Reprinted from

Symposium on

Machine Processing of

Remotely Sensed Data

June 3 - 5, 1975

The Laboratory for Applications of
Remote Sensing

Purdue University
West Lafayette
Indiana

IEEE Catalog No.
75CH1009-0 -C

DIVERGENCE ANALYSIS OF BENDIX FEATURE EXTRACTION AND CLASSIFICATION SYSTEM

R. M. Dye, C. S. Chen

Bendix Aerospace System Division, Ann Arbor, Michigan;
The University of Akron, Akron, Ohio

ABSTRACT

Many investigators reporting on feature extraction algorithms expect a decrease in recognition accuracy as an inevitable consequence of information loss. A feature extraction procedure is introduced which through empirical study indicates an improvement in recognition rate beyond that of a maximum likelihood classifier while permitting computational economy as a result of dimensionality reduction. Average divergence is shown to have increased after the application of the feature extraction procedure.

I. Introduction

To realize computational economy of a pattern recognition system, the measurement data is generally treated in order to extract effective feature and to reduce the dimensionality of the data. Many investigators reporting on dimension reducing algorithms expect a reduction in accuracy as an inevitable consequence of information loss.

In this paper, a new feature extraction and decision algorithm is introduced which is different from others in that the data transformation is performed for each category statistics. After the transformation, the discriminant function in the transformed domain for each category is a sum of squares of the transformed vector components, and hence could not decrease as the individual components are computed, squared and summed. It is thus possible to reject the hypothesis that the measurement data belongs to category ω_i. Whenever the discriminant function $S(\omega_i)$ exceeds the lowest S so far encountered it proceeds to the next group without completing the calculation for category ω_i. Computation economy is then realized. It is shown that the average divergence increases as a result of this transformation indicating possible gain in information for discrimination.

II. Bendix Feature Extraction Algorithm

If a measurement vector consisting of n components $\underline{X}$ is given, the maximum likelihood classifier assigns that vector to the i^{th} pattern ω_i if the conditional probability $P(\omega_i/\underline{X})$ is the greatest among all h patterns [1,2], that is,

$$\text{if } P(\omega_i/\underline{X}) \geq P(\omega_j/\underline{X}),\ j=1,2,..,h$$

then decide $\underline{X} \in \omega_i$, where $P(\omega_i/\underline{X}) = \dfrac{p(\underline{X}/\omega_i)P(\omega_i)}{p(\underline{X})}$

$p(\underline{X})$ and $p(\underline{X}/\omega_i)$ are probability density and conditional probability density functions, $P(\omega_i)$ is a priori probability for pattern ω_i.

If the conditional probability density function is assumed to be Gaussian as

$$p(\underline{X}/\omega_i) = \frac{\exp\{-\frac{1}{2}(\underline{X}-\underline{M}_i)^T \underline{K}_i^{-1}(\underline{X}-\underline{M}_i)\}}{(2\pi)^{n/2}|\underline{K}_i|^{\frac{1}{2}}}$$

where $\underline{K}_i$ = covariance matrix and $\underline{M}_i$ = mean vector of the i^{th} category of the objects, then it is shown [1,2] that the discriminant function for the i^{th} category is

$$S_i(\omega_i) = \log_e|\underline{K}_i| + (\underline{X}-\underline{M}_i)^T \underline{K}_i^{-1}(\underline{X}-\underline{M}_i)$$

$$= C_i + (\underline{X}-\underline{M}_i)^T K_i^{-1}(\underline{X}-\underline{M}_i) \qquad (1)$$

The evaluation of the quadratic form in eq.(1) requires (n+1) times n multiply and add operations for each category. For a twenty-category problem with ten measurement variables, 2200 multiply and add operations would be performed for each pixel. In the search for computational economy, the following feature extraction algorithm is used[3,4,5].

(A) The procedure begins with a preliminary transformation

$$\underline{Y}_i = \underline{\Lambda}_i^{-\frac{1}{2}} \underline{\phi}_i^T (\underline{X} - \underline{M}_i) \quad i=1,2,3,..,h$$

where $\underline{\Lambda}_i$ is the diagonal eigenvalue matrix; and $\underline{\phi}_i$, the normalized eigenvector matrix and $\underline{M}_i$ the mean vector of the i^{th} category. The eigenvalues are so arranged that for

$$\underline{\Lambda}_i = \begin{bmatrix} (\lambda_i)_1 & 0 & \cdots & 0 \\ 0 & (\lambda_i)_2 & & \vdots \\ \vdots & 0 & & \\ \vdots & \vdots & & 0 \\ 0 & 0 & & (\lambda_i)_n \end{bmatrix}$$

$$(\lambda_i)_1 \geq (\lambda_i)_2 \geq \cdots \geq (\lambda_i)_n$$

The subscript i associated with λ indicates that the transformation is performed with respect to the statistics of the i^{th} category.

(i) When this transformation is applied to the observed vector arising from the i^{th} category, then

$$E\left[\underline{Y}_i \middle| \underline{X} \epsilon \omega_i\right] = 0,$$

$$E\left[\underline{Y}_i \underline{Y}_i^T \middle| \underline{X} \epsilon \omega_i\right] = \underline{\Lambda}_i^{-\frac{1}{2}} \underline{\phi}_i^T E\{(\underline{X} - \underline{M}_i)(\underline{X} - \underline{M}_i)^T\}$$

$$\underline{\phi}_i \underline{\Lambda}_i^{-\frac{1}{2}} = \underline{\Lambda}_i^{-\frac{1}{2}} \underline{\phi}_i^T \underline{K}_i \underline{\phi}_i \underline{\Lambda}_i^{-\frac{1}{2}} = \underline{\Lambda}_i^{-\frac{1}{2}} \underline{\Lambda}_i \underline{\Lambda}_i^{-\frac{1}{2}} = \underline{I} \quad \ldots.(2)$$

The implications are:

(a) The components of the vector $\underline{Y}_i$ are uncorrelated and have zero mean and unit variance.

(b) The expected value of squared Euclidean distance to the origin in the $\underline{Y}_i$ space is n, for

$$E\{\underline{Y}_i^T \underline{Y}_i\} = \sum_{j=1}^{n} E[(y_i)_j^2] = n$$

(ii) However, if the transformation is applied to the observations from non-i^{th} category, then

$$E\left[\underline{Y}_i \middle| \underline{X} \notin \omega_i\right] = 0$$

$$E\left[\underline{Y}_i \underline{Y}_i^T \middle| \underline{X} \notin \omega_i\right] =$$

$$\underline{\Lambda}_i^{-\frac{1}{2}} \underline{\phi}_i^T E\{(\underline{X} - \underline{M}_i)(\underline{X} - \underline{M}_i)^T \middle| \underline{X} \notin \omega_i\} \underline{\phi}_i \underline{\Lambda}_i^{-\frac{1}{2}}$$

$$= \underline{G}_i$$

This means that for all non-i^{th} category samples, the components of $\underline{Y}_i$ vector are correlated. It is the purpose of the next transformation to diagonalize the $\underline{G}_i$ matrix.

(B) Let $(g_i)_1, (g_i)_2, \ldots, (g_i)_n$ be the eigenvalues of $\underline{G}_i$ arranged into the descending order and $(\underline{v}_i)_1, (\underline{v}_i)_2, \ldots, (\underline{v}_i)_n$ be their corresponding normalized eigenvectors.

The eigenvector matrix of $\underline{G}_i$ is formed as

$$\underline{V}_i = [(\underline{V}_i)_1, (\underline{V}_i)_2 \cdots (\underline{V}_i)_n]$$

and the following transformation is used

$$\underline{Z}_i = \underline{V}_i^T \underline{Y}_i$$

(i) If the observation arises from the i^{th} category, the covariance matrix is still an identity matrix as

$$E\left[\underline{Z}_i \underline{Z}_i^T \middle| \underline{X} \epsilon \omega_i\right]$$

$$= \underline{V}_i^T E\left[\underline{Y}_i \underline{Y}_i^T \middle| \underline{X} \epsilon \omega_i\right] \underline{V}_i$$

$$= \underline{V}_i^T \underline{V}_i = \underline{I}$$

Therefore the components of $\underline{Z}_i$ are uncorrelated and each has unit variance. The expected value of squared Euclidean distance to the origin is n for

$$E\{(z_i)_1^2 + (z_i)_2^2 + \ldots + (z_i)_n^2\}$$

$$= \sum_{j=1}^{n} E[(z_i)_j^2] = n$$

(ii) However, if the observation is from the non-i^{th} category,

$$E\left[\underline{Z}_i \underline{Z}_i^T \middle| \underline{X} \notin \omega_i\right] = \underline{V}_i^T E\left[\underline{Y}_i \underline{Y}_i^T \middle| \underline{X} \notin \omega_i\right] \underline{V}_i$$

$$= \underline{V}_i^T \underline{G}_i \underline{V}_i = \begin{bmatrix} (g_i)_1 & 0 & \cdots & 0 \\ 0 & (g_i)_2 & & \vdots \\ \vdots & 0 & & 0 \\ \vdots & \vdots & & \\ 0 & 0 & & (g_i)_n \end{bmatrix} \quad (3)$$

The components of $\underline{Z}_i$ vector are also uncorrelated and

$$E[(z_i)_1^2] = (g_i)_1 \geq E[(z_i)_2^2]$$

$$= (g_i)_2 \geq \ldots \geq E[(z_i)_n^2] = (g_i)_n$$

The above transformations are performed for each of the h categories statistics. This amounts to calculating for each category through training the following:

(a) the eigenvalue matrix $\underline{\Lambda}_i$ and the normalized eigenvector matrix $\underline{\phi}_i$ of the i^{th} category covariance matrix $\underline{K}_i$.

(b) the mean vector $\underline{M}_i$ of the i^{th} category.

(c) the eigenvalues $(g_i)_j$, j=1,2,3,..., h arranged into the descending order and the normalized eigenvector matrix $\underline{V}_i$ of the non-i^{th} category distribution.

Once the training is completed, those information can be stored and are used in the pattern classification stage.

III. Decision Process and the Computational Economy.

The discriminant function $S(\omega_i)$ in $\underline{Z}$ domain is identical to that in $\underline{X}$ domain as

$$S(\omega_i) = C_i + \underline{Z}_i^T \underline{Z}_i$$
$$= C_i + (\underline{X}-\underline{M}_i)^T \underline{K}_i^{-1} (\underline{X}-\underline{M}_i)$$

Which indicates that the characteristics of the maximum likelihood classifier is retained if all components of $\underline{Z}_i$ vector are used in the pattern classification stage.

To realize the computational economy, note that the discriminant function $S(\omega_i)$ in the $\underline{Z}$ domain is a sum of squares of the components of $\underline{Z}_i$ and hence could not decrease as the individual components of $\underline{Z}_i$ are computed, squared and summed. It is thus possible to reject the hypothesis that $\underline{X} \varepsilon \omega_i$ whenever $S(\omega_i)$ exceeds the lowest S so far encountered and proceed to the next group without completing the calculation for category ω_i.

It is proven in the previous section that if the hypothesis $\underline{X}\varepsilon\omega_i$ is correct, the expected value of $S(\omega_i)$ is

$$E[S(\omega_i) \,|\, \underline{X}\varepsilon\omega_i] = E[C_i] + \sum_{j=1}^{n} E[(z_i)_j^2]$$
$$= C_i + n$$

on the other hand, if the hypothesis is incorrect, the expected value of $S(\omega_i)$ is

$$E[S(\omega_i) \,|\, \underline{X}\not\varepsilon\omega_i] = C_i + \sum_{j=1}^{n} (g_i)_j$$

If the eigenvalue $(g_i)_1$ is large, say greater than n, then the hypothesis can be rejected upon the evaluation of the first component of $\underline{Z}_i$ requiring only 11 multiply and add operations for the 10 variable example in the earlier section. The full benefits from such early rejection will be realized only when the first few components of $\underline{Z}_i$ computed are those associated with the highest relative variance as indicated by the elements of g_i.

IV. Divergence Analysis

The divergence $J(\omega_i,\omega_j)$ between two classes ω_i and ω_j is defined [6,7] as

$$J(\omega_i,\omega_j) = \int [P(\underline{X}|\omega_i) - P(\underline{X}|\omega_j)] \log \frac{P(\underline{X}|\omega_i)}{P(\underline{X}|\omega_j)} \, d\underline{X}$$

which is in a sense a measure of the degree of difficulty of distinguishing ω_i and ω_j. The larger the value of $J(\omega_i,\omega_j)$ is, the less the degree of difficulty of distinguishing between classes ω_i and ω_j.

If $P(\underline{X}|\omega_i)$ and $P(\underline{X}|\omega_j)$ are of Gaussian distributions with mean vectors $\underline{M}_i$, $\underline{M}_j$ and covariance matrices $\underline{K}_i$ and $\underline{K}_j$. $J(\omega_i,\omega_j)$ becomes [6,7],

$$J(\omega_i,\omega_j)$$
$$= \tfrac{1}{2}\mathrm{trace}[\underline{K}_i^{-1}(\underline{K}_j + (\underline{M}_i-\underline{M}_j)(\underline{M}_i-\underline{M}_j)^T)]$$
$$+ \tfrac{1}{2}\mathrm{trace}[\underline{K}_j^{-1}(\underline{K}_i + (\underline{M}_i-\underline{M}_j)(\underline{M}_i-\underline{M}_j)^T)] - n \qquad (4)$$

For the two-class Gaussian distributions, the recognition rate appears to be bounded above and below by an empirical relationship with the divergence [8,9,10, 11,12].

The use of divergence is extended to the multicategory case by taking the average overall class pairs as, for h classes,

$$J_{ave} = \frac{1}{h(h-1)} \sum_{i=1}^{h-1} \sum_{j=i+1}^{h} J(\omega_i,\omega_j)$$

For Gaussian distribution, J_{ave} becomes

$$J_{ave} = \frac{1}{h(h-1)} \left[\sum_{i=1}^{h-1} \sum_{j=i+1}^{h} J(\omega_i,\omega_j)\right]$$
$$= \frac{\mathrm{trace}}{2h(h-1)} \left[\sum_{i=1}^{h} \underline{K}_i^{-1} \sum_{\substack{j=1 \\ j\neq i}}^{h} (\underline{K}_j + (\underline{M}_i-\underline{M}_j)(\underline{M}_i-\underline{M}_j)^T)\right] \qquad (5)$$

If the Bendix transformation described in the previous section is used,

$$\underline{z}_i = \underline{V}_i^T \underline{\Lambda}_i^{-\frac{1}{2}} \underline{\Phi}_i^T (\underline{X} - \underline{M}_i)$$

$J(\omega_i, \omega_j)$ becomes

$$J(\omega_i,\omega_j) = \tfrac{1}{2}\{\sum_{K=1}^{n} E[(\underline{z}_j)_K^2 | \omega_i] + \sum_{K=1}^{n} E[(\underline{z}_i)_K^2 | \omega_j]\} - n \qquad (6)$$

and the average divergence J_{ave} is

$$J_{ave} = \frac{1}{2h(h-1)} \sum_{K=1}^{n} \sum_{i=1}^{h} \sum_{\substack{j=1 \\ j\neq i}}^{h} E[(\underline{z}_i)_K^2 | \omega_j] - \frac{n}{2} \qquad (7)$$

It is shown earlier that (eq.(3))

$$E[\underline{z}_i \underline{z}_i^T | \underline{z}_i \in \omega_i] = \sum_{\substack{j=1 \\ j\neq i}}^{h} P(\omega_j) E[\underline{z}_i \underline{z}_i^T | \omega_j]$$

$$= \begin{bmatrix} (g_i)_1 & 0 & \cdots & 0 \\ 0 & (g_i)_2 & & \vdots \\ \vdots & 0 & & \vdots \\ \vdots & \vdots & & 0 \\ 0 & 0 & & (g_i)_n \end{bmatrix} \qquad (8)$$

and if $P(\omega_j)$ in eq. (8) is equal to constant $\frac{1}{h-1}$, then from eq. (8)

$$\sum_{\substack{j=1 \\ j\neq i}}^{h} E[(\underline{z}_i)_K^2 | \omega_j] = (h-1)(g_i)_K \qquad (9)$$

Substitution of eq. (9) into eq. (7) yields

$$J_{ave} = \frac{1}{2h} \sum_{K=1}^{n} \sum_{i=1}^{h} (g_i)_K - \frac{n}{2} \qquad (10)$$

After the Bendix transformation, if only the first m components of $\underline{z}_i$ are retained for dimensionality reduction purpose, the average divergence becomes

$$J_{ave}^1 = \frac{1}{2h} \sum_{K=1}^{m} \sum_{i=1}^{h} (g_i)_K - \frac{m}{2} \qquad (11)$$

and the change in the average divergence is

$$\Delta J_{ave} = J_{ave} - J_{ave}^1 = \frac{1}{2h} \sum_{K=m+1}^{n} \sum_{i=1}^{h} (g_i)_K - \frac{(n-m)}{2} \qquad (12)$$

In a problem with n original variables, most of the n eigenvalues $(g_i)_K$, K=1,2,...., n will be substantially greater than unity and usually some will be less than unity. If only the first m components of $\underline{z}_i$ the corresponding eigenvalues $(g_i)_K$, K=1,2,..,m of which are greater than unity are retained, then ΔJ_{ave} of eq. (12) becomes

$$\Delta J_{ave} < \frac{1}{2h} \sum_{K=m+1}^{n} \sum_{i=1}^{h} 1 - \frac{(n-m)}{2} = 0 \qquad (13)$$

which means the average divergence after the application of the Bendix algorithm increases in value, and thus indicates possible information gain for discrimination.

V. Conclusion

A new feature extraction algorithm is introduced which is different from many other reported schemes in that the data transformation is performed as many times as there are number of categories. Hypothesis that each measurement belongs to a particular category is either accepted or rejected upon computation of the discriminant function of each category.

The average divergence is shown to have increased in value after the application of the algorithm indicating possible gain of information.

REFERENCES

1 Fukunaga, Keinosuke
Introduction to Statistical Pattern Recognition, Academic Press, New York, 1972.

2 Nilsson, N.J.
Learning Machines, McGraw-Hill, New York, 1965.

3 Dye, R.H.
"Multivariate Categorical Analysis-Bendix Style," BSR 4149. Bendix Corporation, Ann Arbor, Michigan, 1974.

4 Crawford, C.L., et al.
"Signature Data Processing Study," BSR 2949, Bendix Corporation, Ann Arbor, Michigan.

5 Chen, C.S.
"Bendix System of Multivariate Categorical Classification," Report to NASA Goddard Space Flight Center, Greenbelt, Maryland, June, 1974.

6 Kullback, A.
Information Theory and Statistics, Dover Publishing Co., N.Y., 1960.

7 Jeffreys, H.
Theory of Probability, Oxford University Press, 1948.

8 Kailath, T.
"The Divergence and Bhattacharyya Distance Measures in Signal Selection," IEEE Trans. Communication Technology, Vol. COM-15, pp. 52-60, February, 1967.

9 Toussaint, G.
"Comments on the Divergence and Bhattacharyya Distance in Signal Selection," IEEE Trans. Communication Technology, Vol. COM-20, p. 485, June, 1972.

10 Marill, J. & D. Green
"On the Effectiveness of Receptors in Recognition Systems," IEEE Trans. Information Theory, Vol. IT-9, No. 1, pp. 11-17, January, 1963.

11 Fu, K.S., et al.
"Feature Selection in Pattern Recognition," IEEE Trans. Systems Science and Gybernetics, Vol. SSC-6, No. 1, pp. 33-39, January, 1970.

12 Chen, C.H.
"Theoretical Comparison of a class of feature selection criteria in pattern recognition," IEEE Trans. Comput., Vol. C-20, pp. 1054-1056, September, 1971.

13 Chen, C.S.
"Divergence Analysis of Bendix Multivariate Categorical Classification System," Report to NASA Goddard Space Flight Center, Greenbelt, Maryland, August, 1974.

14 Whitsitt, S.J. & D.A. Landgrebe
"Simulation techniques for estimating error in the classification of normal patterns," LARS Information Note 040174, Purdue University, West Lafayette, Indiana, 1974.

15 Swain, P.H. & R.C. King
"Two effective feature selection criteria for multispectral remote sensing," International Conference on Pattern Recognition Proceeding, pp. 536-540, Washington, D.C., 1973.

16 Swain, P.H., et al.
"Comparison of the divergence and B-distance in Feature Selection," Information Note 020871, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana, February, 1971.