Phi(x)$ is the well-known normal probability

density function. For the Quadratic Classifier, each class density has independent variance as well: $D_i = \Phi(\mu_j, \sigma_j)$. In this case it is well known that standard parametric estimation techniques will converge. However, it is often the case that simple parametric models cannot be assumed. It then becomes necessary to employ either complex parametric models (such as finite mixtures) or nonparametric estimators (such as the kernel estimator).

For finite mixture models[9] we assume each class density D_i is a sum of basic terms. For the normal mixture case, where each term in the mixture is parameterized by a proportion π, a mean μ, and a standard deviation σ, we have

$$\widehat{D}_i(x) = \sum_{j=1}^{m} \pi_j \Phi(x; \mu_j, \sigma_j). \quad (3)$$

Estimation via finite mixtures can often be performed recursively (as opposed to iteratively) by maximum likelihood techniques. In this case, the estimate based on n teaching observations is simply an update of the previous estimate (based on n –1 observations) and the newest observation ζ_n.

The added model complexity of a mixture of normals yields more power than in the case of a simple normal assumption (linear or quadratic classification), but still requires a priori knowledge (or a guess) as to the true character of the class densities. In this case, the class of functions that can be correctly modeled is the class of all m (or fewer) normal mixtures.

Kernel estimation[10] is a powerful nonparametric density estimation technique. We let our density estimate, based on $\{\zeta_1, \ldots, \zeta_n\}$, be

$$\widehat{D}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - \zeta_i}{h}\right) \quad (4)$$

where h is the window width, which depends on n, and K(•) is the kernel function, often taken to be the normal distribution function. Kernel estimators have quite favorable asymptotic properties under very general conditions on the true density D(x). These properties then translate into asymptotic discriminant properties as described above. Unfortunately, as the formulation (4) suggests, kernel estimation requires the storage and processing of each teaching observation. Thus, while high-quality density estimates can be expected when n is large, the computational and storage requirements quickly become unmanageable, rendering kernel estimation inapplicable for many high data rate problems of interest. This drawback must be weighed against the nonparametric property of the kernel estimator, which allows estimation of a rich class of densities.

The reduced kernel, or Parzen, estimator[11] is a method designed to address this difficulty associated with kernel estimation. Here, we consider a sub-kernel estimator based on some subset (of size N << n, say) of the original teaching set. Hence

$$\widehat{D}_{n,N}(x) = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{h} K\left(\frac{x - \zeta_j}{h}\right). \quad (5)$$

The estimate quality is reduced by fixing the number of kernels at N, but the process is computationally feasible. Unfortunately, the process of choosing the appropriate N observations to retain is iterative and quite computationally intensive. The choice of N also reduces the technique to a parametric method, with its inherent limitation on discriminant complexity.

The complexity of our chosen model directly impacts our performance potential. The probability of misclassification obtainable via a discrimination procedure $P_d(e)$ is directly linked to the ability to approximate the true discriminant $d^{OPT}(\bullet)$. The complexity of the model used in estimating the density translates directly into the discriminant complexity that one may attain. In particular, the parametric mixture model approach (3), with m fixed and small, can yield only a relatively small class of discriminants. (Linear and Quadratic discriminant functions fall in this category.) The same is true with the reduced Parzen estimator (5). On the other hand, the kernel estimator (4) can yield arbitrarily complex discriminants as the number of terms in the model increases. The computational costs for this arbitrary complexity can be quite high, however, and will be addressed below.

Approach

There are several different methods used to compute fractal dimension. The method used in this paper, the covering method, first appeared in the literature in the paper of Peleg.[12] Our implementation makes use of the modified version of the covering method developed by Peli.[13] The covering method views the image as a two manifold embedded in R^3 in order to estimate the surface area of the image in a window surrounding a pixel.

To each pixel (i,j) we associate an integer gray scale value g(i,j). We wish to define a sequence of upper and lower surfaces, U(i,j,ε) and L(i,j,ε) respectively, which bound the original image in grey scale intensity. For a scale value of 0 define

$$U(i,j,0) = L(i,j,0) = g(i,j) \quad (6)$$

For other scale values we recursively define U and L as follows:

$$U(i,j,\varepsilon + 1) = \max\{U(i,j,\varepsilon) + 1, \max_{k,m \in \eta} U(k,m,\varepsilon)\} \quad (7)$$

$$L(i,j,\varepsilon + 1) = \min\{L(i,j,\varepsilon) - 1, \max_{k,m \in \eta} U(k,m,\varepsilon)\} \quad (8)$$

In equations (7) and (8) we define η as {(k,m) | Euclidean distance((i,j),(k,m)) ≤ 1}.

Using a window R centered at the pixel (i,j) we define an area estimate at that pixel as follows:

$$A(i,j,\varepsilon) = (2\varepsilon)^{-1} \sum_{(k,l)\in R} [U(k,l,\varepsilon) - L(k,l,\varepsilon)] \qquad (9)$$

Once this area estimate is obtained for several values of ε we can use them along with Richardson's law to compute D. D is computed as the slope obtained via a regression of log [A(ε)] vs ε as illustrated in Figure 3.

Given A as a function of ε there are several features that can be extracted for use in the pixel discrimination process. In particular, if we consider

$$\log A(\varepsilon) = \beta_0 + \beta_1 \log \varepsilon$$

then $\beta_0 \sim \log(K)$ and $\beta_1 \sim d - D$ are obvious regression-derived features. Furthermore, the F-statistic (or the logarithm thereof) for the significance of the regression is a third feature which intuitively lends itself to discrimination in our problem. The use of each of these three features will be motivated independently. The slope is directly related to the previously mentioned texture feature of fractal dimension. Since there is a simple shift constant relating the two quantities, one should be as good a feature as the other. The y-intercept is another naturally occuring property of the regression line. In some sense it represents a scale factor on the rate of area change with respect to ε. The F statistic represents the significance of our regression.[14] The larger the value of the F statistic the greater the degree of significance that can be attributed to the regression. Since a purely fractal object would exactly fit the regression line,

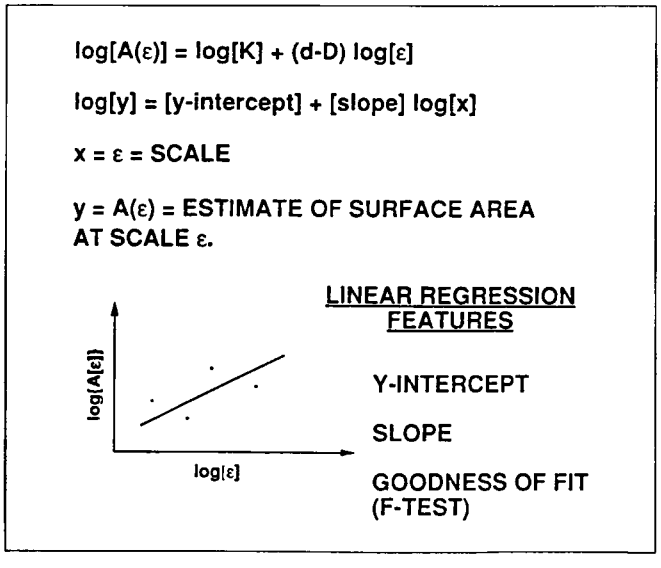


Figure 3. Power law features extracted from Richardson's Law.

it is expected that pixels in our gray-scale image that lie in natural terrain would in general have larger values of the F-test statistic.

It should be noted at this point that the use of power law signatures for pixel classification has appeared previously in the literature. Stein developed a discrimination scheme employing the slope and standard error of fit of the regression line.[5] Besides the use of different features our approach differs from his in two respects. The first is that the set of man made pixels in Steins work represented a small percentage of the overall image, whereas they occupy a large area of our image. Secondly, the decision rules used by Stein were basically heuristic in nature. We propose advanced density estimation techniques in an effort to fully characterize the decision surfaces which originate from our power law signatures.

The adaptive mixtures density estimation procedure[15] used in the simulations below is a cross between the kernel estimator and the mixture model. The recursive nonparametric character of the estimator makes it appropriate for computationally intensive, high data-rate problems such as the discrimination application considered herein. We model the density as

$$\hat{D}(x) = \sum_{j=1}^{m} \pi_j \Phi(x; \mu_j, \sigma_j). \qquad (10)$$

Here m is not fixed as in the mixture model, but can be increased based on the input observations. Thus the parameter list is not fixed (or even bounded!), making (10) a nonparametric estimation technique. The scheme employs the stochastic approximation procedure

$$\hat{D}_{n+1} = \hat{D}_n + [1 - P_n(\zeta_{n+1}; \hat{D}_n)]\, U_n(\zeta_{n+1}; \hat{D}_n) + P_n(\zeta_{n+1}; \hat{D}_n)\, C_n(\zeta_{n+1}; \hat{D}_n, n)$$

which is used to recursively develop density estimates for discriminant analysis. P(•) represents a (possibly stochastic) create decision (a decision to add a new term to the mixture) and takes on values 0 or 1. U(•) updates the current parameters using maximum likelihood techniques, as in mixture modelling, while C(•) adds a new component to the model, incrementing m, analogous to a kernel estimation approach.

Thus the method of adaptive mixtures has the capability to model a rich class of probability density functions, while at the same time having computational efficiencies inherited from its mixture model origins. It is this method that is used to develop the densities, and hence discriminants, used in the simulations below.

It is argued that the ability to model a richer class of densities, inherent in the adaptive mixture procedure, provides more powerful pattern recognition capabilities than simple parametric approaches, as well as insight

into the feature-space structure of the application. It should be noted that the development of complex discriminants is a major focus of neural network research[16]. The relationship between the techniques described above and neural networks is easily seen. In particular, kernel estimators[17] and finite mixture models[18] have been discussed from the perspective of artificial neural implementation, as has adaptive mixtures[19].

Results

Our simulations are based on the gray-scale image shown in Figure 1. We are concerned with the problem of discriminating the three classes found in the image: grass, bush, and tank. In addition, we analyze the performance based on a two-class discrimination problem: natural (grass & bush) vs. man-made (tank).

The first step in the experiment is to choose n scale factors $\varepsilon_1, \ldots, \varepsilon_n$ around each pixel of an image in which we estimate the image area. In these experiments, n = 8. Step two is to determine the measures A(ε), using the covering method, as described a bove, and to perform a least squares regression estimate of the line fitting the 8 measures. The three features described above thus determine a point in 3-space for each pixel.

The training set must now be determined. Here, 50 points from each class (grass, bush, and tank) are chosen at random to train the classifier, and density estimates are produced based on this training sample. It is important to note here that the training set comprises less then one tenth of one percent of the total number of points in the image.

For simplicity, we present first a univarite example of the adaptive mix ture estimator densities in Figure 4. This figure uses the log of the F-statistic as its feature. We see that the class densities break out as expected. In particular, the "bush" density falls between the "grass" and "tank" densities, and the 'tank' density lies to the left of the 'bush' and 'grass' densities. We see also that some level of correct discrimination could be expected based on this single feature.

It is expected, based on the Power-Law theory, that adding the additional information carried by the y-intercept (β_1) feature will improve performance. In Figure 5, we present the density plots for the three classes in the bivariate feature space of y-intercept vs. log(F-statistic). We see that, for this sample at least, the structure of the estimates are quite non-normal. To compare the performance of this estimator vs. the quadratic classifier described above we consider Figures 6 and 7. Figure 7 shows the closer adherence of the adaptive mixtures approach to the structure of the data, and we deduce that this nonparametric approach is more powerful than using an estimator with strict parametric assumptions. This argument is predicated

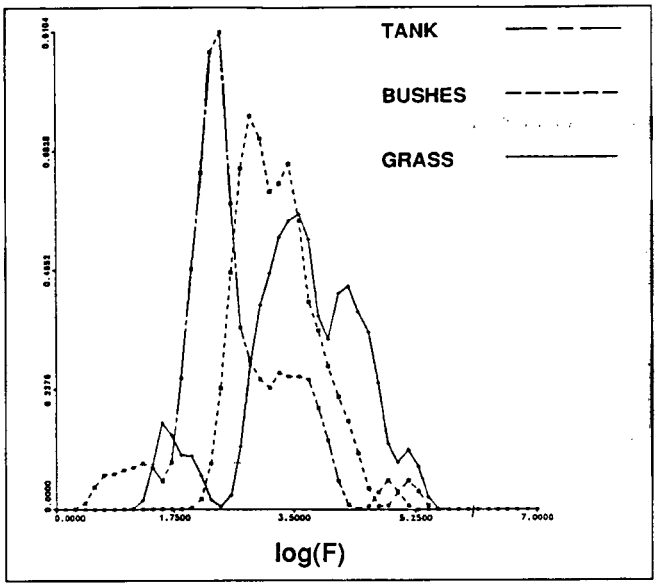


Figure 4. Univariate probability density function for the significance of the regression feature.

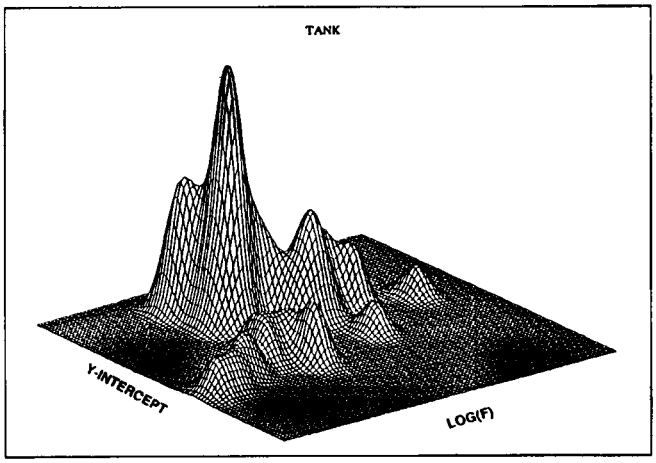


Figure 5a. Bivariate probability density function for the tank class extracted from power law features.

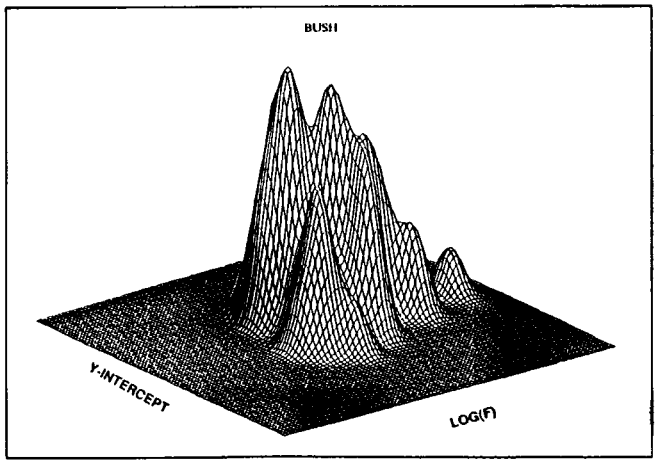


Figure 5b. Bivariate probability density function for the bush class extracted from power law features.

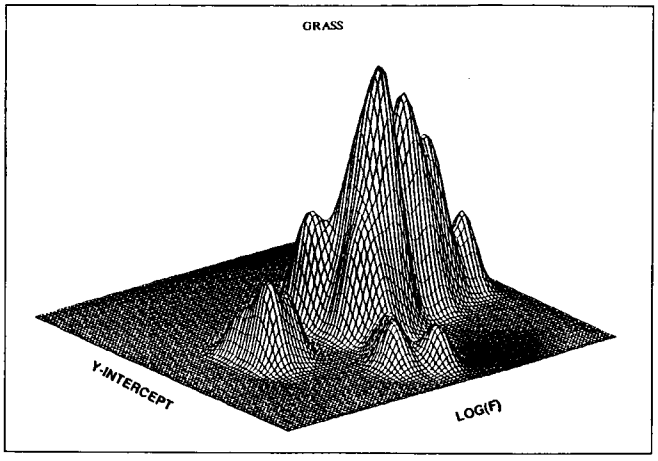


Figure 5c. Bivariate probability density function for the grass class extracted from power law features.

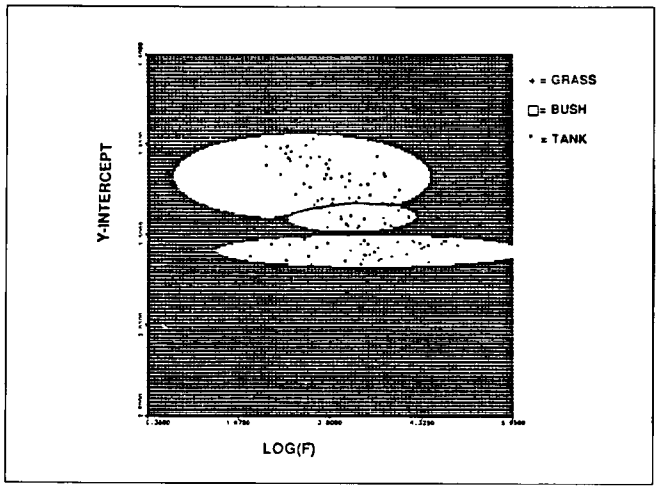


Figure 6. Projected quadratic discriminant surface.

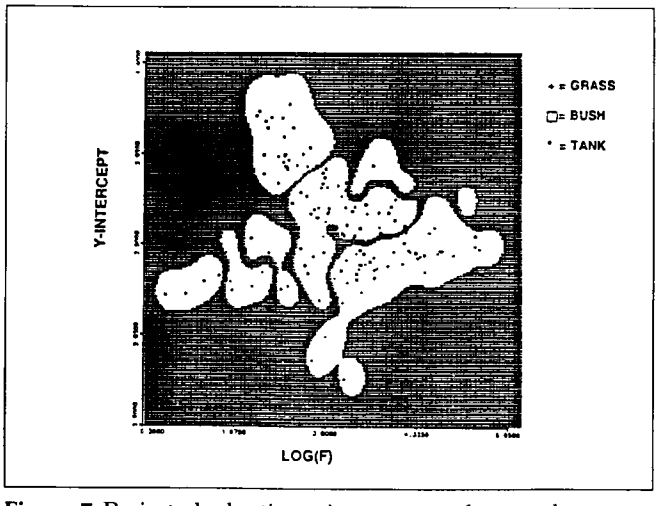


Figure 7. Projected adaptive mixtures neural network discriminant surface.

on the fact that the discriminant surfaces produced via adaptive mixtures more closely model the training data than quadratic discriminants can, while seemingly maintaining generalization capabilities. Detailed simulations are necessary to quantify these results in terms of probability of misclassification.

Some preliminary results as to the performance of the three classifiers on a data set consisting of 1000 points from each of the three classes are given in Table 1. The results were obtained by training and testing the linear and quadratic classifiers on the full 1000 observation sets. The P(C) values represent the probability of correctly classifying a tank pixel. P(FA) is the probability of incorrectly classifying a bush or grass pixel as tank. The results for the adaptive mixtures classfier were obtained by training on 500 observations and testing on the remaining 500. This process was then repeated with the roles of the observations switched. As can be seen from the table the adaptive mixtures classifier improved the probability of correct classification by about eight percent while at the same time lowering the probability of false alarm by about the same amount. A classification threshold of T = 0.5 was used for the tabular results where pixel was classified as tank if M > T where

$$M = P(x|\text{ tank})/\{P(x|\text{ tank}) + \max\{P(x|\text{ bush}), P(x|\text{ grass})\}\}.$$

Table 1. Comaprison of classifier performance for threshold T = 0.5

	P(C)	P(FA)
LC	0.6690	0.1255
QC	0.6680	0.1170
AM	0.7360	0.0840

These results are obtained with an adaptive mixtures model of approximately 20 terms per class. This is a computationally reasonable number of terms compared to a kernel estimator approach, and justifies the consideration of this approach for high data-rate applications.

Figures 8, 9, and 10 show the results of processing the entire image (Figure 1) using the discriminants produced via the estimation approach (Figure 7), with the highlighted pixels being those classified as man-made. We see that, as one would expect based on the univariate densities shown in Figure 4, there are few false alarms in the grassy regions, more false alarms in the bushes (the densities for "bush" and "tank" are less distinct) and a high percentage of correctly classified pixels in the tank itself. As can be seen from the figures, as T is made larger both the false alarm rate and the probability of correct classification is lowered. Depending on the scenario this low false alarm rate may be

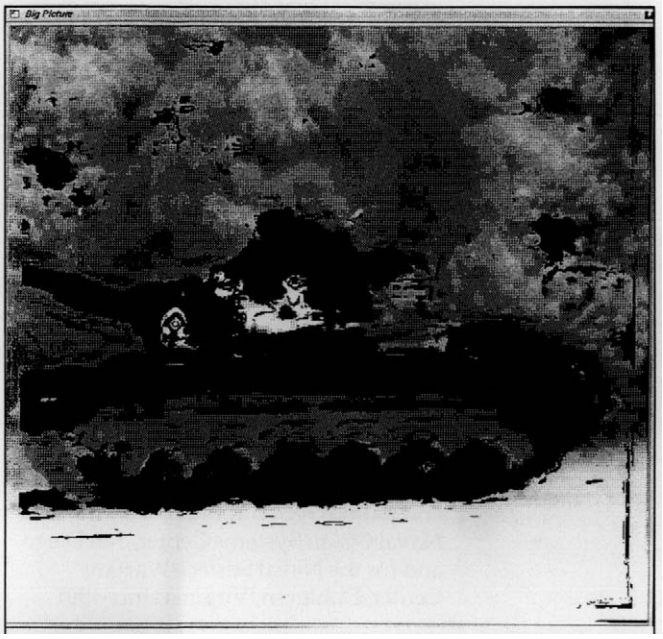


Figure 8. Adaptive mixtures classification results where the pixels classified as manmade are highlighted (T = .58).

Figure 9. Adaptive mixtures classification results where the pixels classified as manmade are highlighted (T = .88).

appropriate so as to not overload a group of human photo-analysis. There is however still a sufficient number of correctly classified manmade pixels to indicate further inspection of this region by the analyst.

Conclusions

The conclusions to be drawn from this experiment are twofold. First, power-law based features can be useful

Figure 10. Adaptive mixtures classification results where the pixels classified as manmade are highlighted (T = .98).

in performing object discrimination. While the probability density functions presented above do have nontrivial overlap, they are sufficiently distinct to indicate the potential for a low probability of misclassification. Second, there is a need to investigate complex discriminant structures for the features obtained in such a process. The densities presented require relaxed parametric assumptions, and the discriminant surfaces derived from these densities may be significantly better than the corresponding linear and quadratic discriminants. Future work will include a quantitative analysis of the performance of the adaptive mixtures estimator on the power law signatures of testing sets obtained from the original tank image and other field collected images.

Acknowledgements

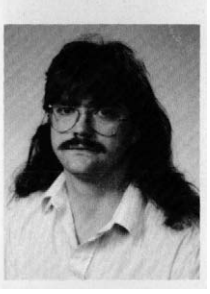
We would like to thank the Advanced Mission Planning Technology Program of the Dahlgren Division of the Naval Surface Warfare Center without whose support this work would not have been possible.

References

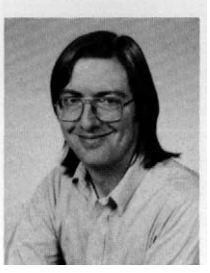
1. Peitgen, H., Saupe, D., Eds., *The Science of Fractal Images*, New York, (1988), Springer-Verlag.
2. Barnsley, M. F., *Fractals Everywhere*, Boston, (1988), Academic Press.
3. Prusinkiewicz, P., and Hanan, J., *Lindenmayer Systems, Fractals, and Plants*, New York, (1980), Springer-Verlag.

4. Mandelbrot, B., B., *The Fractal Geometry of Nature*, New York, (1983), W. H. Freeman.
5. Stein, M. C., "Fractal Image Models and Object Detection", *Visual Communication and Image Processing II*, SPIE, Vol. 845, (1987).
6. Pentland, A. P., "Fractal-based Description of Natural Scenes", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6), (1984)
7. Duda, R.O. and Hart, P.E., *Pattern Classificaion and Scene Analysis*, Wiley, 1973.
8. Wolverton, C.T. and Wagner, T.J., Asymptotically Optimal Discriminant Functions for Pattern Classification", *IEEE Trans. Info. Theory.*, Vol. IT-15, No. 2, pp 258-265, 1969.
9. Titterington, D.M., Smith, A.F.M., and Makov, U.E., *Statistical Analysis of Finite Mixture Distributions*, Wiley, 1985.
10. Silverman, B.W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, 1986.
11. Fukunaga, K. and Hayes, R.R., "The Reduced Parzen Classifier", *IEEE Trans. PAMI*, Vol II, No. 4, pp 423-425, 1989.
12. Peleg, S., Naor, J., Hartley, R., and Avnir, D., "Multiple Resolution Texture Analysis and Classification", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-6, No. 4, (July 1984).
13. Peli, T., "Multiscale Fractal Theory and Object Characterization", *Journal Optical Society of America A.*, Vol. 7, No. 6, (June 1990).
14. Montgomery, D.C., and Peck, E.A., *Introduction to Linear Regression Analysis*, New York, (1982), John Wiley and Sons.
15. Priebe, C.E. and Marchette, D.J., "Adaptive mixtures: recursive nonparametric pattern recognition", *Pattern Recognition*, Vol. 24, No. 12, pp. 1197-1209, 1991.
16. P.A. Shoemaker, M.J. Carlin, R.L. Shimabukuro, and C.E. Priebe, "Least-squares learning and approximation of posterior probabilities on classification problems by neural networks", *Proceedings Second Workshop on Neural Networks*, SPIE Vol 1515, 1991.
17. Specht, D.F., "Probabilistic Neural Networks", *Neural Networks*, Vol. 3, 1990.
18. Perlovsky, L. I. and McManus, M. M., "Maximum likelihood neural networks for sensor fusion and adaptive classification," *Neural Networks*, 1991.
19. Marchette, D. J. and Priebe, C. E., "The adaptive kernel neural network," *Mathl. & Comput. Modelling*, Vol. 14, pp. 328-333, 1990.

JEFFREY L. SOLKA received his B.S. degree in Mathematics and Chemistry from James Madison University in 1978, his M.S. degree in Mathematics from James Madison University in 1981, and his M.S. degree in Physics from Virginia Polytechnic Institute and State University in 1989. Since 1984, Mr. Solka has been working in the areas of strategic defense and artificial neural systems for the Dahlgren Division of the Naval Surface Warfare Center.



CAREY E. PRIEBE received his B.S. degree in Mathematics from Purdue University in 1984 and his M.S. degree in Computer Science from San Diego State University in 1988. Since 1985, Mr. Priebe has been working in adaptive systems and pattern recognition, first for the Naval Ocean Systems Center, San Diego, and for the Naval Surface Warfare Center, Dahlgren, Virginia since 1991.



GEORGE W. ROGERS received his B.S. degree from Georgia Southern College in 1977 and the Ph.D. degree in Theoretical Physics from the University of South Carolina in 1984. Since 1985 he has been employed at the Dahlgren Division of the Naval Surface Warfare Center where he first worked in orbit computation and more recently in the field of artificial neural networks. His current research interests are in composite neuronal dynamics and adaptive pattern recognition.