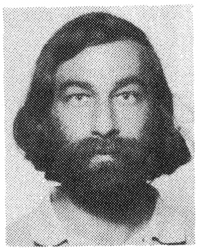
Martin D. Levine (S'59–M'66–SM'74) was born in Montreal, P.Q., Canada, on March 30, 1938. He received the B.Eng. and M.Eng. degrees in electrical engineering in 1960 and 1963, respectively, from McGill University, Montreal, P.Q., Canada, and the Ph.D. degree in electrical engineering in 1965 from the Imperial College of Science and Technology, University of London, London, England.

He is currently a Professor Electrical Engineering in the Department of Electrical Engineering at McGill University. From 1972 to 1973 he was a member of Technical Staff at the Image Processing Laboratory of the Jet Propulsion Laboratory, Pasadena, CA. During the 1979–1980 academic session, he was a Visiting Professor in the Department of Computer Science, Hebrew University, Jerusalem, Israel. His research interests are computer vision, biomedical pattern recognition, and artificial intelligence.

Dr. Levine is the Treasurer of the International Association for Pattern Recognition and an Associate Editor of *Computer Graphics and Image Processing*. He is also the General Chairman of the Seventh International Conference on Pattern Recognition to be held in Montreal during the summer of 1984. He is a member of the Order of Engineers of Quebec, the Computer Society, the Pattern Recognition Society, and the Association for Computing Machinery.



Steven W. Zucker (S'71–M'75) received the B.S. degree in electrical engineering from Carnegie-Mellon University, Pittsburgh, PA, in 1969, and the M.S. and Ph.D. degrees in biomedical engineering from Drexel University, Philadelphia, PA, in 1972 and 1975, respectively.

From 1974 to 1976 he was a Research Associate at the Computer Science Center, University of Maryland, College Park. He is currently an Associate Professor in the Computer Vision and Graphics Laboratory, Department of Electrical Engineering, McGill University, Montreal, P.Q., Canada. His research interests are in computer vision, human perception, and artificial intelligence.

Dr. Zucker is a member of Sigma Xi and the Association for Computing Machinery.

Systematic Feature Extraction

KENNETH A. BRAKKE, JAMES M. MANTOCK, MEMBER, IEEE, AND KEINOSUKE FUKUNAGA, FELLOW, IEEE

Abstract—**A systematic feature extraction procedure is proposed. It is based on successive extractions of features. At each stage a dimensionality reduction is made and a new feature is extracted. A specific example is given using the Gaussian minus-log-likelihood ratio as a basis for the extracted features. This form has the advantage that if both classes are Gaussianly distributed, only a single feature, the sufficient statistic, is extracted. If the classes are not Gaussianly distributed, additional features are extracted in an effort to improve the classification performance. Two examples are presented to demonstrate the performance of the procedure.**

Index Terms—**Gaussian classifier, nonlinear feature extraction, nonlinear mappings, quadratic feature extraction, quotient space determination, sequential feature extraction.**

I. Introduction

FEATURE extraction can be considered as a problem of finding a mapping (linear or nonlinear) that maps an n-dimensional measurement space down to an m-dimensional feature space without significantly increasing the degree of overlap between different class distributions. The determination of the m-dimensional feature space can frequently be viewed as a collection of m features. The manner in which most feature extraction algorithms determine the m features can be divided into two categories. The first approach is to determine all of the m features simultaneously. Examples of these are discriminant analysis [1] and feature extraction using the Bhattacharyya distance [2] or divergence [3]. The second approach can be found in Foley and Sammon [4]. Their procedure can be divided into steps as follows.

1) Extract a feature.

2) Map the data down to a lower dimensional space. This space should not contain any information about class separability present in the preceding extracted feature.

3) In the lower dimensional space a feature is extracted and tested for its contribution to classification performance. If the extracted feature shows little chance of improving the classification performance, the feature is discarded, and the procedure is terminated. If the extracted feature shows the potential to improve classification performance, the feature is retained, and the procedure is iterated [go to step 2)].

Features were extracted in Foley and Sammon's procedure by simply projecting n-dimensional data down to one axis, which was used as the feature. Thus, the form of the mapping was

$$z_j = V^T X_j + w \tag{1}$$

Manuscript received April 20, 1981; revised December 21, 1981. This work was supported in part by the National Science Foundation under Grant ECS-80-05482.

K. A. Brakke is with the Department of Mathematics, Purdue University, West Lafayette, IN 47907.

J. M. Mantock was with the Department of Electrical Engineering, Purdue University, West Lafayette, IN 47907 and the Aerospace Corporation, El Segundo, CA. He is now with Texas Instruments, Lewisville, TX 75067.

K. Fukunaga is with the Department of Electrical Engineering, Purdue University, West Lafayette, IN 47907.

0162-8828/82/0500-0291$00.75 © 1982 IEEE

where V is a constant column vector, X_j is the jth sample, and w is a scalar. We define the form in (1) as linear feature extraction.

We are stimulated by Foley and Sammon's idea and wondered if an extension to quadratic features was possible. That is, features of the form

$$z_j = X_j^T U X_j + V^T X_j + w \tag{2}$$

where U is a square matrix.

Let us introduce as our quadratic feature for the two class problem

$$h(X_j) = \frac{1}{2}(X_j - M_1)^T \Sigma_1^{-1} (X_j - M_1) - \frac{1}{2}(X_j - M_2)^T \cdot \Sigma_2^{-1} (X_j - M_2) + \frac{1}{2} \ln \frac{|\Sigma_1|}{|\Sigma_2|} \tag{3}$$

where M_i and Σ_i are the mean vector and covariance matrix of class ω_i and $|\Sigma_i|$ is the determinant of Σ_i. If the two distributions are Gaussian, $h(X)$ is a sufficient statistic for classification. In addition, $h(X)$ is a sufficient statistic for a wide class of unimodal symmetric distributions.

Jones [5] proved (see the Appendix) that regardless of the distributions for class ω_1 and class ω_2

$$E\{h(X) | X \epsilon \omega_1\} \leqslant 0 \text{ and } E\{h(X) | X \epsilon \omega_2\} \geqslant 0. \tag{4}$$

This indicates that $h(X)$ may carry classification information regardless of the distributions of class ω_1 and class ω_2. This property is particularly important, since the distributions in the lower dimensional subspaces are not likely to be Gaussianly distributed.

Based on the preceding it should be clear that the extension to the quadratic $h(X)$ in (3) has definite advantages.

Section II is devoted to mathematical preliminaries. Section III specifies a general version of the algorithm. This version makes no assumptions about the functional form of the extracted features. In Section IV a detailed feature extraction procedure is developed using $h(X)$ in (3) as a basis. Experimental results are presented in Section V and a summary is contained in Section VI.

II. Mathematical Preliminaries

A *partition* P of a set S is a collection of disjoint subsets of S whose union is S. A partition P corresponds to an equivalence relation $\equiv$, wherein two elements of S are related iff they are in the same member subset of the partition. The members of P are also called *equivalence classes* and P is also called the *quotient space* of S by $\equiv$, denoted $P = S \div \not\equiv$. The *quotient map* $q: S \rightarrow P$ is defined by

$$X \epsilon q(X) \qquad \text{for all} \quad X \epsilon S. \tag{5}$$

If μ is a measure on S and f: $S \rightarrow W$ is a function from S to a set W, then there is an *induced measure* $f_* \mu$ on W defined by

$$f_* \mu(A) = \mu(f^{-1}(A)) \qquad \text{for all} \quad A \epsilon W. \tag{6}$$

Theorem: If μ_1 and μ_2 are measures on S and f: $S \rightarrow S$ is such that $f_* \mu_1 = \mu_2$ and f preserves equivalence classes (i.e., $q(f(X)) = q(X)$), then

$$q_* \mu_1 = q_* \mu_2.$$

Proof: Let A be a subset of P. Then

$$\begin{aligned} q_* \mu_1(A) &= \mu_1(q^{-1}(A)) \\ &= \mu_1(f^{-1}(q^{-1}(A))) \\ &= f_* \mu_1(q^{-1}(A)) \\ &= \mu_2(q^{-1}(A)) \\ &= q_* \mu_2(A). \end{aligned} \tag{7}$$

Q.E.D.

III. The Feature Extraction Algorithm

The algorithm we propose for two class sequential general feature extraction is as follows.

1) Select a model for the distribution of class ω_1 and ω_2. Let μ_1 and μ_2 be the probability measures in S that the models require for classes ω_1 and ω_2, respectively. If the samples in fact are distributed according to our assumed model, then μ_1 and μ_2 allow the computation of a sufficient statistic in S for classification. If this is not the case, then sufficiency of the statistic in S is not guaranteed. Regardless, this statistic is used as the first extracted feature.

2) Find a partition P of S and a map f: $S \rightarrow S$ that preserves P and satisfies $f_* \mu_1 = \mu_2$. That is, we seek a mapping f and a partition P so that the distributions of the two classes in P are identical, if the samples are distributed according to our assumed models with μ_1 and μ_2.

3) Map the two classes to the quotient space P. Since $q_* \mu_1 = q_* \mu_2$ in P, a feature extracted in P cannot include any classification information contained in the previously extracted feature (computed using μ_1 and μ_2).

4) A feature is extracted in P. If the extracted feature provides a significant increase in classification performance, it is retained and the algorithm is iterated. If this is not the case, the extracted feature is discarded and the algorithm is terminated.

It is important to understand that it is not required that the assumed model in step 1) be correct. If it is, the algorithm extracts, as the first feature, a sufficient statistic. Subsequent extracted features do not add any information, so the algorithm terminates extracting the sufficient statistic as the only feature. When the assumed model is incorrect, we expect much of the classification information not contained in the statistic to be mapped to the quotient space P. The algorithm is then iterated using P as a basis.

IV. Application Using a Gaussian Statistic

In this section we develop the algorithm of Section III using a Gaussian model for the distributions. Suppose that ω_1 and ω_2 are two sample distributions in R^n (S from Section III) with mean vectors M_1 and M_2 and covariance matrices Σ_1 and

Σ_2 for classes ω_1 and ω_2, respectively, i.e., $\mu_1 = \{M_1, \Sigma_1\}$, $\mu_2 = \{M_2, \Sigma_2\}$. Our first feature is the sufficient statistic used for classifying two Gaussian distributions. One form of this is the Gaussian-minus-log-likelihood ratio $h(X)$ in (3). Since this statistic is invariant under nonsingular linear transformations, we may assume without loss of generality that the sample space is linearly transformed so that $\Sigma_1 = I$, where I is the identity matrix, and $\Sigma_2 = \Lambda$, where Λ is a diagonal matrix with positive diagonal entries.

We now need to determine $f\colon R^n \to R^n$ such that $f_*\mu_1 = \mu_2$. This can be accomplished by finding a linear transformation A^T such that

$$M_2 = A^T M_1 \tag{8}$$

and

$$\Lambda = A^T I A. \tag{9}$$

This is most easily achieved by initially introducing a translation

$$X = \tilde{X} + C \tag{10}$$

where C is chosen so that

$$M_2 = \tilde{M}_2 + C = A^T(\tilde{M}_1 + C) = A^T M_1$$

or

$$C = (A^T - I)^{-1}(\tilde{M}_2 - A^T\tilde{M}_1) \tag{11}$$

where $\tilde{M}_i$ is the mean vector of class ω_i prior to the translation. This translation satisfies (8). We shall choose A to be the diagonal matrix,

$$A = \Lambda^{1/2} \tag{12}$$

to satisfy (9). Henceforth, we assume (8) and (9) are satisfied, so that $f_*\mu_1 = \mu_2$.

The case where (11) has no solution ($\Lambda^{1/2} - I$ is singular) is discussed more fully later.

We now define a partition P and show that the transformation $f(X) = A^T X$ has the equivalence preserving property we require.

Define the set P of equivalence classes to be the set of trajectories of the differential equation

$$\dot{X}(t) = BX(t) \tag{13}$$

where

$$A^T = e^B. \tag{14}$$

Such a B exists by our earlier assumption of the positive defniteness of A. Solving (13) we get trajectories of the form

$$X(t) = e^{Bt}X(0) = \Lambda^{t/2}X(0). \tag{15}$$

Equation (15) shows that our map $f(X) = \Lambda^{1/2}X$ does map trajectories to trajectories. Thus, we have met all of the assumptions of the theorem. So for the quotient map $q\colon R^n \to P$ we have equality of the induced measures

$$q_*\mu_1 = q_*\mu_2. \tag{16}$$

The significance of (16) is that it guarantees that P contains no classification information represented in the $h(X)$ computed in S. As a result, if $h(X)$ is a sufficient statistic (the Gaussian model is correct), there will be no classification information *at all* in P. Hence, there is no feature that can be extracted in P that will increase classification performance.

Having completed our work on f we now turn to P. In its present form it is a set of trajectories. To perform the mapping into quotient space in a practical way, we need to find some method to individually and uniquely specify almost all of the trajectories. Observe that we need only specify which trajectory a sample is on, not its location on the trajectory. To accomplish this it is necessary to parameterize P as an $(n - 1)$-dimensional space. An obvious method is to choose a hypersurface in R^n that intersects each trajectory exactly once, except possibly for a subset of trajectories with probability measure zero. Another possibility is to introduce homogeneous coordinates as follows. In component form we know from (15) that

$$x_i(t) = \lambda_i^{t/2} x_i(0) \qquad j = 1, \cdots, n. \tag{17}$$

(Note that the sign of $x_i(t)$ remains constant along a trajectory.) By our earlier assumption that $A^T - I = \Lambda^{1/2} - I$ is invertible, we know that $\lambda_i \neq 1$ for all i. Thus, we may rewrite (17) as

$$\begin{aligned} &\operatorname{sgn}(x_i(t))\,|x_i(t)|^{\alpha/\ln\lambda_i} \\ &\qquad = e^{\alpha t/2}\operatorname{sgn}(x_i(0))\,|x_i(0)|^{\alpha/\ln\lambda_i} \end{aligned} \tag{18}$$

where $\operatorname{sgn}(\cdot)$ is the signum function and α is an adjustable constant convenient for numerical calculations. If we introduce the new variables

$$y_i(t) = \operatorname{sgn}(x_i(t))\,|x_i(t)|^{\alpha/\ln\lambda_i} \qquad i = 1, \cdots, n \tag{19}$$

(15) becomes

$$y_i(t) = e^{\alpha t/2} y_i(0)$$

or

$$\frac{y_i(t)}{y_i(0)} = e^{\alpha t/2}. \tag{20}$$

We observe that in the Y coordinate system all of the trajectories are lines beginning at the origin. This is true because $e^{\alpha t/2}$ is independent of i. This is not the case in the X coordinate system, since (17) clearly shows that $x_i(t)/x_i(0)$ is not independent of i. The fact that almost all of the trajectories, except for a subset of trajectories with probability measure zero, are lines beginning at the origin, suggests that a spherical coordinate decomposition will solve the parameterization problem. By retaining the angular components of any sample we uniquely associate it with its trajectory. Note that t specifies where on the trajectory the sample is located. Since we need only specify the trajectory, t need not be computed.

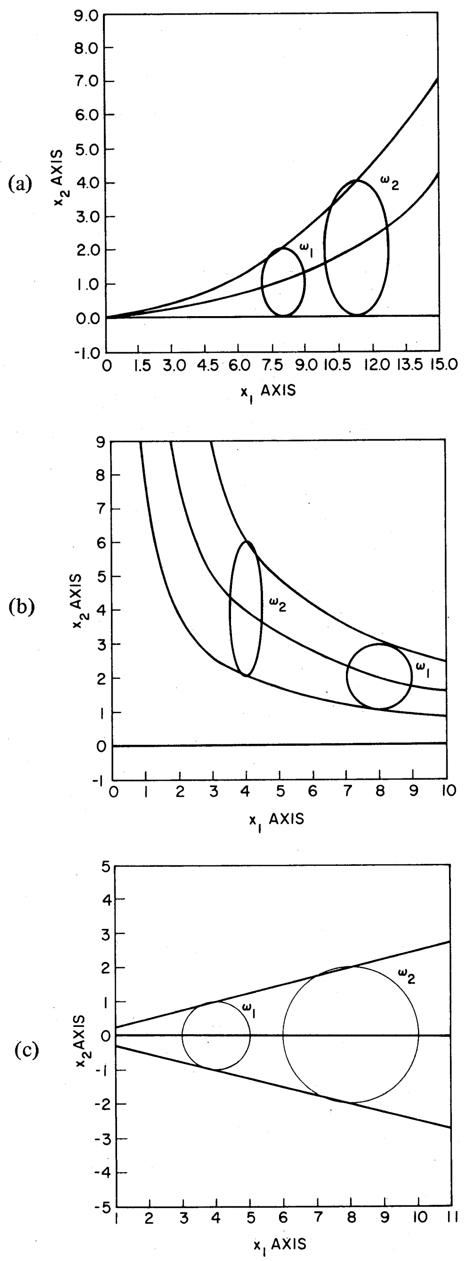


Fig. 1. (a) Three trajectories for $M_1 = \begin{bmatrix}8.0\\1.0\end{bmatrix}$, $\Sigma_1 = I$ and $M_2 = \begin{bmatrix}11.3\\2.0\end{bmatrix}$, $\Sigma_2 = \begin{bmatrix}2.0 & 0.0\\0.0 & 4.0\end{bmatrix}$ with $\alpha = \ln 2$. (b) Three trajectories for $M_1 = \begin{bmatrix}8.0\\2.0\end{bmatrix}$, $\Sigma_1 = I$ and $M_2 = \begin{bmatrix}4.0\\4.0\end{bmatrix}$, $\Sigma_2 = \begin{bmatrix}0.25 & 0.0\\0.0 & 4.0\end{bmatrix}$ with $\alpha = \ln 2$. (c) Three trajectories for $M_1 = \begin{bmatrix}4.0\\0.0\end{bmatrix}$, $\Sigma_1 = I$ and $M_2 = \begin{bmatrix}8.0\\0.0\end{bmatrix}$, $\Sigma_2 = 4I$ with $\alpha = \ln 2$.

The conversion to the angular components of spherical coordinates is easily performed as

$$\begin{aligned}
\theta_1 &= \arctan\left(\frac{y_1}{y_2}\right)\\
\theta_2 &= \arctan\left[(y_1^2 + y_2^2)^{1/2}/y_3\right]\\
\vdots\quad & \quad\vdots\\
\theta_{n-1} &= \arctan\left[\left(\sum_{i=1}^{n-1} y_i^2\right)^{1/2}/y_n\right].
\end{aligned} \tag{21}$$

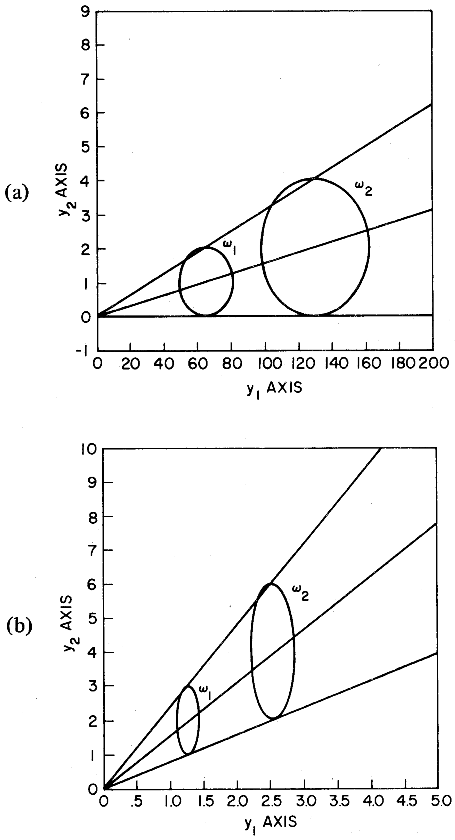


Fig. 2. (a) Y coordinate representation of Fig. 1(a). (b) Y coordinate representation of Fig. 1(b).

Note that the range of θ_1 is $(0, 2\pi)$ and the range of $\theta_2, \cdots, \theta_{n-1}$ is $(0, \pi)$.

Figs. 1 and 2 show three examples of Gaussian distributions with three trajectories labeled in the X and Y coordinate spaces, respectively. Note that for Fig. 1(c) the Y coordinate representation is identical to the X coordinate representation [for this reason there is no Fig. 2(c)]. This is possible because of the introduction of the control parameter α. By trying to get $\alpha/\ln\lambda_i$ as close to one as possible for all i, we can minimize the distortion of the original data. Unfortunately, for wide ranges of $1/\ln\lambda_i$ and/or different signs for the various i it is not clear how to best choose α.

We now detail the proposed feature extraction procedure.

1) Compute the sample mean vector and covariance matrix for each class.

2) Compute $h(X)$ in (3), and select this as a feature.

3) If $h(X)$ is an effective feature, retain it and continue. Otherwise, stop.

4) Simultaneously diagonalize the data.

5) Compute C in (11) and use it to translate the data sets. Select an α so that $\alpha/\ln\lambda_i$ is as close to one as possible for all i. (This is to minimize data distortion.)

6) For each sample X in the data set compute Y using (19).

7) Convert Y to multidimensional spherical coordinates, retaining only the angular components, using (21).

8) Go to step 2).

After the first pass through the procedure the sample space will be bounded. As a result, it is impossible for the data to be reduced to two dimensions using the nonlinear mapping algorithm. The NN error estimate in the two-dimensional subspace indicated an error rate of 49 percent, which implies that the two distributions were essentially identical. When a second feature was extracted from the transformed data space and combined with the first feature, there was no change in the NN error estimate.

B. Experiment 2

In the second experiment the first distribution was 100 Gaussian samples with $M_1^T = [-1.0\ -4.0\ -1.0]$ and $\Sigma_1 = I$. The second distribution was formed using two Gaussian distributions with parameters

$$M_{21} = \begin{bmatrix} 2.0 \\ 0.0 \\ 0.0 \end{bmatrix}, \quad \Sigma_{21} = \begin{bmatrix} 0.1 & 0.0 & 0.0 \\ 0.0 & 4.1 & 0.0 \\ 0.0 & 0.0 & 4.1 \end{bmatrix}$$

and

$$M_{22} = \begin{bmatrix} -2.0 \\ 0.0 \\ 0.0 \end{bmatrix}, \quad \Sigma_{22} = \begin{bmatrix} 0.1 & 0.0 & 0.0 \\ 0.0 & 4.1 & 0.0 \\ 0.0 & 0.0 & 4.1 \end{bmatrix}$$

where the second subscript serves as an identifier for the two distributions. Fifty samples from each distribution were combined to provide 100 samples. The NN error estimate of the data set indicated an error rate of 8 percent. The first feature was extracted and the NN error estimate indicated an 11 percent error rate. The data set was reduced to two dimensions using the nonlinear mapping algorithm. The NN error estimate in the two-dimensional subspace produced an error rate of 40 percent indicating that improvement was possible. A second feature was extracted using the two-dimensional subspace. The resulting NN error estimate of the two features was 8 percent. Since this was equal to the error rate of the original data, the feature extraction process was terminated.

VI. Summary

We have proposed a method to perform systematic feature selection. The specific version of the algorithm is presented that is based on the minus-log-likelihood ratio under a Gaussian assumption. This approach provides the advantage that if the classes are Gaussianly distributed, only one feature will be extracted. This was substantiated in an experiment that was presented. If the classes are not Gaussianly distributed, additional features may be extracted. Since they are all based on the Gaussian minus-log-likelihood ratio, an interesting interpretation can be made. The first feature is based on first- and second-order moment information in the original data space. Due to the nonlinear nature of the mapping algorithm, higher order information is mapped so that it can be measured, in some sense, in the first- and second-order moments. Therefore, even though subsequent features are also only based on first and second moments in lower dimensional spaces, they contain higher order information from the original space. An experiment was presented suggesting the validity of this statement.

Appendix

Theorem: We define $h(X)$ as

$$h(X) = \frac{1}{2}(X - M_1)^T \Sigma_1^{-1} (X - M_1) - \frac{1}{2}(X - M_2)^T \Sigma_2^{-1} \cdot (X - M_2) + \frac{1}{2} \ln \frac{|\Sigma_1|}{|\Sigma_2|} \tag{A1}$$

where M_1 and M_2 are mean vectors, and Σ_1 and Σ_2 are covariance matrices of two distributions, in classes ω_1 and ω_2, respectively, in n-dimensional Euclidean vector space. With $h(X)$ so defined

$$E\{h(x) | X \epsilon \omega_1\} \leqslant 0 \tag{A2}$$

and

$$E\{h(X) | X \epsilon \omega_2\} \geqslant 0. \tag{A3}$$

regardless of the distribution of X.

Proof: Since $h(X)$ is invariant under any nonsingular linear transformation, we first transform the data as

$$Y = A^T X. \tag{A4}$$

The nonsingular matrix A is chosen so that

$$\Sigma_1 = I \quad \text{and} \quad \Sigma_2 = \Lambda \tag{A5}$$

where I is the identity matrix and Λ is a diagonal matrix. Thus, we have

$$h(X) = h(Y) = \frac{1}{2}(Y - D_1)^T (Y - D_1) - \frac{1}{2}(Y - D_2)^T \Lambda^{-1} (Y - D_2) + \frac{1}{2} \ln \frac{|I|}{|\Lambda|} \tag{A6}$$

where D_1 and D_2 are the mean vectors for classes ω_1 and ω_2, respectively, in the transformed space. The conditional expectations are computed as

$$E\{h(X) | X \epsilon \omega_1\} = E\{h(Y) | Y \epsilon \omega_1\} = \frac{1}{2} \operatorname{tr} [I - \Lambda^{-1}] - \frac{1}{2}(D_1 - D_2)^T \Lambda^{-1} (D_1 - D_2) + \frac{1}{2} \ln \frac{|I|}{|\Lambda|} \tag{A7}$$

and

$$E\{h(X) | X \epsilon \omega_2\} = E\{h(Y) | Y \epsilon \omega_2\} = \frac{1}{2} \operatorname{tr} [\Lambda - I] + \frac{1}{2}(D_1 - D_2)^T (D_1 - D_2) + \frac{1}{2} \ln \frac{|I|}{|\Lambda|} \tag{A8}$$

Expressing these in terms of the components λ_i, the ith diagonal element of Λ, and d_{ji}, the jth element of the D_i, we get

$$E\{h(Y) | Y \epsilon \omega_1\} = \frac{1}{2} \sum_{i=1}^{n} \left[\left(1 - \frac{1}{\lambda_i}\right) - (d_{2i} - d_{1i})^2 / \lambda_i + \ln \frac{1}{\lambda_i} \right] \tag{A9}$$

and

$$E\{h(Y)|Y\epsilon\omega_2\} = \tfrac{1}{2}\sum_{i=1}^{n} [(\lambda_i - 1) + (d_{2i} - d_{1i})^2 + \ln 1/\lambda_i]. \tag{A10}$$

We consider the *i*th term of the summation in (A9) first. Since $-(d_{2i} - d_{1i})^2/\lambda_i$ is never positive, we need only establish that

$$f(\lambda_i) = 1 - \frac{1}{\lambda_i} - \ln \lambda_i \tag{A11}$$

is never positive. This is readily confirmed by evaluating the first and second derivative of (A11). We have shown that every term in the summation of (A9) is nonpositive, hence

$$E\{h(X)|X\epsilon\omega_1\} \leqslant 0. \tag{A12}$$

In a similar fashion the proof of (A3) requires establishing that

$$g(\lambda_i) = \lambda_i - 1 - \ln \lambda_i \tag{A13}$$

is never negative. Again, simple differentiation readily confirms this. Since all of the terms of the summation of (A10) are nonnegative, we have

$$E\{h(X)|X\epsilon\omega_2\} \geqslant 0. \tag{A14}$$

Q.E.D.

References

[1] O. R. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* New York: Wiley, 1973, pp. 114-121.

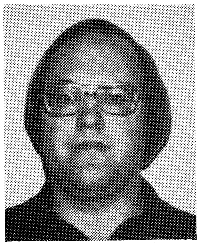
[2] K. Fukunaga and T. F. Krile, "Calculation of Bayes recognition error for two multivariate Gaussian distributions," *IEEE Trans. Comput.*, vol. C-18, pp. 220-229, Mar. 1969.

[3] J. T. Tou and R. P. Heyden, "Some approaches to optimum feature selection," in *Computer and Information Sciences*, vol. II, J. T. Tou, Ed. New York: Academic, 1967, pp. 57-89.

[4] D. H. Foley and J. W. Sammon, "An optimal set of discriminant vectors," *IEEE Trans. Comput.*, vol. C-24, pp. 281-289, Mar. 1975.

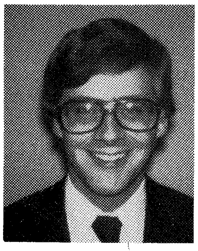
[5] L. K. Jones, private communication.

[6] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* New York: Academic, 1972.



Kenneth A. Brakke received the Ph.D. degree in mathematics from Princeton University, Princeton, NJ, in 1975.

Since 1975 he has been with the Department of Mathematics, Purdue University, West Lafayette, IN. His principal interest is geometric measure theory applied to shapes produced by surface tension.



James M. Mantock (S'78-M'82) was born in Columbus, OH, on December 19, 1954. He received the B.S.E.E., M.S.E.E., and Ph.D. degrees from Purdue University, West Lafayette, IN, in 1976, 1977, and 1981, respectively.

From 1977 to 1981 he was with the Aerospace Corporation, El Segundo, CA, working in satellite image understanding. During the period from 1979 to 1981, he was awarded a corporate fellowship for pattern recognition research at Purdue University. He is currently with Texas Instruments, Lewisville, TX, working in FLIR image understanding. His research interests include pattern recognition and signal processing.

Dr. Mantock is a member of Eta Kappa Nu and Tau Beta Pi.

Keinosuke Fukunaga (M'66-SM'74-F'79) received the B.S. degree in electrical engineering from Kyoto University, Kyoto, Japan, in 1953, the M.S.E.E. degree from the University of Pennsylvania, Philadelphia, in 1959, and the Ph.D. degree from Kyoto University, in 1962.

From 1953 to 1966 he was with the Mitsubishi Electric Company, Japan, first with the Central Research Laboratories, working on computer applications in control systems, and then with the Computer Division, where he was in charge of hardware development. Since 1966 he has been with Purdue University, West Lafayette, IN, where he is currently a Professor of Electrical Engineering. In the summers he has worked with a number of organizations including the IBM Corporation at Endicott, NY (1967 and 1968), IBM at Rochester, MN (1969), the Army White Sands Missile Range, NM (1977), and Exxon Production Research, TX (1981). He spent his sabbatical year of 1974-1975 with the Central Research Laboratories, Mitsubishi Electric Company. He has served as a Consultant to the Lincoln Laboratory, Massachusetts Institute of Technology, Cambridge, and the Health Science Center, University of Oklahoma. He was an Associate Editor of the IEEE Transactions on Information Theory for Pattern Recognition from 1977 to 1980. He is the author of *Introduction to Statistical Pattern Recognition* (New York: Academic, 1972).

Dr. Fukunaga is a member of Eta Kappa Nu.