Open access • Journal Article • DOI:10.1109/TPAMI.1984.4767560

Estimation of Error Rates in Classification of Distorted Imagery — Source link

M. J. Lahart

Institutions: United States Naval Research Laboratory

Published on: 01 Apr 1984 - IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE)

Topics: Contextual image classification, Feature extraction, Feature vector, Fourier transform and Gaussian

A. Analysis of the Expression of $\hat{\mu}_1$

As shown above, $\hat{\mu}_1$ is the sum of three terms and is of the form

$$\hat{\mu}_1 = \alpha A_1 + (1 - \alpha)\mu_{01} + \beta(A_2 - \mu_{02})$$

where

$$\alpha = \frac{n_1\sigma_{01}^2(\sigma_2^2 + n_2\sigma_{02}^2(1 - \rho^2))}{(\sigma_1^2 + n_1\sigma_{01}^2)(\sigma_2^2 + n_2\sigma_{02}^2) - n_1 n_2 \rho^2 \sigma_{01}^2 \sigma_{02}^2}$$

$$\beta = \frac{n_2\rho\sigma_1^2\sigma_{01}\sigma_{02}}{(\sigma_1^2 + n_1\sigma_{01}^2)(\sigma_2^2 + n_2\sigma_{02}^2) - n_1 n_2 \rho^2 \sigma_{01}^2 \sigma_{02}^2}.$$

For $n_1 = 0$, $\alpha = 0$ and the first term of $\hat{\mu}_1$ (contribution of the observations from ω_1) is equal to 0 as could be expected. When n_1 becomes very large, α tends to 1 and β tends to 0, so that the contribution of the second and third terms of $\hat{\mu}_1$ becomes negligible. Finally, if $\rho = 0$, β is equal to 0 and there is no contribution of the observations from Class ω_2; this could also be predicted since in that case the random variables μ_1 and μ_2 are independent. β increases with $|\rho|$, so that the contribution of the observations from Class ω_2 increases with the amount of correlation of μ_1 and μ_2.

B. Analysis of the Mean-Squared Error

A simple way to evaluate the performance of the EMAP estimation procedure is by analyzing the expression of the mean-square error between the random variable μ_1 and its estimate $\hat{\mu}_1$.

If no observations have been obtained from Class ω_1, (i.e., $n_1 = 0$), we note from (15) that the mean-squared error of $\hat{\mu}_1$ is equal to

$$r_1^2 = E(\hat{\mu}_1 - \mu_1)^2 = \frac{\sigma_{01}^2(\sigma_2^2 + n_2\sigma_{02}^2(1 - \rho^2))}{\sigma_2^2 + n_2\sigma_{02}^2}. \tag{17}$$

From (17), we see that r_1^2 is always smaller than σ_{01}^2, the *a priori* variance of ω_1 and is a decreasing function n_2. As n_2 becomes large, r_1^2 approaches its lower bound of $\sigma_{01}^2(1 - \rho^2)$. As could be expected, this lower bound decreases when the cross-covariance between μ_1 and μ_2 increases.

Finally, let us consider the case where $n_1 = n_2 = n$, $\sigma_1^2 = \sigma_2^2 = \sigma^2$, and $\sigma_{01}^2 = \sigma_{02}^2 = \sigma_0^2$. In that case

$$r_1^2 = \sigma_0^2 \frac{\sigma^2(\sigma^2 + n\sigma_0^2(1 - \rho^2))}{(\sigma^2 + n\sigma_0^2)^2 - n^2\rho^2\sigma_0^4}. \tag{18}$$

Fig. 5 shows the evolution of r_1^2 as a function of n for different values of ρ when $\sigma^2 = \sigma_0^2 = 1$.

We can see that the expected mean-square error asymptotes to 0 in all curves as n becomes large. We can also see that the most significant improvement of the mean square error for small n is obtained in the case of highly correlated mean values. We note from (16) that for a given n, r_1^2 is an increasing function of σ^2/σ_0^2. Therefore, the most advantageous conditions for the use of the extended MAP estimate occur when 1) the correlation of the mean values across classes is high, 2) the ratio of the variance of the data to the variance of the mean, σ/σ_0 is low, and 3) the number of observations is small.

V. Summary and Conclusions

In this paper we reviewed the classical MAP estimation procedure for updating the probability density functions of Gaussian random mean vectors from a set of labeled observations. We extended the procedure so that it could take into consideration not only the feature-to-feature correlations within a decision class but also the correlations of the features' means from one class to another. We formally evaluated this procedure for a simple two-class and one-feature case. We showed that the mean-squared error of estimates of the mean vectors is always smaller when the class-to-class correlations are taken into account, and that the greatest improvement afforded by the exploitation of class-to-class correlations is obtained when the number of observed samples is small, the class-to-class correlations of the means are high, and the ratio of the variance of the data to the variance of the mean is large.

The formulation of the estimation procedure was simplified by the use of a set of notational conventions that capture the covariances of the feature mean values within a given class as well as the cross-covariances between the mean vectors of different classes.

Acknowledgment

The authors wish to thank Dr. R. A. Cole for the initial motivation for this research, as well as general help and advice on feature-based speech recognition, Dr. B. V. K. V. Kumar for useful discussions, and Dr. D. R. Reddy for providing a stimulating working environment and computing resources.

References

[1] R. A. Cole, R. M. Stern, M. S. Philips, S. M. Brill, A. P. Pilant, and P. Specker, "Feature-based speaker-independent recognition of isolated English letters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Boston, MA, 1983, pp. 731-733.

[2] H. Cramer, *Mathematical Methods of Statistics*. Princeton, NJ: Princeton Univ. Press, 1946.

[3] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[4] D. G. Keehn, "A note on learning for Gaussian processes," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 126-132, 1965.

[5] R. M. Stern and M. J. Lasry, "Dynamic speaker adaptation for isolated letter recognition using MAP estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Boston, MA, 1983, pp. 734-737.

[6] H. L. Van Trees, *Dectection, Estimation, and Modulation Theory, Part I*. New York: Wiley, 1973, pp. 60-85.

Estimation of Error Rates in Classification of Distorted Imagery

M. J. LAHART

Abstract—**This correspondence considers the problem of matching image data to a large library of objects when the image is distorted. Two types of distortions are considered: blur-type, in which a transfer function is applied to Fourier components of the image, and scale-type, in which each Fourier component is mapped into another. The objects of the library are assumed to be normally distributed in an appropriate feature space. Approximate expressions are developed for classification error rates as a function of noise. The error rates they predict are compared with those from classification of artificial data, generated by a Gaussian random number generator, and with error rates from classification of actual data. It is demonstrated that, for classification purposes, distortions can be characterized by a small number of parameters.**

Index Terms—**Image classification, image matching, feature extraction, pattern classification, pattern recognition.**

Manuscript received May 26, 1983; revised August 22, 1983.

The author was with the U.S. Naval Research Laboratory, Washington, DC 20375. He is now with the National Institutes of Health, Bethesda, MD 20209.

I. Introduction

The ability to classify data is determined not only by the features that describe it, but by the number of categories into which the data are to be classified and by noise and distortion that influence it. The measure by which classifiability is judged is error rate, which is a function of all of these parameters. Evaluation of a feature set must include computation of the factors that influence error rates and, ultimately, the estimation of the error rates themselves.

In this paper we compute Bayes error rates that arise when each member of a library of classes is equally likely and the object to be classified is the sum of a library feature vector and Gaussian noise. Data, represented by the measured feature vector $\vec{x}$, are classified to the class ω by maximizing the *a posteriori* probability $p(\omega|\vec{x})$. When the *a priori* probability $P(\omega)$ is the same for all classes, this can be shown by application of the Bayes rule to be equivalent to maximizing $p(\vec{x}|\omega)$. This function is Gaussian, and it is a maximum when its exponent d^2, given by

$$d^2 = \tfrac{1}{2}(\vec{x} - \vec{x}_0)\vec{N}^{-1}(\vec{x} - \vec{x}_0) \tag{1}$$

is a minimum. Here, $\vec{x}_0$ is a library feature vector representing a class ω, $\vec{x}$ is a measured feature vector to be classified, and $\vec{N}$ is the noise covariance matrix, assumed to be the same for each class.

Often the data are defined by the addition of noise to a distorted version of a library feature vector. This occurs, for example, when the library consists of measurements of images and the data $\vec{x}$ are derived from magnified, blurred, or otherwise distorted versions of these same images. Typically, the *a posteriori* probability of the occurrence $p(\vec{x}|\omega)$ is not known or is not tractable. Application of the simple distance measure defined in (1) results in classification error rates that are larger than the Bayes rate. This distance measure can still be used, however, if the library is enlarged to include distorted versions $\vec{x}_0'$ of each object, if simple assumptions can be made concerning the probability $p(\omega|\vec{x}_0')$ of the occurrence of class ω given the distorted data $\vec{x}_0'$. In the nearest neighbor method of classification [1], [2] a measurement $\vec{x}$ is assigned to the same class as its nearest library member $\vec{x}_0'$ is assigned under application of the Bayes decision rule. This requires only that the class ω is known for each x_0' for which $p(\omega|\vec{x}_0')$ is largest.

Bounds on error rates for the nearest neighbor decision rule have been computed as a function of the Bayes error rate under the assumption that $p(\omega|\vec{x}_0') \approx p(\omega|\vec{x})$ holds, where $\vec{x}$ is the measured feature vector and $\vec{x}_0'$ is the library feature vector nearest it. The assumption requires that the library have enough distorted members so that the distance (1) between a measurement and its nearest neighbor is negligible. The distance requirement has been relaxed somewhat to a small distance by Short and Fukunaga [3], who defined a local metric that is a function of the gradient of $p(\omega|\vec{x}_0')$.

Successful classifier design often requires knowledge of the amount of distortion that can be tolerated. To provide this, we will compute, as a function of noise variance, error rates that arise when the distance measure of (1) is used to classify against a library of distorted data. We compute them separately for different levels of distortion, noting that the error that is expected in a classifier would be an average of these errors, each weighted by the probability of occurrence of its level of distortion. Even without exact knowledge of this probability, the computed error rates can indicate how far apart must be the levels of distortion represented in a library of data. The distortion problem may also be approached by attempting to remove the effects of distortion from a measured feature vector before comparing it to a library of undistorted data. In this case the error curves will indicate how accurate the restoration must be.

We will consider error rates that arise when the number of classes is large. Although multiclass classification problems are encountered often in practice, problems that are peculiar to it have not been investigated extensively. Some time ago Johns [4] derived an upper bound on multiclass error rates as a sum of pairwise error rates. More recently, Fukunaga and Flick [5] made a number of Monte Carlo simulations in which error rates were computed as a function of noise level, average distance between members of the data library, and other factors.

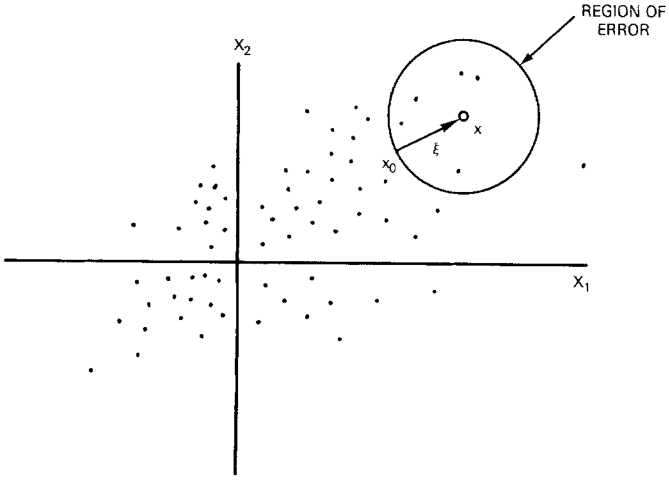

Fig. 1. Relationship between library member x_0 and measurement x in feature space with Gaussian distribution of data.

We will assume that the features that characterize the data library are normally distributed. In Section II, we will derive an upper bound on error rates for this data as a function of noise and number of classes when the data is undistorted. In Section III this analysis will be applied to classification by correlation detection, and the influence of distortions on error rates will be discussed. In Section IV, the analysis of Section II will be extended to distorted data and an approximation expression for the error rate in the case of zero noise will be derived. In Section V the classification error rates for both actual and contrived data will be computed and compared with the analytical results of the preceding sections.

II. Multiclass Error

We consider a library of M feature sets, each of which is derived from a noiseless prototype of one of M classes. We assume that these are distributed normally in a feature space. A set of classes is shown schematically in Fig. 1, in which each class is represented by a point. A measurement datum is equivalent to a library vector $\vec{x}$ to which noise $\vec{\xi}$ has been added and is indicated in Fig. 1 by the small circle. The measurement is classified by computing the distance $\boldsymbol{d}$ to each member of the library and assigning the measurement to the class for which the distance is the smallest. The probability that it will be misclassified is the probability that at least one other library prototype is located in the hypersphere of radius $|\vec{\xi}|$ whose center is located at $\vec{x} + \vec{\xi}$; it is a function of both $\vec{x}$ and $\vec{\xi}$. We will compute the average value of this probability, the error rate, under the assumption that the library has $\boldsymbol{n}$ features, each characterized by the same standard deviation σ and that the noise that is added to each feature is described by standard deviation σ_N. The probability $f(\vec{\xi}, \vec{x})$ that a given class will be mistaken for the correct one at $\vec{x}$ is the integral of its distribution function over the hypersphere:

$$f(\vec{\xi}, \vec{x}) = \frac{1}{(2\pi)^{n/2}\sigma^n} \int_0^{|\xi|} \exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i + \xi_i + s_i)^2\right] d\vec{s}. \tag{2}$$

In the two class problem, the probability that the wrong class is not closer to the measurement is $[1 - f(\vec{\xi}, \vec{x})]$. If there are M classes, the probability that none of the $M - 1$ other classes is closer to $\vec{x}$ is this quantity raised to the $M - 1$ power. The classification error at point $\vec{x}$ for noise $\vec{\xi}$ is then

$$p_e(\vec{\xi}, \vec{x}) = 1 - [1 - f(\vec{\xi}, \vec{x})]^{M-1}. \tag{3}$$

The average error probability E is the integral of $p_e(\xi, x)$ over all possible noise values $\vec{\xi}$ and all classes in the feature space:

$$E = 1 - \frac{1}{(2\pi)^n (\sigma_N \sigma)^n} \int \exp\left[-\frac{\xi^2}{2\sigma_N^2}\right] \exp\left[-\frac{1}{2\sigma^2}\sum_i x_i^2\right] \times [1 - f(\vec{\xi}, \vec{x})]^{M-1}\, d\vec{x}\, d\vec{\xi}, \tag{4}$$

This integral is difficult to evaluate in a multidimensional feature space, but it can be approximated by performing the integration over $\vec{x}$ before computing $p_e(\vec{\xi}, \vec{x})$. The approximate error rate is

$$E \cong 1 - \frac{1}{(2\pi)^{n/2} \sigma_N^n} \int \exp\left[-\frac{\xi^2}{2\sigma_N^2}\right] [1 - g(\vec{\xi})]^{M-1}\, d\vec{\xi} \tag{5}$$

where $g(\vec{\xi})$ is

$$g(\vec{\xi}) = \frac{1}{(2\pi)^n \sigma^{2n}} \int_0^{|\xi|} \int_{-\infty}^{\infty} \cdot \exp\left\{-\frac{1}{2\sigma^2}\sum_i [x_i^2 + (x_i + \xi_i + s_i)^2]\right\} d\vec{x}\, d\vec{s}. \tag{6}$$

This expression can be simplified. It is evaluated in the Appendix and shown to be

$$g(\vec{\xi}) = C \int_0^{\xi} s^{n-1}\, ds \int_{-\pi/2}^{\pi/2} \cdot \exp\left[-\frac{1}{4\sigma^2}(\xi^2 + s^2 + 2s\xi \sin\theta)\right] \cos^{n-2}\theta\, d\theta \tag{7}$$

where C is a normalization constant that is defined in the Appendix, s and ξ are the lengths of the vectors $\vec{s}$ and $\vec{\xi}$, and θ is the angle that the vector $\vec{s}$ makes with the perpendicular to $\vec{\xi}$.

We can show that the approximation used in (5) always leads to an upper bound on the true error rate. The integral respect to $\vec{x}$ of (4) is less than or equal to the integrand of (5). The integral with respect to $\vec{x}$ of (4) can be written

$$\int p(\vec{x})[1 - f(\vec{\xi}, \vec{x})]^{M-1}\, d\vec{x} = \int [p(\vec{x})^{1/(M-1)}]^{M-1}\, d\vec{x} \cdot \int \{p(\vec{x})^{1/(M-1)}[1 - f(\vec{x}, \vec{\xi})]\}^{M-1}\, d\vec{x} \tag{8}$$

where the Gaussian distribution of classes has been written as $p(\vec{x})$, whose integral over all $\vec{x}$ is unity. Use of Hölder's inequality leads to

$$\int p(\vec{x})[1 - f(\vec{x}, \vec{\xi})]^{M-1}\, d\vec{x} \geqslant \left\{\int p(\vec{x})^{2/(M-1)}[1 - f(\vec{x}, \vec{\xi})]\, d\vec{x}\right\}^{M-1}. \tag{9}$$

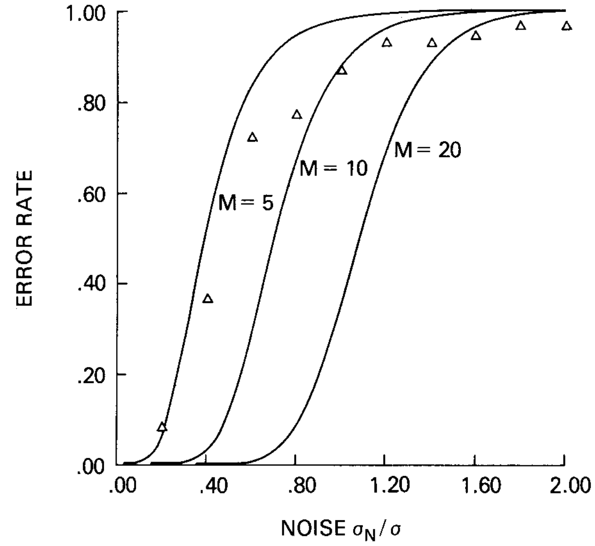

Fig. 2. Approximate error rates as a function of noise for 255 classes with dimensionalities of 5, 10, and 20.

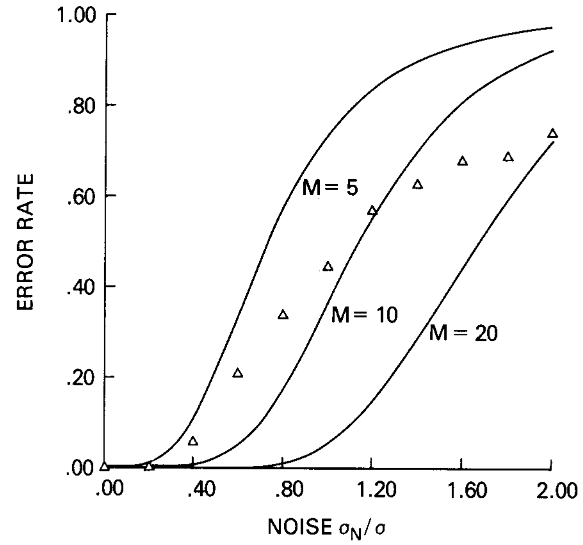

Fig. 3. Approximate error rates as a function of noise for 20 classes with dimensionalities of 5, 10, and 20.

Because $p(x)$ is bounded by zero and one, the inequality

$$\int p(\vec{x})[1 - f(\vec{x}, \vec{\xi})]^{M-1}\, d\vec{x} \geqslant \left\{\int p(\vec{x})[1 - f(\vec{x}, \vec{\xi})]\, dx\right\}^{M-1} \tag{10}$$

holds for $M \geqslant 3$. When M is 2, the equality holds trivially. We note that the equality holds when $f(\vec{x}, \vec{\xi})$ is zero, i.e., when the error rate is zero, and that the approximation is best for small error rates.

If an appropriate change of variables is made in (5) and (6), the error rate E can be shown to be a function of the ratio σ_N/σ, and not of the standard deviations individually. Figs. 2 and 3 show error rates when the number of classes is 255 and 20, respectively, for dimensionalities of 5, 10, and 20. Each figure also contains the results of classification of randomly generated data with a dimensionality of 5. To compute these, the data libraries were generated randomly and data measurements were created by adding Gaussian random noise to each member of the library. (For 20 classes, the results are the average error for classifications of 20 sets of randomly generated data. Errors computed for a single set of data are statistically unstable when the number of classes is low.) The approximation of (5) is demonstrated to be a reasonably tight upper bound which holds best for small error rates.

III. Application–Correlation Matching

One of the most widely used methods of comparing and classifying images is correlation. Typically, a measurement is compared with a prototype image, and identification is accomplished by subtracting or correlating the measurement from the prototype. When images are subtracted, the distance measurement of (1) is the integral of the difference over the area of interest. This is

$$d^2 = \frac{1}{\sigma_N^2} \int [y(\vec{z} - \vec{z}_0) - y_r(\vec{z})]^2 \, d\vec{z} \tag{11}$$

where $y_r(\vec{z})$ is the measurement, $y(\vec{z} - \vec{z}_0)$ is the prototype, and σ_N^2 is the noise variance, which we assume to be independent of position. This integral can be written

$$d^2 = \frac{1}{\sigma_N^2} \int y^2(\vec{z} - \vec{z}_0) \, d\vec{z} + \frac{1}{\sigma_N^2} \int y_r^2(\vec{z}) \, d\vec{z} - \frac{2}{\sigma_N^2} \int y(\vec{z} - \vec{z}_0) \, y_r(\vec{z}) \, d\vec{z}. \tag{12}$$

If the prototype library and the measurements are of generally similar imagery, the first two integrals may be relatively independent of the choice of the library member and the measurement that is being classified. Often these conditions hold, and classification can be accomplished solely by means of the correlation function that constitutes the last integral: the distance d^2 is smallest when the correlation peak is largest.

If the statistics of the imagery are Gaussian, the analysis of the previous section can be used to estimate error rates for classification through application of (12). The axes $\{X_i\}$ of the feature space are eigenvectors of the correlation matrix of the imagery. The measurement $y_r(\vec{z})$ and the library member $y(\vec{z})$ are represented as vectors $\vec{x}$ and $\vec{x}'$ in this feature space, and the quantity d^2, expressed in terms of these is

$$d^2 = \frac{1}{\sigma_N^2} |\vec{x} - \vec{x}'|^2. \tag{13}$$

If the measurement is a truncated version of what may be assumed to be stationary imagery, and if its truncating aperture is large, the eigenfunctions $\{X_i\}$ are approximately Fourier components of $y(\vec{z})$, defined over a region the size of the limiting aperture. This may be demonstrated by computing the Fourier transform of the correlation function $\langle y(\vec{z}) \, y(\vec{z}') \rangle$ of the library images. Because of stationarity, the correlation function is a function only of the difference s between $\vec{z}$ and $\vec{z}'$. Its Fourier transform, for large apertures, is, in one dimension,

$$T(\omega, \omega') = \frac{1}{(2\pi)^2} \iint e^{i(\omega - \omega')z} e^{i\omega' s} \langle y(z) \, y(z - s) \rangle \, dz \, ds \tag{14}$$

or

$$T(\omega, \omega') = \frac{1}{2\pi} \delta(\omega - \omega') \int e^{i\omega' s} \langle y(z) \, y(z - s) \rangle \, ds$$

where $\delta(\cdot)$ is a Dirac delta function. The correlation matrix is diagonal in the Fourier domain, implying that Fourier components are eigenfunctions. The eigenvalues σ_i^2 are (approximately) values of the Wiener spectrum $T(\omega_i)$.

A distorted image may be represented by the product $\vec{A}\vec{x}$, where $\vec{A}$ is a matrix whose coefficients describe the distortion. When $\vec{A}\vec{x}$ is substituted for $\vec{x}$ in (13), the distance may not be zero when $\vec{x}$ and $\vec{x}'$ describe the same class. The error rate with zero noise is calculated by determining the distance $\boldsymbol{r}$ between $\vec{x}'$ and $\vec{A}\vec{x}$ and computing the probability that an incorrect class is within a hypersphere of this radius centered at the measurement point. If noise is present, the radius of the hypersphere may be larger or smaller, depending on the direction of the noise vector $\vec{\xi}$.

The matrix $\vec{A}$ defines distortion generally, including not only geometric distortions, but transformations such as blur or contrast loss which change imagery that is to be classified. Blur is a multiplication in the Fourier domain and may be described by a diagonal matrix $\vec{A}$, whose elements a_{ii} are values of the transfer function $a(\omega_i)$. In the case of contrast loss, all a_{ii} are the same. Off-diagonal elements a_{ij} of $\vec{A}$ are zero for transformations of this type.

Rotations and scale distortions have proved to cause serious errors in correlation matching, and their effects on the correlation peak have been analyzed by several workers [6]–[9]. Under these transformations the spectral component at each frequency ω_i is mapped into a new component, at ω_i/m in the case of magnification $\boldsymbol{m}$. On the average, the correlation function is the Fourier transform of the products of transformed and untransformed spectral components, integrated over the extent of the aperture that limits the reference [6], [7]. The expectation of the correlation peak height is a function of the Wiener spectrum of the imagery and the size of the reference aperture; high spatial frequency components and large apertures make the height of the peak more sensitive to distortions of this kind. Since spectral components are only approximately equivalent to the orthogonal features, each transformed $\vec{x}$ usually has a component along the original. Both this and the component orthogonal to it must be considered in calculating error rates.

The energy of the imagery is conserved under rotations and, if the imagery is limited in extent by a truncating aperture, it is approximately conserved under magnifications. The integral of the Wiener spectrum is the same before and after the transformation in this case. In our notation in which a set of spectral components is denoted by $\vec{x}$ and a set of transformed components by $\vec{A}\vec{x}$, this means

$$|\vec{x}|^2 = |\vec{A}\vec{x}|^2 \tag{15}$$

or

$$\sum_i x_i^2 (1 - a_{ii}^2) = 2 \sum_i \left[a_{ii} x_i {\sum_j}' a_{ij} x_j \right] + \sum_i \left[{\sum_j}' a_{ij} x_j \right]^2$$

where the prime on the summation sign indicates that the term for which $j = i$ holds is omitted from the summation.

Equation (15) can be averaged over the entire data set to give

$$\sum_i \sigma_i^2 (1 - a_{ii}^2) = {\sum_{i,j}}' a_{ij}^2 \sigma_j^2. \tag{16}$$

The right side of this equation is the sum of what all x_i gain under application of the operator $\vec{A}$, while the left side is the sum of what all lose. We will assume that this relationship holds for each x_i individually: that on the average

$$\sigma_i^2 (1 - a_{ii}^2) = {\sum_j}' a_{ij}^2 \sigma_j^2. \tag{17}$$

This relationship will be approximately true if the Wiener spectrum σ_i^2 and the projection quantities a_{ii} vary slowly as a function of the index i, and the off-axis elements of $\vec{A}$ are small when i and j differ greatly.

As described above, the coefficients a_{ii} may be computed analytically in simple cases. If a library of data is available, the a_{ii} may also be computed as the inner product of a covariance eigenvector and a distorted version of itself. This is

$$a_{ii} = \phi_i \vec{A} \phi_i \tag{18}$$

where ϕ_i is a normalized eigenvector of the covariance matrix of the data library.

IV. Distortion Error Rates

As in Section II, we compute error rates by finding the probability that a given class is within a hypersphere centered at the measurement point. The integral of the distribution function is

$$f(\vec{\xi}, \vec{x}) = \frac{1}{(2\pi)^{n/2}\sigma^n} \int_0^r \exp\left[-\sum_i \frac{1}{2\sigma_i^2}(\vec{A}x_i + \xi_i + s_i)^2\right] d\vec{s}. \tag{2'}$$

Here, r is the radius of the hypersphere, i.e., the distance between the correct member of the library and the distorted image with noise $\vec{\xi}$. For scale-type distortions, the exponent may be expressed

$$\sum_i \frac{1}{2\sigma_i^2}\left[a_{ii}x_i + {\sum_j}' a_{ij}x_j + \xi_i + s_i\right]^2. \tag{19}$$

We will not attempt to compute terms involving a_{ij}. Rather we will estimate provisional error rates for given values of the summation $\Sigma' a_{ij}x_j$, assumed constant over the library coordinate $\vec{s}$. The average error rate is approximated as the sum of these provisional error rates, weighted by the probability of occurrence of a given value of $\Sigma' a_{ij}x_j$.

These considerations mean that $\Sigma' a_{ij}x_j$ will be treated as a random quantity that follows an ergodicity assumption. The quantity $\Sigma' a_{ij}x_j$ is a zero mean Gaussian process and, because of the orthogonality of x_i and x_j, would be independent of $a_{ii}x_i$, except for the relationship of (15). Our approximation will assume that it is an independent Gaussian process and has a standard deviation given by right side of (17).

We also make an approximation concerning r. Rigorously, it is a function of x_i given by

$$r^2 = \sum_i [x_i(1 - a_{ii}) + \xi_i]^2, \tag{20}$$

where ξ_i includes both randomly added noise and contributions from off-axis values of the matrix $\vec{A}$. The average value of this, for a given value of ξ is

$$\langle r^2 \rangle = \sum_i \sigma_i^2(1 - a_{ii})^2 + \xi^2. \tag{21}$$

We will use the square root of this average for r in (2′).

With these additional approximations, the computation proceeds as in Section II. As before, we compute error rates under the assumption that the correlation matrix eigenvalues all have the same value σ. We will also assume that the projection coefficients all have the same value a. Because σ_i and a_{ii} are all equal in the computations of this section, the standard deviation of $\Sigma' a_{ij}x_j$ must be also. We denote it σ_M. We are now considering a noise component that consists of a sum of the original $\vec{\xi}$ and the noise-like terms that include a_{ij}. Since these quantities are independent, their combined distribution is a convolution of their individual distributions. It is a Gaussian distribution with variance $(\sigma_N^2 + \sigma_M^2)$.

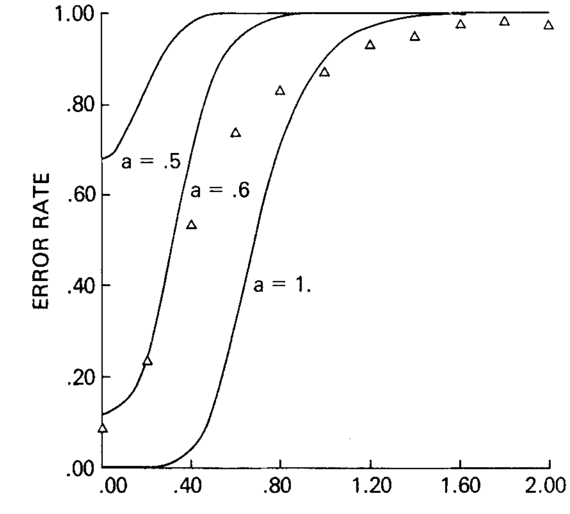

Fig. 4. Blur-type error rate as a function of noise.

The probability of error $p_e(\xi, \vec{x})$ is computed from $f(\xi, \vec{x})$ as in (3). An approximate error rate is computed by integrating over $\vec{x}$ before computing p_e. The result, for scale-type distortions, is

$$E \simeq 1 - \frac{1}{(2\pi)^{n/2}(\sigma_N^2 + \sigma_M^2)^{n/2}} \int_{-\infty}^{\infty} \exp\left[-\frac{\xi^2}{2(\sigma_N^2 + \sigma_M^2)}\right] \cdot [1 - g(\vec{\xi})]^{M-1}\, d\vec{\xi} \tag{5'}$$

where the function $g(\vec{\xi})$ is

$$g(\vec{\xi}) = \frac{1}{(2\pi)^n \sigma^{2n}} \int_0^r \int_{-\infty}^{\infty} \cdot \exp\left\{-\frac{1}{2\sigma^2}\sum_i [x_i^2 + (ax_i + \xi_i + s_i)^2]\right\} d\vec{x}\, d\vec{s}. \tag{6'}$$

The evaluation of $g(\xi)$ in the Appendix leads to

$$g(\vec{\xi}) = C \int_0^r s^{n-1}\, ds \int_{-\pi/2}^{\pi/2} \cdot \exp\left[-\frac{1}{2\sigma^2(a^2+1)}(\xi^2 + s^2 + 2s\xi \sin\theta)\right] \cdot \cos^{n-2}\theta\, d\theta, \tag{7'}$$

where n is the dimensionality and r is computed from (21). The expressions for blur-type distortions are the same, except that σ_M^2 is omitted. This is a consequence of the fact off-diagonal elements of A are zero.

As in Section II, the expressions for error can be expressed in terms of the variable σ_N/σ, instead of in terms of these variables individually. Figs. 4 and 5 are examples of error rate computations as a function of noise for ten-dimensional Gaussian data and 255 classes. The solid lines are the approximations to the error rates for the three indicated values of a. Fig. 4 applies to blur-type distortions for which σ_M is zero, and Fig. 5 to scale-type distortions, where σ_M is as calculated above.

The fact that blur and scale error rates are not zero even at zero noise (when a is not 1) is demonstrated. The error rate

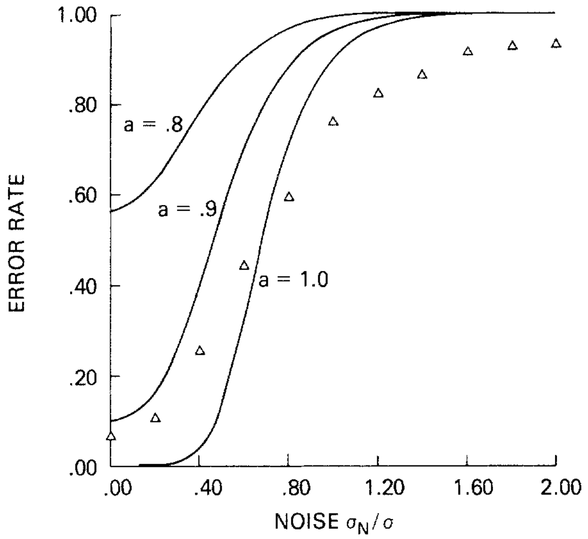

Fig. 5. Scale-type error rate as a function of noise.

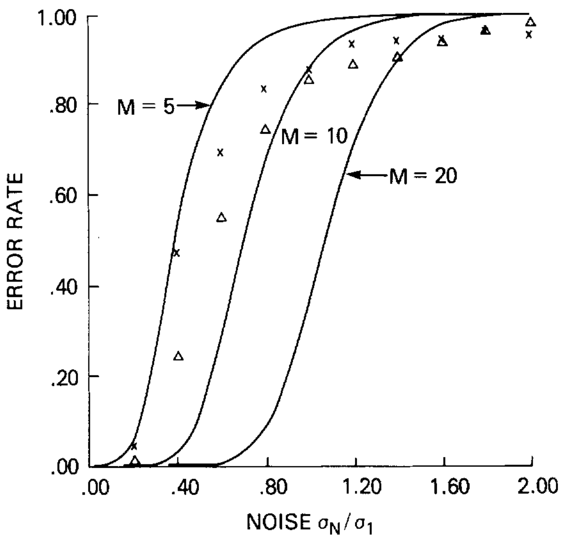

Fig. 6. Error rates curves for ship library (X) and Gaussian data with the same eigenvalues as the ship library (Δ).

curves tend to be flat near the origin, suggesting that a zero noise error rate can be a useful estimate of error over a range of noise values.

Data points are shown in both curves for comparison. Ten independent sets of normally distributed data with variances of unity, each with 255 values, were generated to be used as a ten-dimensional library. Measurements were generated by multiplying each coordinate of a library vector by one of the values of $\boldsymbol{a}$, adding to each coordinate Gaussian random noise from a distribution with variance σ_N^2, and, in the case of scale-type distortions adding additional noise from a distribution with variance $1 - a^2$. In Fig. 4, the value of $\boldsymbol{a}$ was 0.6, and in Fig. 5 it was 0.9. Classification was accomplished by measuring the Euclidian distance from the measurement to each member of the library. The measurement point was classified to the library member corresponding to the smallest distance, and error rates were computed accordingly.

V. Classification Examples

The error rate calculations of the preceding section assumed that the data to be classified are normally distributed in feature space. They assumed that the eigenvalues of the correlation matrix are equal and that distortion, when present, can be characterized by the same multiplicative parameter $\boldsymbol{a}$ for each coordinate. The approximate agreement between computed error rate curves and error rates for contrived Gaussian data suggest that a large body of data may be characterized by eigenvalues σ_i^2 and distortion parameters a_{ii}. In this section we will compare classification of actual data to classification of contrived data that have been generated to have Gaussian distributions with variances σ_i^2 that equal the eigenvalues of the actual data.

We use as a database a set of ship measurements that is maintained at the Naval Research Laboratory. This library contains, in tabular form, measured heights of the superstructure above the deck at twenty equally spaced intervals between bow and stern, as well as heights and positions of masts, and certain data on radars, guns, missile launchers, and directors. Ships of several nations, both military and commercial, are represented in the library. At the time of this writing, 255 ships are included. Our classifications have used only the information of the height above the deck.

Fig. 6 is a comparison of error rates computed by classifying data from the NRL ship library and classifying contrived data that were intended to simulate it. The contrived data consisted of 20 sets of Gaussian data, each with a variance equal to one of the eigenvalues of the height, covariance matrix. The elements M_{ij} of this matrix are defined

$$M_{ij} = \frac{1}{255} \sum_s (x_{is} - x_{iso})(x_{js} - x_{jso}), \tag{22}$$

where x_{is} is the ith height of ship s and x_{iso} is the average of the ith heights over the entire library.

A single feature vector, simulating a ship, consists of one value from each of these 20 sets. Gaussian noise with standard deviation σ_N was added to each coordinate of the ship library or contrived data library to create measurements. The abscissa is the ratio of the standard deviation of the noise to σ_1, the square root of the largest eigenvalue of the data. For reference, the three curves of Fig. 2 are reproduced, showing, for three dimensionalities, error rates when all σ are the same.

We note that the error rates that pertain to the actual data are somewhat higher than those for contrived data. This probably arises from clustering within the data set—there are sister ships, for example, that are fairly similar. The agreement is good enough, however, to suggest the correlation matrix eigenvalues as rough descriptors of the classification properties of the data.

The measured error rate curves are fairly close to those that might be expected when the data consist of $\boldsymbol{n}$ identical σ. This suggests that the data can be characterized by a dimensionality even when the σ are not the same—a dimensionality of about 7 would describe the Gaussian data of Fig. 6. The use of dimension to characterize data has been suggested by several workers, including Bennett [10] and Fukunaga and Flick [5], who computed it in different ways from that described here.

We can compute error rates under scale changes by magnifying each ship in the NRL library and comparing it with the unmagnified versions. The 20 heights above deck are expanded via linear interpolation so that each ship is represented by 22 heights and the center 20 of these are selected as a distorted feature vector. The measurement is created by adding noise to this feature vector. The vectors are classified by finding the distance between them and each of the 255 members of the data library. Results are in Fig. 7, where error rates are calculated versus σ_N/σ_1, defined previously.

The a_{ii} are calculated from (18). The normalized eigenvectors are magnified and truncated as described above and the inner product formed with an original eigenvector. Fig. 8 shows the a_{ii} computed in this way for three scale changes: 5, 10, and

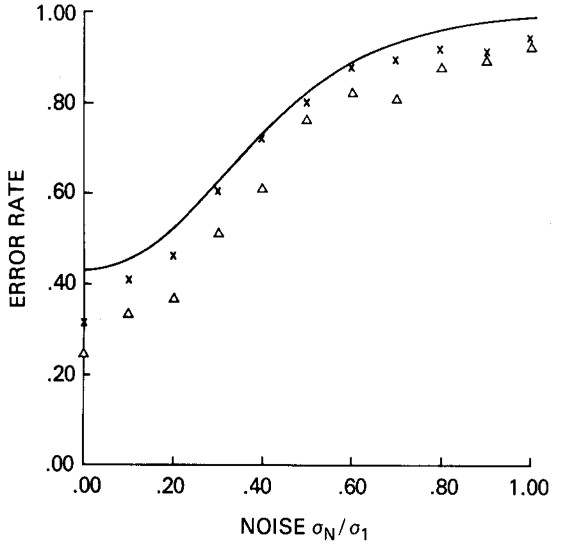

Fig. 7. Error rates for classifying magnified ship data (X), classifying contrived data (Δ), and calculated error rates.

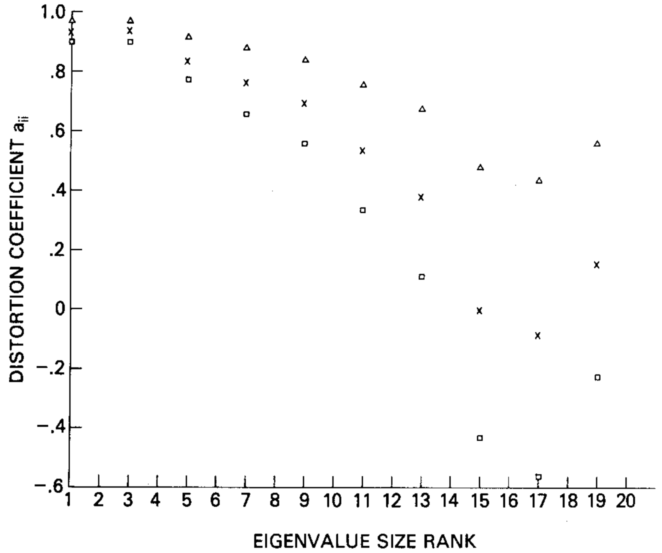

Fig. 8. Measured a_{ii} as a function of the rank of eigenvalue size for 5 percent (Δ), 10 percent (X) and 15 percent distortion (□).

15 percent. From the discussion in Section III, we know the eigenfunctions correspond roughly to Fourier components—they would correspond exactly if the images were stationary. The results in Fig. 8 show that high spatial frequency components are affected most by scale changes.

For comparison, contrived data with Gaussian distributions were generated as described above. As before, the variance of each coordinate was equal to one of the eigenvalues of the covariance matrix of the ship library. To simulate magnification of 10 percent, each coordinate of the contrived data was multiplied by the appropriate a_{ii} from Fig. 8 and Gaussian random noise with a variance $\sigma_i^2(1 - a_{ii}^2)$, from (17), was added to each coordinate to simulate the effects of off-diagonal terms of the distortion matrix $\vec{A}$. An additional amount of Gaussian noise, with a variance σ_N^2 was added to simulate measurement noise. A given coordinate i of the contrived data thus was a sum of three elements from Gaussian distributions. The variances of these distributions were: $(a_{ii}\sigma_i)^2$, $\sigma_i^2(1 - a_{ii}^2)$, and σ_N^2. This corresponds to the coordinate in the library with a standard deviation of σ_i.

The contrived data were classified using the least distance criterion of (1). Error rates for actual data are slightly higher, possibly for the same reasons as in Fig. 6, but the results are close enough to encourage the characterization of data by the parameters a_{ii}.

The solid line of Fig. 7 shows the result of a plot of (5′) for a dimensionality of 7 and an $\boldsymbol{a}$ of 0.897. This single number to characterize distortion was the weighted average of the first 7 values in Fig. 8 that correspond to 10 percent distortion.

$$a^2 = \frac{\sum_{i=1}^{7} \sigma_i^2 a_{ii}^2}{\sum_{i=1}^{7} \sigma_i^2}. \tag{23}$$

The analytical result, based on only two parameters, dimensionality $\boldsymbol{n}$ and average distortion $\boldsymbol{a}$, is in qualitative agreement with the classification results for both actual and contrived data.

VI. Summary

We have shown that the ability of a feature extraction algorithm to classify data is determined by a relatively small number of parameters when the data are Gaussian. The results of Fig. 7 demonstrate this especially; error rates computed from σ_i and a_{ii} agree fairly well with those obtained by classifying actual data, and reasonable agreement is seen with the analytical curves, based only on σ, $\boldsymbol{n}$, and $\boldsymbol{a}$. It is worth noting that the computational methods used to arrive at the error rates based on the parameters σ_i and a_{ii} are significantly different from those used to classify the actual data. The former method included only combinations of Gaussian data, while the latter had to simulate the actual magnification process for each ship.

The approach that we have followed has demonstrated how classifiability depends both on the feature set used and on the data that are classified. We believe that the application of feature sets to specific data can be analyzed in a straightforward way by computing the parameters described above. Feature sets can be easily compared in this way. Moreover, the performance of feature sets can be extrapolated to situations not represented by the data.

Appendix

The function $g(\vec{\xi})$ from (6′) is

$$g(\vec{\xi}) = \frac{1}{(2\pi)^n \sigma^{2n}} \int_0^r \int_{-\infty}^{\infty} \cdot \exp\left\{-\frac{1}{2\sigma^2} \sum_i \left[x_i^2 + (ax_i + \xi_i + s_i)^2\right]\right\} d\vec{x}\, d\vec{s}. \tag{A1}$$

In (6), $\boldsymbol{r}$ is replaced by $|\xi|$ and $\boldsymbol{a}$ is 1. The inner integral is a convolution that can be evaluated in each coordinate x_i separately. This evaluation leads to

$$g(\vec{\xi}) = \frac{1}{(2\pi)^{n/2} \sigma^n (a^2 + 1)^{n/2}} \int_0^r \cdot \exp\left[\frac{1}{2\sigma^2(a^2 + 1)} \sum_i (\xi_i + s_i)^2\right] d\vec{s}. \tag{A2}$$

If ξ and s are the lengths of $\vec{\xi}$ and $\vec{s}$, and θ is the angle between $\vec{s}$ and the perpendicular to $\vec{\xi}$, the expression for $g(\vec{\xi})$ is

$$g(\vec{\xi}) = \frac{1}{(2\pi)^{n/2}\sigma^n(a^2+1)^{n/2}} \int_0^r \cdot \exp\left[-\frac{1}{2\sigma^2(a^2+1)}(\xi^2 + s^2 + 2s\xi \sin\theta)\right] d\vec{s}. \quad \text{(A3)}$$

The volume element $d\vec{s}$ is a product of individual volume elements

$$d\vec{s} = ds_1\, ds_2 \cdots ds_n. \quad \text{(A4)}$$

We will define our coordinate system so that $\vec{s}_1$ is parallel to $\vec{\xi}$ and $\vec{s}_2$ through $\vec{s}_n$ are perpendicular to $\vec{\xi}$. If $\vec{t}$ is the projection of $\vec{s}$ on this subspace, these are related to the angle θ through

$$s_1 = s \sin\theta$$
$$t = s\cos\theta. \quad \text{(A5)}$$

A volume element $d\vec{s}$ can be expressed in terms of a product of volume elements in each of these subspaces. The integrand is spherically symmetric in the subspace perpendicular to $\vec{\xi}$, and the integration can be performed over the variable $\boldsymbol{t}$. The corresponding volume element is expressed in spherically symmetric form. The total volume element $d\vec{s}$ is

$$d\vec{s} = \frac{(n-1)\pi^{(n-1)/2}}{\Gamma\left(\frac{n+1}{2}\right)} t^{n-2}\, dt\, ds_1. \quad \text{(A6)}$$

The variables t, s_1 can be transformed to s, θ_i:

$$dt\, ds_1 = sd\, sd\theta. \quad \text{(A7)}$$

Combining (A5)-(A7) gives for the volume element $d\vec{s}$

$$d\vec{s} = \frac{(n-1)\pi^{(n-1)/2}}{\Gamma\left(\frac{n+1}{2}\right)} \cos^{n-2}\theta\, d\theta\, s^{n-1}\, ds \quad \text{(A8)}$$

and the expression for $g(\vec{\xi})$ becomes

$$g(\vec{\xi}) = C\int_0^r s^{n-1}\, ds \int_{-\pi/2}^{\pi/2} \exp\left[-\frac{\xi^2 + s^2 + 2s\xi\sin\theta}{2\sigma^2(a^2+1)}\right] \cdot \cos^{n-2}\theta\, d\theta \quad \text{(A9)}$$

where constant C in (7′) is

$$C = \frac{(n-1)}{\sqrt{\pi}\,\Gamma\left(\frac{n+1}{2}\right)} \cdot \frac{1}{(a^2+1)^{n/2}\sigma^n} \quad \text{(A10)}$$

and in (7) is

$$C = \frac{(n-1)}{\sqrt{\pi}\,\Gamma\frac{(n-1)}{2}} \frac{1}{2^{n/2}\sigma^n}. \quad \text{(A11)}$$

Acknowledgment

The author is indebted to K. Fukunaga of Purdue University for many discussions on classification theory during several months when he was a resident at NRL as an IGPA fellow. He is also grateful to D. Kerr of NRL for providing the library of ship data used in the classification experiments in Section V.

References

[1] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 21-27, Jan. 1967.

[2] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* New York: Wiley-Interscience, 1973, p. 100.

[3] R. D. Short and K. Fukunaga, "The optimal distance measure for nearest neighbor classification," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 622-627, Jan. 1981.

[4] S. Johns, "On some classification problems–I," *Sankhya: Ind. J. Statist.*, vol. 22, pp. 301-308, June 1960.

[5] K. Fukunaga and T. E. Flick, "Density identification and risk estimation," Purdue Univ. Tech. Rep. TR-EE 82-36, Nov. 1982.

[6] M. J. Lahart, "Optical area correlation with magnification and rotation," *J. Opt. Soc. Amer.*, vol. 60, pp. 319-325, Mar. 1970.

[7] ——, "Erratum to 'Optical area correlation with magnification and rotation,'" *J. Opt. Sci. Amer.*, vol. 61, p. 985, July 1971.

[8] H. Mostafavi and F. W. Smith, "Image correlation with geometric distortion. Part I: Acquisition performance," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-14, pp. 487-493, May 1978.

[9] D. Casasent and A. Furman, "Sources of correlation degradation," *Appl. Opt.*, vol. 16, pp. 1652-1661, June. 1977.

[10] R. S. Bennett, "The intrinsic dimensionality of signal collections," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 517-525, Sept. 1969.