Conference on

Machine Processing of

Remotely Sensed Data

October 16 - 18, 1973

The Laboratory for Applications of
Remote Sensing

Purdue University
West Lafayette
Indiana

FEATURE EXTRACTION OF MULTISPECTRAL DATA†

R. B. Crane, T. Crimmins, J. F. Reyer

Environmental Research Institute of Michigan (ERIM)*
Ann Arbor, Michigan

I. ABSTRACT

A method is presented for feature extraction of multispectral scanner data. Non-training data is used to demonstrate the reduction in processing time that can be obtained by using feature extraction rather than feature selection.

II. INTRODUCTION

The data gathered by a multispectral scanner must be processed before much of the useful information is available. The amount of data can be inconveniently large when the ground area to be surveyed is large and many spectral channels of data are recorded. This is particularly true when using general purpose, sequential computers. One method of circumventing this problem, is to limit the number of spectra channels in the scanner itself. This method has the disadvantage of requiring that the spectral channels be chosen before or during the gathering of the data. The choice of spectral regions is dependent on the atmospheric conditions and the state of the various ground covers. A problem with this approach is that the factors needed to choose the spectral channels for a particular mission may not be accurately known in advance.

Another method of spectral band selection, now being used with most aircraft multispectral scanners, is to record a reasonably large number of spectral channels of data, and determining from the training data during processing which subset of channels should be used. Previous experience (Crane, 1972) has shown that 4 to 6 spectral channels provide almost as much recognition information on test data as do all of the channels available.

The experiments that will be described in this paper were designed to show an alternative method of data reduction. Subsets of linear combinations of recorded channels (Crane, to be published; Crane, 1972; Jegewski, 1973; Hsia, 1973; Quirein, 1972) rather than a subset of pure channels, were tested to see if a fewer number of linear combinations of channels could be used without sacrificing recognition accuracy. If the computation giving the linear combinations of channels can be carried out in less time than the difference in time in using for recognition a larger number of pure channels and a subset of linear combinations then a very practical advantage is achieved for sequentially organized computers. The practical impact may be that many applications may be reduced to using 2 or 3 channels (linear combinations) giving nearly the same recognition information as if all pure channels were used but requiring less time (cost) and storage.

Of course a pure channel is a special type of linear combination, with the advantage that the pure channel is available without additional processing. The additional processing required to form the linear combinations may be a problem. With analog data available, the linear combinations can be formed at the time the data is digitized. For data in digital format, it may be convenient to form the linear combinations when the data is converted into a format suitable for recognition, or during the preprocessing operation.

†This work was supported under NASA Contract NAS9-9784.
*Formerly Willow Run Laboratories of The University of Michigan.

A simple example can illustrate the formation of linear combinations and the performance that is possible. Consider the problem of recognizing one of four possible classes using one linear combination of two channel data. The data channels are assumed to be independent, each with variance σ^2. The location of the mean values for the four classes is shown in Figure 1. The calculation of the average probability of misclassification for this geometry was made by assuming that a linear decision rule would be used, that the data were normal and described by the means and variance, (i.e. the covariance terms are zero) and that the a priori probabilities of occurrence of each class are equal. The calculations were limited to pairwise evaluations, whereby the probabilities of misclassification possible for each pair of classes were not affected by the presence of the remaining class.

The calculation results are shown in Figure 2, where the average probability of misclassification is shown as a function of the angle between the abcissa and projection line that determines the linear combination. The performance that would be obtained by using either pure channel is found for angles of 0° and 90°. The lowest average probability of misclassification occurs for a linear combination described by an angle of approximately 15°. Thus for this particular example, a linear combination would be more desirable for recognition than a subset (i. e. a subset of one) of pure channels.

A procedure for finding the best linear combination in this case might be to start with any combination, and compute the average probability of misclassification. Then, repeat the calculation for a linear combination described by an angle close to the first angle. This procedure would then be repeated, always using an angle close to that angle which in the previous computation had provided the lowest average probability of misclassification, until a minimum was found. For this example, where there are two minima, the lowest would be the one that would be found most of the time.

A slight variation of the geometry previously described is shown in Figure 3. The only difference is the change in the variance, σ. Figure 4 depicts the calculated average probability of misclassification. The averages are all lower than those shown previously, a result of the reduced variance. In addition, there are now 5 minima, rather than 2, so that a minimum seeking technique is more dependent on the starting linear combination. The number of easily detectable minima can be reduced by artificially increasing the variance. This phenomena may lead to an improved minimum seeking technique.

III. FORMING LINEAR COMBINATIONS

The problem of finding a good method of choosing linear combinations is primarily one of finding a workable algorithm in three distinct steps: (1) develop a measure of performance; (2) develop a minimum seeking technique; and (3) find suitable starting points for initiating the minimum seeking technique. In addition, the algorithm should not require an excessive amount of computational time.

The performance measure used is similar to that employed to find a subset of pure channels and is derived from the linear decision rule now used routinely in this laboratory. The measure can be expressed as:

$$M = \sum_{i,j} \Phi\{1/2[(\mu_i - \mu_j)^t A^t (A \frac{R_i + R_j}{2} A^t)^{-1} A(\mu_i - \mu_j)]^{1/2}\} \qquad (1)$$

where the summation is for all signatures, the i-th class is distributed normally with mean vector μ_i and covariance matrix R_i, and $\Phi(X)$ is the normal distribution function. The row vectors of the mxn matrix A represent the linear combinations in question. (n is the number of pure channels and m is the number of linear combinations.) An advantage of Eq. (1) is that it can be developed directly from the maximum likelihood decision rule, so the approximations used can be enumerated and evaluated. In fact, Eq. (1) is approximately proportional to a constant minus the average probability of misclassification that would be measured.

A method has been developed to find a local minimum of a function of several variables by starting at a point and following a path of steepest descent by steps of variable but controllable size. Both the local gradient and the local curvature are used to estimate the path of steepest descent.

Finding starting points, the third step, is more difficult. The following are suggested starting points.

Best Subset of Channels Starting Point

Each individual channel can be thought of as a linear combination of channels. (The vector representing this combination has a 1 in the appropriate coordinate and 0's elsewhere.) Therefore, a subset of m channels can be thought of as a set of m linear combinations. Since there can be a very large number of subsets of m channels, rather than check through all of them to find the best one, we use a stepwise procedure to find a "good" one. Experience has shown that this good subset is usually either the best or second best one. This stepwise procedure successively adds the one channel which gives the lowest average probability of misclassification when used with the channels already selected. The linear combinations represented by this subset are then used as a starting point.

Norm Squared Starting Point

By replacing each covariance matrix by the average of all of them, the problem is reduced to minimizing the function

$$M(P) = \sum_i \phi(1/2||Pw_i||) \tag{2}$$

where the w_i are a fixed set of vectors and P ranges over all orthogonal projections of rank m. The number of vectors w_i is the total number of pairwise combinations of signatures. Each projection P correspoinds in a simple fashion to a matrix A in the original formulation (see [1]). The projection P which maximizes

$$\sum_i ||Pw_i||^2 \tag{3}$$

is found analytically and the corresponding A is used as a starting point.

Principal Eigenvector Starting Point

First calculate the average of all the covariance matrices. Then transform the data so that the average covariance matrix is the identity matrix. Let $N(\mu,R)$ be the distribution of all the transformed data in the training area lumped together. This distribution can be calculated from the distributions of the various materials if we can estimate the frequency of occurrence of each material. The starting point A is then taken as the matrix whose row vectors are the m orthogonal eigenvectors correspoinding to the m largest eigenvalues of the cavariance matrix R.

Clustered Starting Point

This method is based on the fact that if there are only two signatures and we are using linear discrimination, then there always exists a single linear combination channel which distinguishes exactly as well as all n channels no matter what the value of n. If there are many signatures, then for each pair S_i, S_j $(i \neq j)$ let v_{ij} be the unit vector corresponding to this best single linear combination. In general, the number of vectors $v_{i,j}$, will be greater than m. The $v_{i,j}$ are then clustered into m clusters. For each cluster C_k, k=1,...,m, a weighted average, w_k, of the $v_{i,j}$ in that cluster is computed. The starting point A is formed from the w_k as row vectors. The weights can be made to reflect the sensitivity of the recognition accuracy to the decision rule.

There is one additional problem concerning the determination of A that should be mentioned. If A is an m x n matrix, there are mn components to be determined. This number of components can be reduced to m(n - m) by the choice of a suitable canonical form for A. A canonical form is possible because the value of M obtained for any A is not changed if PA is substituted for A, where P is any nonsingular matrix. We actually use PA, where P is chosen to scale the average covariance matrix and the mean vectors of the materials to our data format. The canonical form we chose is:

$$A = \begin{matrix} & \tan\theta_{11} \tan\theta_{12} \cdots \\ I_m & \tan\theta_{21} \\ & \vdots \end{matrix} \qquad (4)$$

where I_m is the identity matrix with rank m. (For a specific example see Figure 1, where m = 3, n = 10, and the 10 pure channels have been rearranged in order of the wavelength).

The canonical form with the θ_{ij} has two advantages. The first is that, in general, a minimum number of unknown scalars must be found. The second advantage is that the minimization process can be accomplished by varying the θ_{ij} with a nearly uniform step size. It is not necessary to have large jumps in the values of the unknown scalars, which occur if the tan θ_{ij} are considered to be the unknown scalars.

IV. EXPERIMENTAL RESULTS

An experiment was devised to compare subsets of linear combinations with subsets of pure channels. The data used were one of the data sets previously employed to test our linear decision rule (Ref. 1). This particular set of data was chosen because of the difficulty we have noticed in obtaining satisfactory recognition with it. We felt that with relatively poor recognition accuracy, the test results would represent greater statistical accuracy. If only a few data points were incorrectly recognized, the test results would be too dependent on those few points.

The test procedure we used was to first select data that corresponded to 20 training fields. From these fields we developed statistics (means and covariance) for each of the 7 classes of materials. The statistics or signatures were then used to develop the decision rules which were applied to data that corresponded to 23 test fields different from the training fields. We then found the average correct recognition for each field, and then the average for each material. Finally we averaged recognition accuracies for the materials to obtain an average recognition accuracy for the data set. The computer programs were merely functional, not optimized for minimum computation time, so meaningful comparisons were not made.

The material classes consisted of bare soil and six vegetative species; alfalfa, barley, lettuce, sugar, safflower, and rye. The bare soil data tended to be atypical, because three or more pure channels of data provided almost perfect recognition, whereas all of the subsets of 3 linear combinations of channels provided reduced accuracy. Note that the various subsets of linear combinations were chosen to optimize over all the species, therefore it is not surprizing that they did less well for one of them. There is some evidence that one infrared channel or ratio of channels can be used to separate vegetative and non-vegetative materials. Thus for some applications of layered or sequential classifiers, bare soil may not be considered as a class to be recognized when discriminating among vegetation types. However, for this study, we retained bare soil as a class.

The test results are shown in Figure 5. The top line of each bar indicates the average recognition accuracy obtained for the 7 classes. The next line indicates accuracies for the 6 vegetative classes. Note that the recognition accuracy for the subset of 3 linear combinations was better than that obtained when subsets of either 3 or 4 best pure channels were used. In fact, the accuracy approached that obtained using all 10 pure channels, especially when only the 6 vegetative materials are considered.

Figure 5 shows the average recognition accuracy obtained for one subset of 3 linear combinations only. We actually tested three subsets of 3 linear combinations. Two of the subsets resulted from minimizing our measure function with two different starting points, (the first of these was used for Figure 5) and the third subset was an unweighted addition of channels. We obtained approximately the same average recognition accuracy for each of the subsets of linear combinations.

In Figure 6, the average recognition accuracies we obtained from the 3 subsets of 3 linear

combinations are compared with each other and with subsets of pure channels. Note the correspondence with the predicted accuracies, especially for subsets of pure channels without the bare soil class.

The first subset of linear combinations is shown in Figure 7, each row represents one combination. This matrix is not determined uniquely, because premultiplication by any nonsingular matrix results in a new set of linear combinations which would produce identical recognition performance.

The starting point for this set was the subset of 3 pure channels that we used for comparison.

The tentative conclusion that we drew from our test program was that the use of linear combinations may be a feasible method of spectral feature extraction to reduce overall processing time. The tests should be extended to include more data sets and different starting points.

Acknowledgements

The authors received valuable assistance through discussions with Dr. Quinten Holmes of NASA-MSC and with staff members H. Horwitz, R. Kauth, and J. Erickson of ERIM.

REFERENCES

Crane, R. B. and W. Richardson, "Performance Evaluation of Multispectral Scanner Classification Methods," Proceedings of the Eighth International Symposium on Remote Sensing of Environment, The University of Michigan, Ann Arbor, 2-6 October 1972.

Crane, R. B., W. Richardson and R. Hieber, "A Study of Techniques for Processing Multispectral Scanner Data," Environmental Research Institute of Michigan Technical Report 31650-155-T (to be published).

Crane, R. B., "Linear Combinations," presented at Quarterly Supporting Research and Technology Review, 26-28 September 1972, and reported in Environmental Research Institute of Michigan Progress Report 31650-147-L, September 15 to October 15, 1972.

Hsia, W. S. and J. P. de Figueiredo, "Optimal Feature Extraction - The Two Class Case," Institute for Computer Services and Applications, Rice University, Houston, Texas, ICSA-275-025-010, May 1973.

Jegewski, D. J., "Optimal Feature Extraction by a Linear Transformation, Revision 1," MSC Internal Note No. 73-FM-19, 5 March 1973.

Quirein, J. A., "An Interactive Approach to the Feature Selection Classification Problem," TRW Systems, Earth Resources Technology, Houston Operations, Technical Note 99900-H019-RO-00, December 1972.

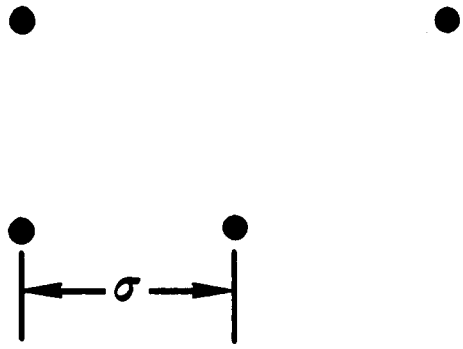

Figure 1. Location of Means for Example with Large Variance

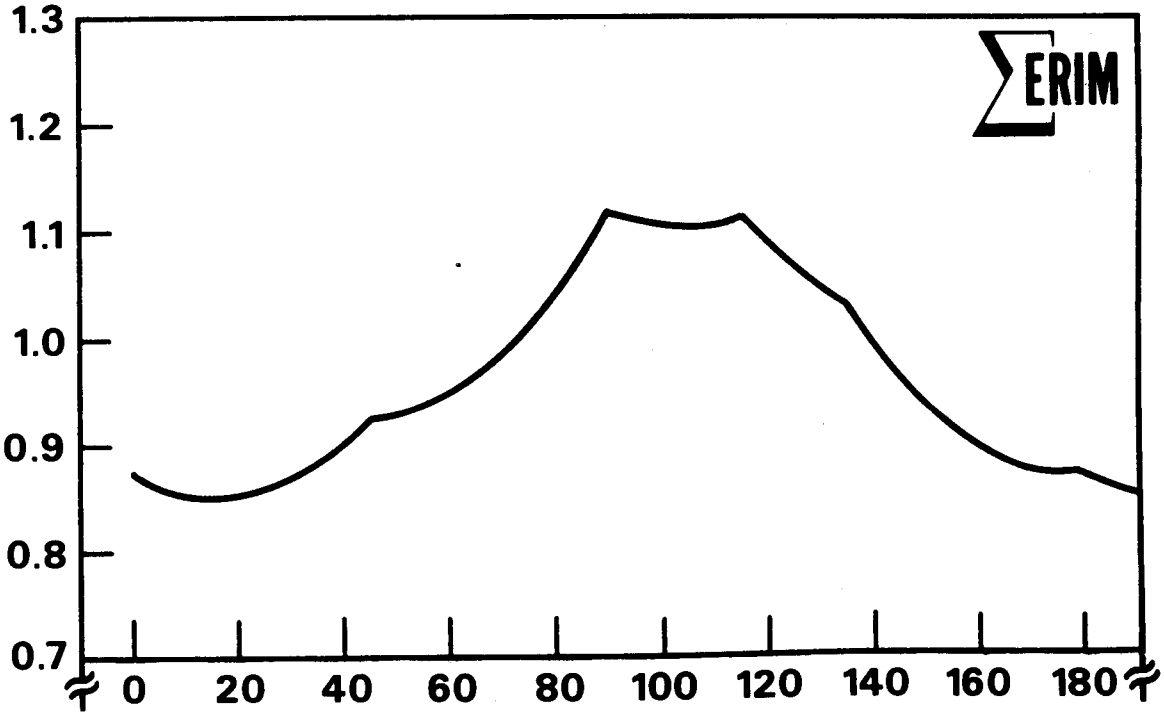

Figure 2. Calculated Recognition Accuracy for Different Projection Angles for Example with Large Variance

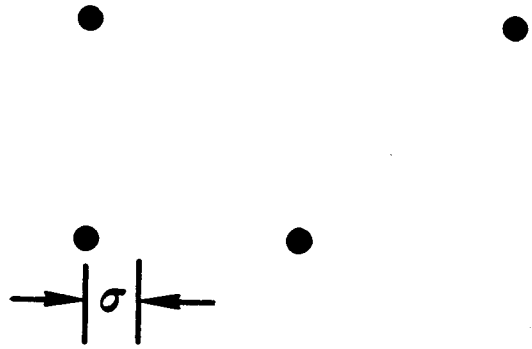

Figure 3. Location of Means for Example with Small Variance

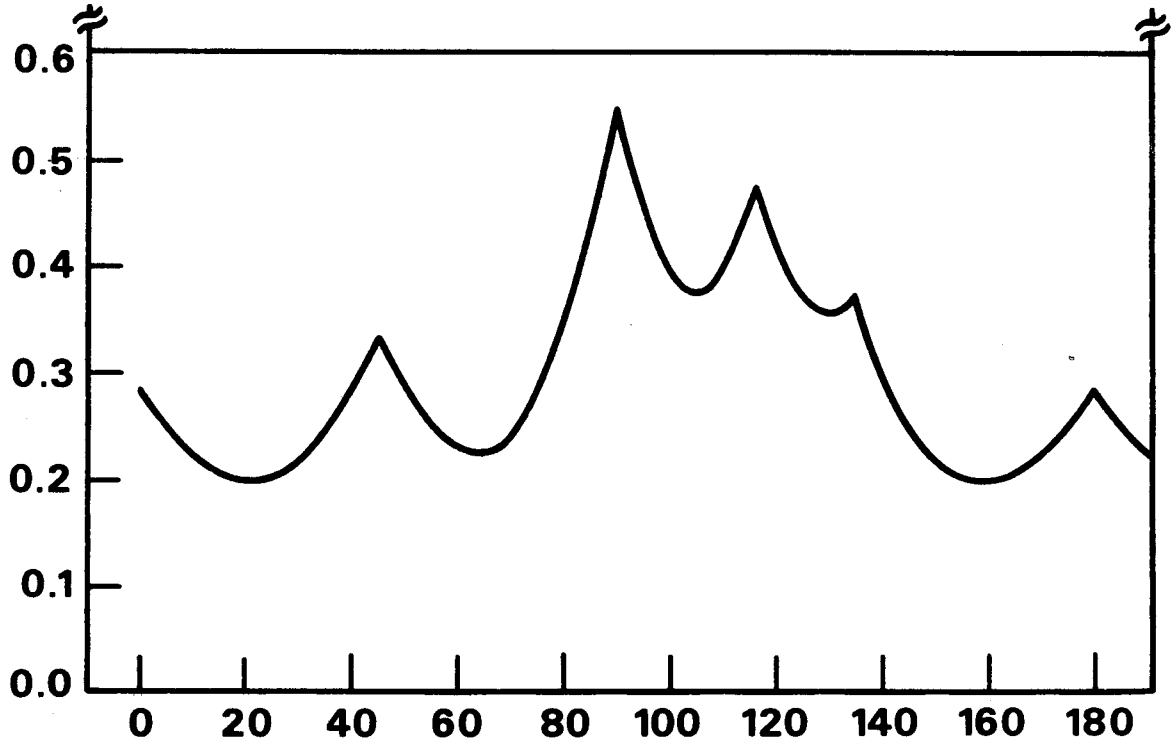

Figure 4. Calculated Recognition Accuracy for Different Projection Angles for Example with Small Variance

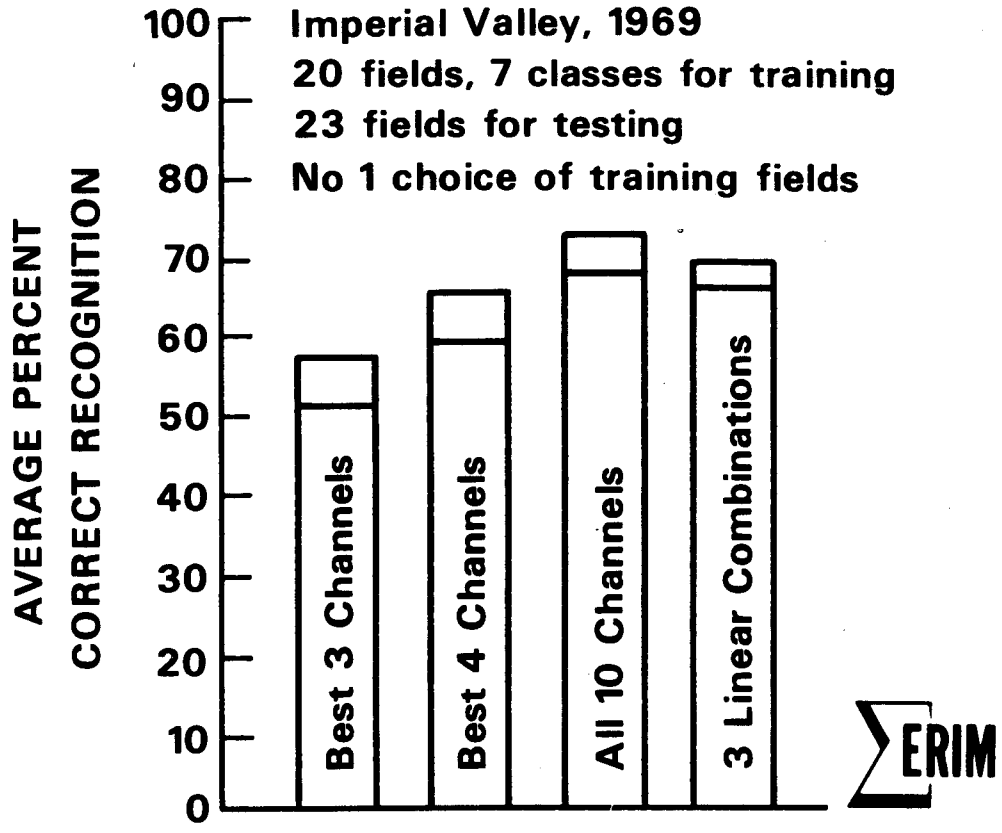

Figure 5. Comparison of Linear Combinations with Subsets of Channels Test Fields

	LINEAR COMBINATIONS				SUBSETS		
	No. 1	No.2	No. 3		3	4	10
Without Soil	66	66	64		51	59	68
With Soil	67	70	69		58	65	73
Predicted	71	75			51	62	70

Figure 6. Percentage Recognition Accuracy Obtained by Using Linear Combinations, Subsets of Channels, and by Analytic Prediction

1	.29	.45	-.74	-.42	-.06	.07	0	0	-.20
0	.23	.65	-.02	.60	.76	.82	1	0	-.01
0	.18	.57	-.34	.20	.31	.11	0	1	-.09

Figure 7. Matrix Describing 3 Linear Channels When the Starting Point is the Best Subset of 3 Channels