An Iterative Clustering Procedure

ROBERT M. HARALICK, MEMBER, IEEE, AND ITS'HAK DINSTEIN, STUDENT MEMBER, IEEE

Abstract—In many remote sensing applications millions of measurements can be made from a satellite at one time, and many times the data is of marginal value. In these situations clustering techniques might save much data transmission without loss of information since cluster codes may be transmitted instead of multidimensional data points. Data points within a cluster are highly similar so that interpretation of the cluster code can be meaningfully made on the basis of knowing what sort of data point is typical of those in the cluster. We introduce an iterative clustering technique; the procedure suboptimally minimizes the probability of differences between the binary reconstructions from the cluster codes and the original binary data.

The iterative clustering technique was programmed for the GE 635 KANDIDATS (Kansas Digital Image Data System) and tested on two data sets. The first was a multi-image set. Twelve images of the northern part of Yellowstone Park were taken by the Michigan scanner system, and the images were reduced and run with the program. Thirty-thousand data points, each consisting of a binary vector of 25 components, were clustered into four clusters. The percentage difference between the components of the reconstructed binary data and the original binary data was 20 percent.

The second data set consisted of measurements of the frequency content of the signals from lightning discharges. One hundred and thirty-four data measurements, each consisting of a binary vector of 32 components, were clustered into four clusters. The ground truth divides the elements into two classes: cloud-to-ground and cloud-to-cloud discharges. The best performance of a trained classifier provided 82 percent of correct classification. Associating two clusters with each class yielded 74.5 percent of correct classification.

I. Introduction

A. Clustering

CLUSTERING, when properly used (it has been used as long ago as 1939 by Tryon [53]) and interpreted, can provide meaningful information about a data set. Clustering is often used as a tool to help inform the researcher where the "action" in his data set lies. A cluster is usually thought to be a subset of data points which are highly similar or associated and relatively unassociated with data points outside the subset [11]. With this interpretation, data points within a cluster are viewed as measurements which have been made of some one kind of environmental object or process. Hence clusters may inform the researcher what measurements come from distinct environmental objects or processes, that is, distinct as seen through the eyes of the instrument used to measure the environment.

Manuscript received May 11, 1970; revised January 30, 1971. This work was supported by NASA under Grant NGR 17-004-019.

The authors are with the Center for Research, Inc., and the Department of Electrical Engineering, University of Kansas, Lawrence, Kans. 66044.

B. Literature Review

There are a number of proposed clustering techniques currently available. Widely used in numerical taxonomy [45] are agglomerative and divisive hierarchical clustering schemes. Here, a small and fixed set of patterns is given and a matrix is computed whose (i,j)th entry is the association or similarity between the ith and jth patterns. The measure of association or similarity can be the correlation coefficient, Yules Q, or some inverse metric function. The agglomerative procedures [26] generally link together the most similar patterns. Then the similarities between the groups of linked patterns and the remaining groups (or patterns) are recomputed using the minimum, maximum or mean similarity between the two groups. The procedure continues in this manner linking together the most similar patterns or groups. Michener and Sokal [33], Ward [54], and McQuitty [29] were early users of this scheme.

The divisive procedures [26] begin with all the patterns in the same group and split the group into the two most dissimilar groups. Splitting, as proposed by Edwards and Cavalli-Sforza [7], can be done by examining all possible partitions of two parts for each group and selecting the partition of that group which reduces the within group variance the most. Lance and Williams [25] suggest successively splitting the groups by thresholding that variable in a way which is expected to reduce the variance the greatest for the split groups. Mattson and Dammann [30] suggest successively splitting each group by thresholding the dominant eigenvector of the covariance matrix for that group. Wirth *et al.* [55] suggest thresholding the association or similarity matrix and defining the components of the resulting graph as the clusters. Thresholding is done successively from strict thresholds to more liberal thresholds.

Nonhierarchical schemes have also been used in clustering a fixed small set of patterns. Most popular among these have been those iterative schemes beginning with an arbitrary set of exhaustive and mutually exclusive clusters and successively improving the set of clusters by transferring patterns from one cluster to another until no further improvement is available [10]. In the ISODATA technique [2] each pattern is put into the cluster for which the squared distance between it and the cluster mean is least. Then the new cluster means are computed and the whole procedure repeated. Jones and Jackson [22] suggested an iterative technique where clusters are found one at a time. An initial pattern is picked to be the first pattern in the cluster. Patterns are successively transferred into and out of the cluster in a way which increases the within cluster similarities and decreases in-cluster to out-cluster similarities. Another type of nonhierarchical scheme [4] starts out by thresholding the association or similarity matrix and defining as "core clusters" the maximal complete subgraphs (cliques) of the resulting graph. Then the smaller core clusters are merged into the large core clusters, and largely overlapping core clusters are merged.

When the number of patterns is not a small and fixed set, all of the preceding methods seem unfeasible since they each involve too many calculations. In the case of a large number of patterns another approach has been developed. Here, the clustering problem is conceived of in a different way. It is assumed that the patterns are generated by a number of different "sources" according to some unique source probability distribution. The probability distribution for the collection of patterns is then a mixture of the probability distributions of the sources. The clustering problem is then concerned with decomposing the mixture by identifying from the mixture distribution the individual probability distributions of the sources and then constructing a minimum risk Bayes decision rule to assign any pattern to the most probable source. The decision rule, of course, determines a partition, and the cells of the partition are the clusters. Work on the identifiability of mixture distributions has been done by Teicher [50]–[52], Yakowitz [57], [59], Yakowitz and Spragins [58], and Stanat [48]. Application and development of this technique under the name "learning without a teacher" or "unsupervised learning" has been done mainly by Fralick [9], Spragins [47], Patrick and Hancock [36], Patrick [37], Hilborn and Lainiotis [17], and Patrick and Costello [38].

Ruspini [43] has an interesting article on the abstract formulation of the clustering problem.

In this paper we will view a cluster as a subset of highly associated or similar data points, not necessarily requiring the clusters to be unassociated. From this perspective clustering serves as a dimensionality reduction technique. Clusters are formed parametrically from a sample of patterns iteratively improving a cluster assignment function of simple form.

When no more improvement is possible for the cluster assignment function, the entire set of patterns may be quickly clustered using it. An additional feature of this method is the concomitant definition of a simple inverse clustering function which assigns each cluster to some data point representative of the cluster. Hence a data point assigned to a cluster may be reconstructed from the cluster code. A test made of this clustering algorithm with some remote sensor multispectral scanner imagery indicated that it correctly reconstructed 79.3 percent of the binary data components after a 25 to 3 dimensionality reduction. A map is also illustrated indicating the correspondence of the clusters with natural terrain type categories.

C. Development of Iterative Clustering

In our discussion we will be concerned with clustering techniques which are applicable to data sets in which it is appropriate to ignore the order of the data points. For these cases the entire structure of the data is described by the triple (D,P,ρ), where D is a countable set of all possible measurements sometimes referred to as measurement space, P is the empirically observed probability distribution on D, and ρ is a metric on D. A set of M clusters for (D,P,ρ) is defined by any pair $(\mathscr{C},D^*)$, where $\mathscr{C} = \{C_1,C_2,\cdots,C_M\}$ is a set of M subsets of D which cover D, $\bigcup_{i=1}^{M} C_i = D$, and $D^* = \{d_1{}^*,d_2{}^*,\cdots,d_M{}^*\}$ is a subset of distinct elements of D, $D^* \subset D$, such that

$$\sum_{i=1}^{M} \sum_{d \in C_i} \rho(d,d_i^*)P(d)$$

is minimal for all possible sets $\mathscr{C}$ and D^*.

Usually, it is also required that the subsets C_i be mutually exclusive $C_i \cap C_j = \phi$, for $i \neq j$, so that $\mathscr{C}$ is a partition of D. A data point d_i^* in D^* is interpreted as being the most typical point in C_i, and d_i^* is therefore the most representative measurement of the kind of environmental object or process associated with cluster C_i.

The reader might note that measurement d_i^* was not defined to be a member of cluster C_i; however, when $P(d_i^*) > 0$, for every $d_i^* \in D^*$, this fact follows from the minimality of

$$\sum_{i=1}^{M} \sum_{d \in C_i} \rho(d,d_i^*)P(d).$$

(When $P(d_i^*) = 0$ for some $d_i^* \in D^*$, it is of no consequence into which cluster d_i^* is put; putting d_i^* in cluster C_i is intuitively most pleasing, so without loss of generality we assume it there when $P(d_i^*) = 0$.)

In many remote sensing applications millions of measurements can be made from a satellite at one time, and sometimes most of the data is not even worth transmitting. In these situations it is useful to cluster the data first, and then instead of transmitting the data point d, one transmits a code assigned to the cluster to which d belongs. Transmitting a data point d from a multispectral sensor usually involves a choice of one out of at least a million (d, for example, can be a six-tuple where each component is a choice of one out of ten), while transmitting a cluster code usually involves a choice of one out of ten. Then, there can be a 20 to 1 bit compaction achieved by transmitting cluster codes instead of data points. If before or after the transmission of cluster codes an inverse clustering function is transmitted which assigns each data point to some cluster code, then the received cluster codes can be interpreted as having come from measurements of an environmental object or process of which the data point specified by the transmitted inverse clustering function must be a typical example.

Let D be the set of possible data points, and $\mathscr{C} = \{C_1,\cdots,C_M\}$ be the set of clusters. The clustering function $f, f\colon D \to \mathscr{C}$, is defined by

$$f(d) = C_i, \qquad \text{if and only if } d \in C_i.$$

Let $D^* = \{d_1^*,\cdots,d_M^*\}$ be the set of representative data points for the cluster set $\mathscr{C}$. We define the inverse clustering function $g, g\colon \mathscr{C} \to D$, by $g(C_i) = d_i^*$. With this choice of g, by the definition of the clustering function f we must have that

$$\sum_{i=1}^{M} \sum_{d \in C_i} P(d)\rho(d,d_i^*) = \sum_{d \in D} P(d)\rho(d,(g \circ f)(d))$$

is minimal for all $\mathscr{C}$ and D^*, or equivalently, minimal for all functions $f, f\colon D \to \mathscr{C}$, and $g, g\colon \mathscr{C} \to D$.

This later perspective can offer another definition of a set of clusters. Let D be the set of all possible data points, P be a probability distribution on D, ρ be a metric on D, and $C = \{c_1,\cdots,c_M\}$ be a set of cluster codes. Let f and g be functions, $f\colon D \to C$ and $g\colon C \to D$, such that $\sum_{d \in D} P(d)\rho(d,(g \circ f)(d))$ is minimized. The cluster set which partitions D is $\{f^{-1}(c_1),f^{-1}(c_2),\cdots,f^{-1}(c_M)\}$, and the representative data point for cluster $f^{-1}(c_i)$ is $g(c_i)$. The function f is called a clustering function, and the function g is called an inverse clustering function.

There are some interesting facts about f and g which emerge from this perspective: 1) the function f is onto; 2) all the images of g are distinct; 3) if $g(c) = d$, then $f(d) = c$; 4) $f \circ g$ is the identity function on C; and 5) $g \circ f$ is "closest" to the identity function on D.

A formidable problem is to find functions g and f which minimize $\sum_{d \in D} P(d)\rho(d,(g \circ f)(d))$. In fact, besides exhaustive search, there is no known way to find optimal functions g and f. Because of this we propose an iterative suboptimal way for a restricted class of functions defined parametrically. First, we will assume that measurement space D is an N-dimensional binary space. This assumption is really not very restrictive since any real data may be appropriately coded to a binary form.

Initially, the functions f and g are chosen arbitrarily from a class of parametric functions. They are iteratively improved, monotonically decreasing $\sum_{d \in D} P(d)\rho(d,(g \circ f)(d))$ by making small changes in the function parameters on the basis of previous performance. Because the functions f and g are required to be in a restricted class of parametric functions, the minimum found is not guaranteed to be the global minimum. In fact we shall see that the algorithm is suboptimal in the sense that there is no guarantee that a local minimum within the class of parametric functions will be found.

The following notation convention will be used throughout the paper. Let $D = \{-1,1\}^N$ and $C = \{-1,1\}^K$, for $K < N$. For any $d \in D$, we will write

$$d = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_N \end{pmatrix}$$

and, for any $c \in C$, we will write

$$c = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_K \end{pmatrix}.$$

For any real number a, we define sgn (a) by

$$\text{sgn}\,(a) = \begin{cases} +1, & \text{if and only if } a > 0 \\ -1, & \text{otherwise.} \end{cases}$$

For any real N vector

$$r = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{pmatrix}$$

we define sgn (r) by

$$\operatorname{sgn}(r) = \begin{pmatrix} \operatorname{sgn}(r_1) \\ \operatorname{sgn}(r_2) \\ \vdots \\ \operatorname{sgn}(r_N) \end{pmatrix}.$$

Let Q be a $K \times N$ real matrix and T be a $N \times K$ real matrix:

$$Q = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1N} \\ q_{21} & q_{22} & \cdots & q_{2N} \\ \vdots & & & \vdots \\ q_{K1} & q_{K2} & \cdots & q_{KN} \end{pmatrix}$$

$$T = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1K} \\ t_{21} & t_{22} & \cdots & t_{2K} \\ \vdots & & & \vdots \\ t_{N1} & t_{N2} & \cdots & t_{NK} \end{pmatrix}.$$

We define $(Qd)_k = \sum_{i=1}^{N} q_{ki}\delta_i$ and $(Tc)_k = \sum_{i=1}^{K} t_{ki}\gamma_i$. We will restrict our attention to functions $f: D \to C$ and $g: C \to D$, which have the parametric form

$$f(d) = \operatorname{sgn} \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1N} \\ q_{21} & q_{22} & \cdots & q_{2N} \\ \vdots & & & \vdots \\ q_{K1} & q_{K2} & \cdots & q_{KN} \end{pmatrix} \cdot \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_N \end{pmatrix} = c$$

$$g(c) = \operatorname{sgn} \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1K} \\ t_{21} & t_{22} & \cdots & t_{2K} \\ \vdots & & & \vdots \\ t_{N1} & t_{N2} & \cdots & t_{NK} \end{pmatrix} \cdot \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_K \end{pmatrix} = \hat{d}.$$

(If we change the form of g by leaving out the sgn, there is the possibility of extending the method to nonbinary data. Perhaps this can be the topic of a future paper.) By expressing f and g parametrically, we can change the functions f and g a little by perturbing the parameters in the Q and T matrices a little. Because their implementation is easy, linear threshold functions of this sort have been used frequently in supervised pattern recognition learning systems [1], [3], [5], [6], [12], [13], [18], [19], [21], [23], [24], [31], [32], [34], [35], [39]–[41], [49], [56]. Their use here in an unsupervised or clustering mode [14], [15] provides an additional application.

The essence of the iterative clustering algorithm may be seen in the following perspective. Once the Q and T matrices are chosen, we can calculate $\sum_{d \in D} P(d)\rho(d,(g \circ f)(d))$. We can also predict changes in the summation due to small changes in some elements in the Q or T matrices. We will apply only those changes that cause the summation to decrease. We continue the process until any additional small changes in some elements of Q or T, but not both, cause an increase in the summation. At this point, f and g suboptimally minimize $\sum_{d \in D} P(d)\rho(d,(g \circ f)(d))$.

The iteration process is not unique. For any pair of functions f and g arbitrarily chosen, there are usually many ways to improve the Q and the T matrices. Since we cannot consider all the possible changes, it is not guaranteed that the final f and g are local minima of $\sum_{d \in D} P(d)\rho(d,(g \circ f)(d))$.

II. Iterative Clustering Procedure

A. Definition

The iterative clustering procedure is described by its three major components:

1) the clustering function by which it assigns data points to clusters and the inverse clustering function by which it defines a typical or representative data point for each cluster;
2) the performance criterion by which it evaluates any pair of clustering and inverse clustering functions; and
3) the adjustment procedure by which it changes and improves the clustering and inverse clustering function on the basis of their performance.

B. Clustering and Inverse Clustering Function

In our case the clustering function is defined by $f(d) = \operatorname{sgn}(Qd)$; $f(d)$ is interpreted as the binary $(+1,-1)$ code or label for the cluster to which d is assigned. The inverse clustering function is defined by $g(c) = \operatorname{sgn}(Tc)$; $g(c)$ is interpreted as the representative data point typical of the cluster whose label is c.

C. Performance Criterion

The performance criterion for the pair of clustering and inverse clustering functions f and g is

$$\sum_{d \in D} P(d)\rho(d,(g \circ f)(d)) = \sum_{d \in D} P(d)\rho(d, \operatorname{sgn}(T \ \operatorname{sgn} Qd))$$

where ρ is a metric specifying the number of components in which its arguments disagree, and P is a probability distribution on $D = \{-1,+1\}^N$.

D. Adjustment Procedure

The idea of perturbing the Q and T matrices a little is a simple enough idea concerning how to adjust. The problem is how much to adjust and which way to adjust. We solve the problem of how much to adjust by an analysis of "significant change". We solve the problem of which way to adjust by a complete analysis of the effect of a change (in terms of the performance criterion) on any one element of the Q or T matrices.

1) Significant Changes or How Much to Adjust: The binary nature of the signum function makes the f and g functions respond to changes in the parameters of the Q and T matrices in a stepwise manner. Such stepwise behavior implies that changes in the Q or T matrices do not necessarily lead to changes in the f and g functions. We define a significant change in an element of the Q or T matrix as one which leads to a change in the f or g function for some element in D or C, respectively. A significant amount for some parameter is the magnitude of any perturbation on that parameter which produces a significant change.

It is not hard to show that perturbations below a magnitude called the smallest significant amount cannot produce a change. The smallest significant amount by which an element of the ith row of the Q or T matrix must be perturbed in order to produce a change in f or g is just greater than

$$\min_{d \in D} \left| \sum_{j=1}^{N} q_{ij} \cdot \delta_j \right|, \qquad \text{for } f$$

$$\min_{c \in C} \left| \sum_{j=1}^{K} t_{ij} \cdot \gamma_j \right|, \qquad \text{for } g$$

for $P(d) > 0$. It is also not hard to show that when the magnitude of a perturbation on a parameter is increased beyond a certain amount, called the largest significant amount, no additional changes in the f or g function can be obtained. Perturbing a parameter by any larger amount than the largest significant amount causes exactly the same change in f or g as perturbing by the largest significant amount. The largest significant amount for an element in the ith row of the Q or T matrix is

$$\max_{d \in D} \left| \sum_{j=1}^{N} q_{ij} \cdot \delta_j \right|, \qquad \text{for } f$$

$$\max_{c \in C} \left| \sum_{j=1}^{K} t_{ij} \cdot \gamma_j \right|, \qquad \text{for } g$$

for $P(d) > 0$.

The number of distinct significant changes in some parameter of the Q and T matrices is finite bounded by the number of nonzero probability elements in D and C, respectively. However, because changes in a parameter above or below certain amounts produce no further changes in f or g and because f and g have a stepwise behavior under changes of their parameters, the number of distinct significant changes can be expected to be much smaller than the number of elements in D.

2) *Effect of Change or Which Way to Adjust:* We know, now, how to get significant changes in the functions f and g. Each significant change in f or g can cause a change in the performance criterion $\sum_{d \in D} P(d)\rho(d,(g \circ f)(d))$. We are interested in such significant changes of f and g which produce a better performance index. In order to keep track of changes in the performance index we must now define the following sets, functions, and parameters.

ε_{kj}: Consider the ordering of the distinct elements in the sequence $\langle |Qd|_k \mid d \in D \rangle$.[1] Set ε_{kj} equal to just more than the jth element in this ordering; ε_{kj} is therefore the jth largest significant amount for any element in the kth row of the Q matrix.

θ_{nj}: Similarly, consider the ordering of the distinct elements in the sequence $\langle |T \operatorname{sgn}(Qd)|_n \mid d \in D \rangle$. Set θ_{nj} to be just more than the jth element in this ordering. θ_{nj} is therefore the jth largest significant amount for any element in the nth row of the T matrix.

A_n: This is the set of all elements in D whose nth component is different from the nth component of its image under $g \circ f$:

$$A_n = \left\{ d = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_n \end{pmatrix} \in D \mid \delta_n \neq (\operatorname{sgn}(T \operatorname{sgn}(Qd)))_n \right\}, \qquad n = 1,2,\cdots,N.$$

G_n: This is the set of elements in D whose nth component is $+1$:

$$G_n = \left\{ d = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_n \end{pmatrix} \in D \mid \delta_n = +1 \right\}, \qquad n = 1,2,\cdots,N.$$

F_k: This is the set of all elements in D whose image under f has its kth component $+1$:

$$F_k = \{d \in D \mid (Qd)_k > 0\}, \qquad k = 1,2,\cdots,K.$$

D_{kj}^+: This is the set of all elements in D whose image under f has the kth component $+1$ and which would be affected by a significant change of an element in the kth row of the Q matrix by applying the jth largest significant amount:

$$D_{kj}^+ = \{d \in D \mid 0 < (Qd)_k < \varepsilon_{kj}\}, \qquad k = 1,2,\cdots,K, \qquad j = 1,2,\cdots,J.$$

D_{kj}^-: This is the set of all elements in D whose image under f has the kth component -1 and which would be affected by a significant change of an element in the kth row of the Q matrix by applying the jth largest significant amount:

$$D_{kj}^- = \{d \in D \mid -\varepsilon_{kj} < (Qd)_k \leq 0\}, \qquad k = 1,2,\cdots,K, \qquad j = 1,2,\cdots,J.$$

E_{nj}^+: This is the set of all elements in D whose image under $g \circ f$ has the nth component $+1$ and which would be affected by a significant change of an element in the nth row of the T matrix by applying the jth largest significant amount:

$$E_{nj}^+ = \{d \in D \mid 0 < (T \operatorname{sgn}(Qd))_n < \theta_{nj}\}, \qquad n = 1,2,\cdots,N, \qquad j = 1,2,\cdots,J.$$

E_{nj}^-: This is the set of all elements in D whose image under $g \circ f$ has the nth component -1 and which would be affected by a significant change of an element in the nth row of the T matrix by applying the jth largest significant amount:

$$E_{nj}^- = \{d \in D \mid -\theta_{nj} < (T \operatorname{sgn}(Qd))_n \leq 0\}, \qquad n = 1,2,\cdots,N, \qquad j = 1,2,\cdots,J.$$

H_{nk}^+: This is the set of all elements d in D for which 1) its image under f is $+1$, and 2) a significant change of an element in the kth row of the Q matrix changes its image

[1] The number $|Qd|_k$ is the absolute value of the kth component of Qd, that is,

$$|Qd|_k = \left| \sum_{n=1}^{N} q_{kn}\delta_n \right|.$$

under $g \circ f$:

$$H_{nk}^{+} = \left\{ d = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_n \end{pmatrix} \in D \mid \operatorname{sgn}\left((T \operatorname{sgn}(Qd))_n - 2t_{nk}\right) \neq \delta_n \text{ and } (Qd)_n > 0 \right\},$$

$$n = 1,2,\cdots,N, \qquad k = 1,2,\cdots,K.$$

H_{nk}^{-}: This is the set of all elements d in D for which 1) its image under f is -1, and 2) a significant change of an element in the kth row of the Q matrix changes its image under $g \circ f$:

$$H_{nk}^{-} = \left\{ d = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_n \end{pmatrix} \in D \mid \operatorname{sgn}\left((T \operatorname{sgn}(Qd))_n + 2t_{nk}\right) \neq \delta_n \text{ and } (Qd)_n \leq 0 \right\},$$

$$n = 1,2,\cdots,N, \qquad k = 1,2,\cdots,K.$$

Finally, for any set S we denote the complement of S by S^c. Figs. 1 and 2 illustrate Venn diagrams of these sets.

Define the metric ρ by

$$\rho(d,\tilde{d}) = \sum_{i=1}^{N} \frac{|\delta_i - \tilde{\delta}_i|}{2}.$$

The metric ρ counts the number of components of d and $\tilde{d}$ which disagree. For this definition of ρ, minimizing $\sum_{d \in D} P(d) \cdot \rho(d,\hat{d})$, where $\hat{d} = \operatorname{sgn}(T \operatorname{sgn}(Qd))$, is equivalent to minimizing $\sum_{n=1}^{N} P(A_n)$ since A_n is the set of all elements in D for which $\delta_n \neq \operatorname{sgn}(T \operatorname{sgn}(Qd))_n$ and

$$\rho(d,\hat{d}) = \begin{cases} m, & \text{if } \delta_n \neq \operatorname{sgn}(T \operatorname{sgn}(Qd))_n, \\ & \quad \text{for } m \text{ components and all} \\ & \quad \text{other components agree} \\ 0, & \text{if } \delta_n = \operatorname{sgn}(T \operatorname{sgn}(Qd))_n, \\ & \quad \text{for every component.} \end{cases}$$

Any significant change in the Q or the T matrices moves some elements d of D into A_n, for some n, and some other elements d of D out of A_n, for some n. The change is beneficial if and only if $\sum_{n=1}^{N} P(A_n)$ is decreased due to the change. In order to be able to decide whether a change is beneficial or not, we must determine the net number of elements which are moved out of each A_n due to the change. It is not hard to show that the only elements in D which are affected by significant changes of the jth largest significant amount are some of those in the sets D_{kj}^{+}, D_{kj}^{-}, E_{nj}^{+}, and E_{nj}^{-}.

Consider significant changes in the Q matrix. There are four cases to consider depending on whether an element belongs to D_{kj}^{+} or D_{kj}^{-}, and whether we change q_{kn} by adding or subtracting ε_{kj}. For each of these four cases we must answer three questions.

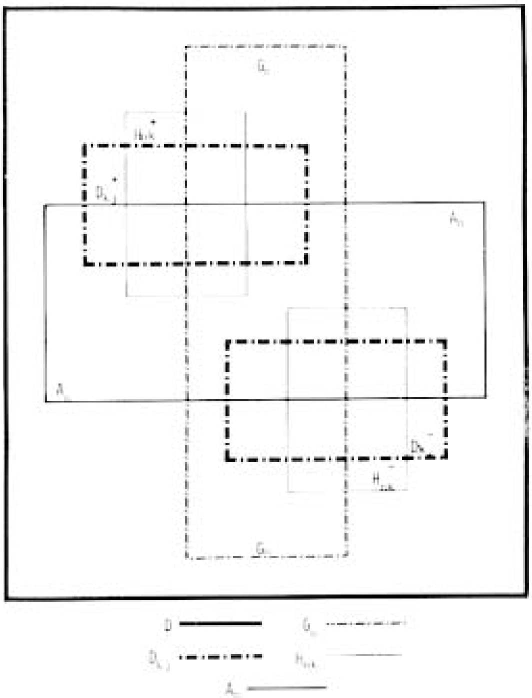

Fig. 1. Venn diagram for sets A_n, G_n, D_{kj}^{+}, D_{kj}^{-}, H_{nk}^{+}, and H_{nk}^{-}.

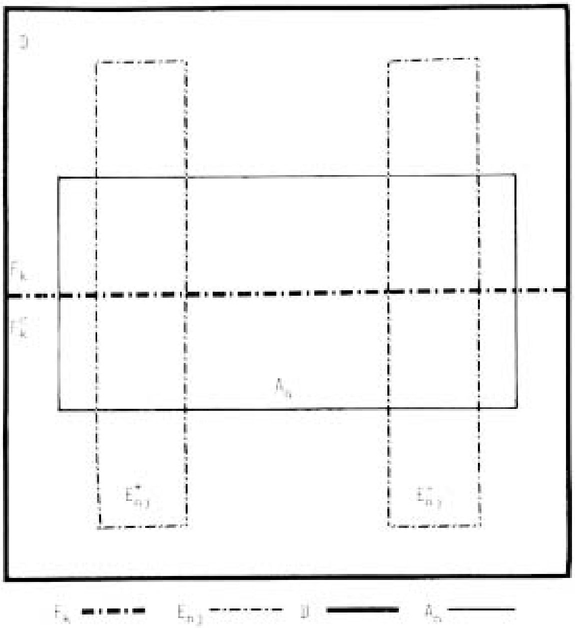

Fig. 2. Venn diagram for sets A_n, F_k, E_{nj}^{+}, and E_{nj}^{-}.

a) For which elements is the change in q_{kn} going to produce a change in $(Qd)_k$?

b) Of the elements for which the change in q_{kn} produces a change in $(Qd)_k$, for which of them is the change going to produce a change in $\operatorname{sgn}(T \operatorname{sgn}(Qd)) = \hat{d}$?

c) What is the net change in $\sum_{n=1}^{N} P(A_n)$ caused by the change in q_{kn}?

We will discuss the question for one case in an outline form and give the results of the other three cases.

Case 1: Consider the case in which we add ε_{kj} to $q_{kn}(q_{kn} \leftarrow q_{kn} + \varepsilon_{kj})$ and $d \in D_{kj}^+$.

a) Since $d \in D_{kj}^+$, we must have that $0 < (Qd)_k < \varepsilon_{kj}$. Adding ε_{kj} to q_{kn} will increase $(Qd)_k$ if $\delta_n = +1$, and decrease $(Qd)_k$ if $\delta_n = -1$. Hence by adding ε_{kj} to q_{kn}, $(Qd)_k$ can be changed from positive to negative for all those elements whose nth component is -1. These elements must be members of $D_{kj}^+ \cap G_n^c$.

b) The elements in D whose $\operatorname{sgn}(T \operatorname{sgn}(Qd))_n$ is affected by a change in $(Qd)_k$ going from positive to negative must belong to H_{nk}^+. Hence the elements in D whose $\operatorname{sgn}(T \operatorname{sgn}(Qd))_n$ is affected by adding ε_{nj} to q_{kn} must belong to $H_{nk}^+ \cap D_{kj}^+ \cap G_n^c$.

c) Of the elements in D for which $\operatorname{sgn}(T \operatorname{sgn}(Qd))_n$ is affected by adding ε_{kj} to q_{kn}, some elements are in A_n, and these will move out of A_n; others are not in A_n, and these will move into A_n. The elements which are moved out of A_n are members of $D_{kj}^+ \cap G_n^c \cap H_{nk}^+ \cap A_n$. The elements which are moved into A_n are members of $D_{kj}^+ \cap G_n^c \cap H_{nk}^+ \cap A_n^c$. Since this change may affect all components of $\operatorname{sgn}(T \operatorname{sgn}(Qd))$, the net beneficial effect of the change can be expressed by

$$\sum_{i=1}^{N} [P(D_{kj}^+ \cap G_n^c \cap H_{ik}^+ \cap A_i) - P(D_{kj}^+ \cap G_n^c \cap H_{ik}^+ \cap A_i^c)].$$

By the same argument, the net beneficial effects of the remaining three cases are found to be as follows.

Case 2: For adding ε_{kj} to $q_{nk}(q_{nk} \leftarrow q_{nk} + \varepsilon_{kj})$ and $d \in D_{kj}^-$ the net beneficial effect is

$$\sum_{i=1}^{N} [P(D_{kj}^- \cap G_n \cap H_{ik}^- \cap A_i) - P(D_{kj}^- \cap G_n \cap H_{ik}^- \cap A_i^c)].$$

Case 3: For subtracting ε_{kj} from $q_{nk}(q_{nk} \leftarrow q_{nk} - \varepsilon_{kj})$ and $d \in D_{kj}^+$ the net beneficial effect is

$$\sum_{i=1}^{N} [P(D_{kj}^+ \cap G_n \cap H_{ik}^+ \cap A_i) - P(D_{kj}^+ \cap G_n \cap H_{ik}^+ \cap A_i^c)].$$

Case 4: For subtracting ε_{kj} from $q_{nk}(q_{nk} \leftarrow q_{nk} - \varepsilon_{kj})$ and $d \in D_{kj}^-$ the net beneficial effect is

$$\sum_{i=1}^{N} [P(D_{kj}^- \cap G_n^c \cap H_{ik}^- \cap A_i) - P(D_{kj}^- \cap G_n^c \cap H_{ik}^- \cap A_i^c)].$$

Fig. 3 illustrates a Venn diagram of the possible transition of elements to and from various sets due to adjustment $(q_{nk} \leftarrow q_{nk} + \varepsilon_{kj})$.

Consider, now, significant changes in the T matrix. There are four cases to consider depending on whether an element belongs to E_{nj}^+ or E_{nj}^-, and whether we change t_{nk} by adding or subtracting θ_{nj}. For each of these cases we must answer two questions.

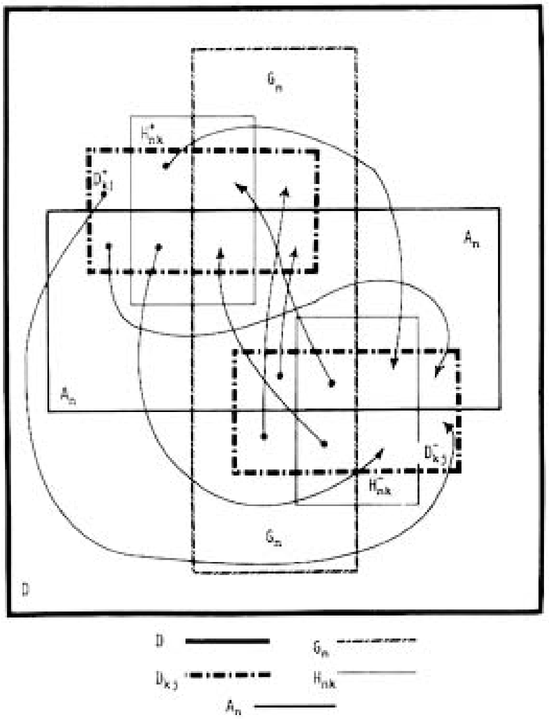

Fig. 3. Some transitions of data elements from and into sets D_{kj}^+ and D_{kj}^- due to reinforcement $q_{nk} \leftarrow q_{nk} + \varepsilon_{kj}$.

d) For which of the elements is the change in t_{nk} going to produce a change in $\operatorname{sgn}(T \operatorname{sgn}(Qd))_n$?

e) What is the net change in A_n caused by the change in t_{nk}? (Note that a change in t_{nk} affects only A_n.)

We will discuss the questions for one case in outline form and give the results for the other three cases.

Case 5: Consider the case in which we add θ_{nj} to t_{nk}, $t_{nk} \leftarrow t_{nk} + \theta_{nj}$, and $d \in E_{nj}^+$.

d) Since $d \in E_{nj}^+$, we must have that $0 < \operatorname{sgn}(T \operatorname{sgn}(Qd))_n < \theta_{nj}$ so that $\operatorname{sgn}(T \operatorname{sgn}(Qd))_n = +1$. Adding θ_{nj} to t_{nk} will increase $(T \operatorname{sgn}(Qd))_n$ if $(Qd)_k > 0$, and decrease it if $(Qd)_k < 0$. Hence by adding θ_{nj} to t_{nk}, $\operatorname{sgn}(T \operatorname{sgn}(Qd))_n$ can be changed from $+1$ to -1 for all the d whose $(Qd)_k$ is negative. Such elements must be in $E_{nj}^+ \cap F_k^c$.

e) Of the elements in D whose $\operatorname{sgn}(T \operatorname{sgn}(Qd))_n$ is affected by adding θ_{nj} to t_{nk}, some elements are in A_n, and these will move out of A_n; others are not in A_n, and these will move in A_n. The elements which are moved out of A_n are members of $E_{nj}^+ \cap F_k^c \cap A_n$. The elements which are moved into A_n are members of $E_{nj}^+ \cap F_k^c \cap A_n^c$. The net beneficial effect of the change can be expressed by

$$P(E_{nj}^+ \cap F_k^c \cap A_n) - P(E_{nj}^+ \cap F_k^c \cap A_n^c).$$

By the same argument, the net beneficial effect of the other three cases are found to be as follows.

Case 6: For adding θ_{nj} to t_{nk}, $t_{nk} \leftarrow t_{nk} + \theta_{nj}$, and $d \in E_{nj}^-$ the net beneficial effect is

$$P(E_{nj}^- \cap F_k \cap A_n) - P(E_{nj}^- \cap F_k \cap A_n^c).$$

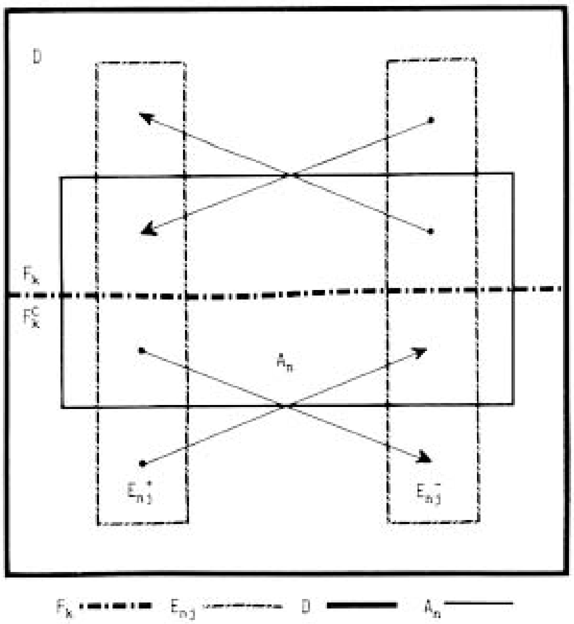

Fig. 4. Transitions of data elements from and into set A_n due to reinforcement $t_{nk} \leftarrow t_{nk} + \theta_{nj}$.

Case 7: For subtracting θ_{nj} from t_{nk}, $t_{nk} \leftarrow t_{nk} - \theta_{nj}$, and $d \in E_{nj}^+$ the net beneficial effect is

$$P(E_{nj}^+ \cap F_k \cap A_n) - P(E_{nj}^+ \cap F_k \cap A_n^c).$$

Case 8: For subtracting θ_{nj} from t_{nk}, $t_{nk} \leftarrow t_{nk} - \theta_{nj}$, and $d \in E_{nj}^-$ the net beneficial effect is

$$P(E_{nj}^- \cap F_k^c \cap A_n) - P(E_{nj}^+ \cap F_k^c \cap A_n^c).$$

Fig. 4 illustrates a Venn diagram of the possible transitions of elements to and from various sets due to the adjustment $t_{nk} \leftarrow t_{nk} + \theta_{nj}$.

Consider now the four possible types of adjustments to some parameters in the Q or T matrices:

$$q_{kn} \leftarrow q_{kn} + \varepsilon_{kj} \tag{1}$$

$$q_{kn} \leftarrow q_{kn} - \varepsilon_{kj} \tag{2}$$

$$t_{nk} \leftarrow t_{nk} + \theta_{nj} \tag{3}$$

$$t_{nk} \leftarrow t_{nk} - \theta_{nj}. \tag{4}$$

We will define improvement numbers so we can compare the benefit due to the four different possibilities of significant changes. The first possibility of significant change is:

$$\begin{aligned} Q_{knj}^{\text{plus}} = \sum_{i=1}^{N} & P\{A_i \cap [(D_{kj}^+ \cap H_{ik}^+ \cap G_n^c) \\ & \cup (D_{kj}^- \cap H_{ik}^- \cap G_n)]\} \\ & - P\{A_n^c \cap [D_{kj}^+ \cap H_{ik}^+ \cap G_n^c) \\ & \cup (D_{kj}^- \cap H_{ik}^- \cap G_n)]\} \end{aligned} \tag{1'}$$

where Q_{knj}^{plus} is the net increase in the probability that $\delta_n = \text{sgn}\,(T\,\text{sgn}\,(Qd))_n$ after the element q_{kn} has been increased by ε_{kj}. The second is

$$\begin{aligned} Q_{knj}^{\text{minus}} = \sum_{i=1}^{N} & P\{A_i \cap [(D_{kj}^+ \cap G_n \cap H_{ik}^+) \\ & \cup (D_{kj}^- \cap G_n^c \cap H_{ik}^-)]\} \\ & - P\{A_i^c \cap [(D_{kj}^+ \cap G_n \cap H_{ik}^+) \\ & \cup (D_{kj}^- \cap G_n^c \cap H_{ik}^-)]\} \end{aligned} \tag{2'}$$

where Q_{knj}^{minus} is the net increase in the probability that $\delta_n = \text{sgn}\,(T\,\text{sgn}\,(Qd))_n$ after the element q_{nk} has been decreased by ε_{kj}. The third is

$$\begin{aligned} T_{nkj}^{\text{plus}} = {} & P\{A_n \cap [(E_{nj}^+ \cap F_k^c) \cup (E_{nj}^- \cap F_k)]\} \\ & - P\{A_n^c \cap [(E_{nj}^+ \cap F_k^c) \cup (E_{nj}^- \cap F_k)]\} \end{aligned} \tag{3'}$$

where T_{nkj}^{plus} is the net increase in the probability that $\delta_n = \text{sgn}\,(T\,\text{sgn}\,(Qd))_n$ after the element t_{nk} has been increased by θ_{nj}. The fourth is

$$\begin{aligned} T_{nkj}^{\text{minus}} = {} & P\{A_n \cap [(E_{nj}^+ \cap F_k) \cup (E_{nj}^- \cap F_k^c)]\} \\ & - P\{A_n^c \cap [(E_{nj}^+ \cap F_k) \cup (E_{nj}^- \cap F_k^c)]\} \end{aligned} \tag{4'}$$

where T_{nkj}^{minus} is the net increase in the probability that $\delta_n = \text{sgn}\,(T\,\text{sgn}\,(Qd))_n$ after the element t_{nk} has been decreased by θ_{nj}.

The improvement numbers are either positive, negative, or zero. A negative improvement number indicates an increase in $\sum_{n=1}^{N} P(A_n)$ due to the respective change. A positive improvement number indicates a decrease in $\sum_{n=1}^{N} P(A_n)$ due to the respective change. A beneficial change in the Q or the T matrix is possible only if at least one of the improvement numbers is positive.

We compute the improvement numbers by counting the elements in the proper sets as defined in the preceding. We apply the smallest significant change which is beneficial. If there is no benefit for any change, then the Q and T matrices cannot be further iteratively improved by this algorithm.

III. Results

In order to test the iterative clustering algorithm it was implemented in a computer program for the GE-635 KANDIDATS,[2] and it was run with a multispectral set of images and an 8-dimensional lightning discharges data set.

A. Multi-Image Set

Twelve images taken by the Michigan scanner system were the multi-image data set. These images were taken of a 2 by 6 mi area in the northern part of Yellowstone Park at approximate coordinates 100°30′ by 44°57′ on September 19, 1967. Fig. 5 shows an old panchromatic photograph of

[2] KANDIDATS (Kansas Digital Image Data System) is a high-efficiency package of image processing programs developed by the Center for Research, Inc., University of Kansas. It is specially designed to operate on the GE-635 computer with minimum time and cost.

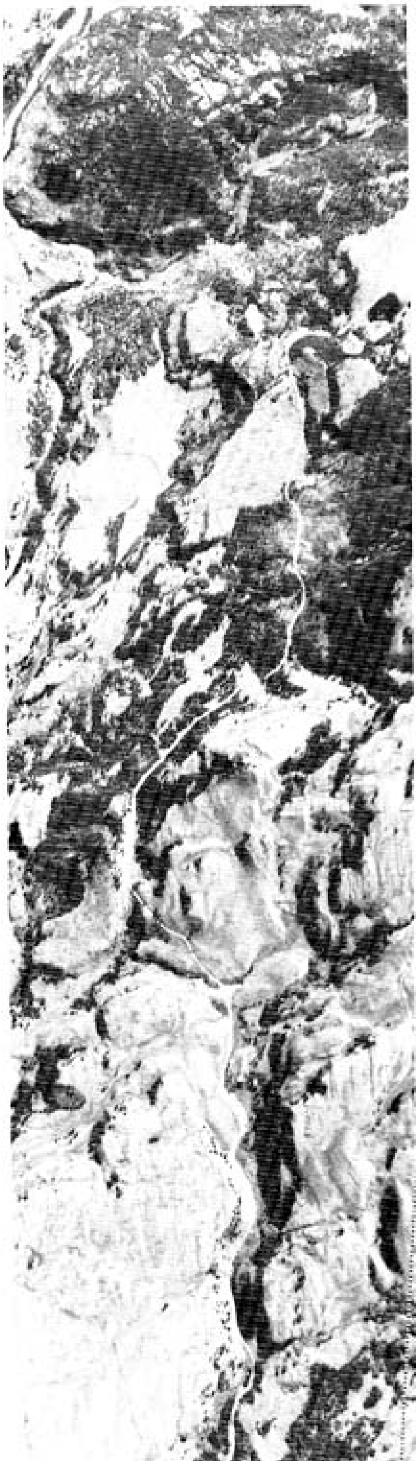

Fig. 5. Panchromatic imagery—Yellowstone area.

TABLE I
Tabulation of Wavelength Band for Images

Image	Wavelength Band (mμ)
1	800–1000
2	720–800
3	660–720
4	620–660
5	580–620
6	550–580
7	520–550
8	500–520
9	480–500
10	460–480
11	440–460
12	400–440

the area taken in 1954. Each image of the multi-image set was, in effect, a picture taken with a different narrow-band filter, where the filters passed light in narrow bandwidths from the near infrared band part of the spectrum through the ultraviolet portion of the spectrum. Table I tabulates these bands. The images were digitized to 256 levels on a grid of 220 by 1260 for a total of about 270 000 resolution cells for each image. Each resolution cell contains the returns from 12 spectral bands coming from a 20 by 20 ft small-area ground patch. Successive resolution cells contain returns from small-area ground patches separated by a gap of 20 ft.

In order to reduce computer time the original 12 images were preprocessed to yield 4 smaller images, but with most of the statistical and spatial structure preserved. The first part of the preprocessing consisted of a principal components analysis [20]. A principal components analysis may be considered in the following fashion. In any image some grey tones occur more frequently than others. We may consider the relative frequency of the grey tones in the image as defining a 1-dimensional probability distribution. This probability distribution has a mean and variance. Similarly, in a 12-dimensional multi-image set each resolution cell has a 12-dimensional vector of grey tones. Some of these grey tone vectors occur more frequently than others, and we may consider the relative frequency of the grey tone vectors as defining a 12-dimensional probability distribution. This probability distribution has a mean with a covariance matrix. The first principal component is that eigenvector of the covariance matrix having the largest eigenvalue. The kth principal component is the eigenvector of the covariance matrix having the kth largest eigenvalue. We will define the kth principal component image as the projection of the multi-image on the 1-dimensional subspace spanned by the kth principal component. Because the sum of variances of the 1-dimensional probability distributions determined by the principal component images equals the total variance of the original 12-dimensional probability distribution, the ratio of the variance of the kth principal component to the total variance is called the variance accounted for by the kth principal component. The variance accounted for by the kth principal component is an indicator

TABLE II
WEIGHTS USED TO OBTAIN LINEAR COMBINATIONS FOR FIRST, SECOND, AND THIRD PRINCIPAL COMPONENTS

	Principal Component		
	First	Second	Third
	Percent of Variance Accounted for		
Wavelength Band	97.4	1.6	0.9
800–1000	0.15702	−0.33153	0.27651
720–800	0.22758	−0.33529	0.26796
660–720	0.28342	−0.31332	0.13188
620–660	0.23184	−0.23019	0.11530
580–620	0.20199	−0.16154	0.010509
550–580	0.17486	−0.094258	0.025696
520–550	0.32559	−0.070442	0.25499
500–520	0.24756	0.031677	0.046381
480–500	0.42727	0.09246	−0.12730
460–480	0.52815	0.22196	−0.60118
440–460	0.25916	0.30966	0.043766
400–440	0.14878	0.65714	0.61121

of how much statistical structure from the original 12 images is preserved by the kth principal component image.

The principal components analysis provided the following results. It was determined that the first component accounted for 97.4 percent of the variance, the second component 1.6 percent of the variance, and the third component 0.9 percent of the variance. The respective weights used in the linear combination are listed in Table II, and the three principal component images are illustrated in Fig. 6.

Table II has an interesting interpretation. The first linear combination has weights which are all positive and which are about the same magnitude. The first principal component image is then very close to what a panchromatic image of the area would be. This should not be surprising since most photo interpreters will prefer a panchromatic image over any narrow-band image because they see more structure in it. The second linear combination weighs the infrared part of the spectrum negatively, the middle of the spectrum hardly at all, and the ultra-violet part of the spectrum positively. This weighting trend from the infrared to the ultra-violet is almost a linear one. It is indicative of the fact that the spectral reflectance curve for most natural objects shows that when infrared reflectance is high, then the ultra-violet reflectance is low, and when the ultra-violet reflectance is high, then the infrared reflectance is low. Hence the weighting done by the second principal component will enhance the difference between those objects with high-infrared and high-ultra-violet reflectance. The third linear combination is perhaps indicative of the spectral reflectance difference between vegetation and rock. The spectral reflectance curve for most vegetation slopes positively at the ultra-violet end of the spectrum, while that for ryolite (a volcanic rock known to be prevalent in the area photographed) is almost flat in that very same region. Hence weighing the 460–480-mμ part of the spectrum negatively and the 400–440-mμ end of the spectrum positively will enhance the difference between vegetation and ryolite. Since the first three principal components accounted for about 99 percent of the variance, there is no need to cluster 12 images. Almost all the information is contained in three linear combinations (principal components) of the original 12.

The next step in the preprocessing consisted of reducing the size of the three principal component images. Each image was 220 resolution cells horizontally by 1200 resolution cells vertically. Each was reduced in size to 73 resolution cells horizontally by 406 resolution cells vertically by taking every third row of resolution cells and every third resolution cell on each such row taken.

The final step in the preprocessing consisted of quantizing the grey tones of the three images. This was done in two parts: the images were first quantized to 64 grey tone classes by a folded-tail linear quantizing procedure and then quantized to 13, 8, and 7 grey tone classes for the first, second, and third principal component images, respectively, by a spatial quantizing procedure. The folded-tail quantizing is essentially a linear quantizing procedure modified to ignore extreme wild points on the tails of the distribution. In other words instead of determining the highest grey tone and the smallest grey tone and then equally dividing the resulting interval into 64 pieces as the linear quantizing does, the folded-tail linear quantizing determines a "high" grey tone less than the highest and a "small" grey tone greater than the smallest and equally divides the resulting interval up into 64 quantized classes. Of course grey tones higher than the determined "high" and smaller than the determined "small" get put in the highest and smallest quantized class, respectively.

The spatial quantizing procedure further reduces the number of quantized classes from 64 to 13 in a way which capitalizes on the spatial dependence of the 64 quantized grey tone classes. A 64 × 64 Markov transition probability matrix is set up where the (i,j)th element is the probability that a resolution cell having a grey tone in the ith grey tone class will be next to a resolution cell having a grey tone in the jth grey tone class. The spatial quantizing uses the information in this matrix to form quantized classes whose grey tones have a high probability of occurring next to each other. Hence spatial contiguity of grey tones tends to be preserved. These quantizing procedures are fully described in [6].

After quantizing, each resolution cell of the multi-image is characterized by a triplet of quantized grey tones. This triplet must be converted to a binary $(-1,+1)$ N-tuple so that it may be used with the iterative clustering algorithm. Since the metric chosen for the clustering was the number of components in which the binary N-tuples disagree, the binary code for the quantized grey tone triplet cannot be any arbitrary binary code. It must be a code which under the metric preserves the ordering structure of triplet. In other words a disagreement of only one component between two binary N-tuples should arise if and only if there is a difference of only one in one of the components for the corresponding quantized grey-tone triplets. One simple binary code which preserves the ordering structure is the following code having 25 binary components. The first k rightmost components of the rightmost 12 components

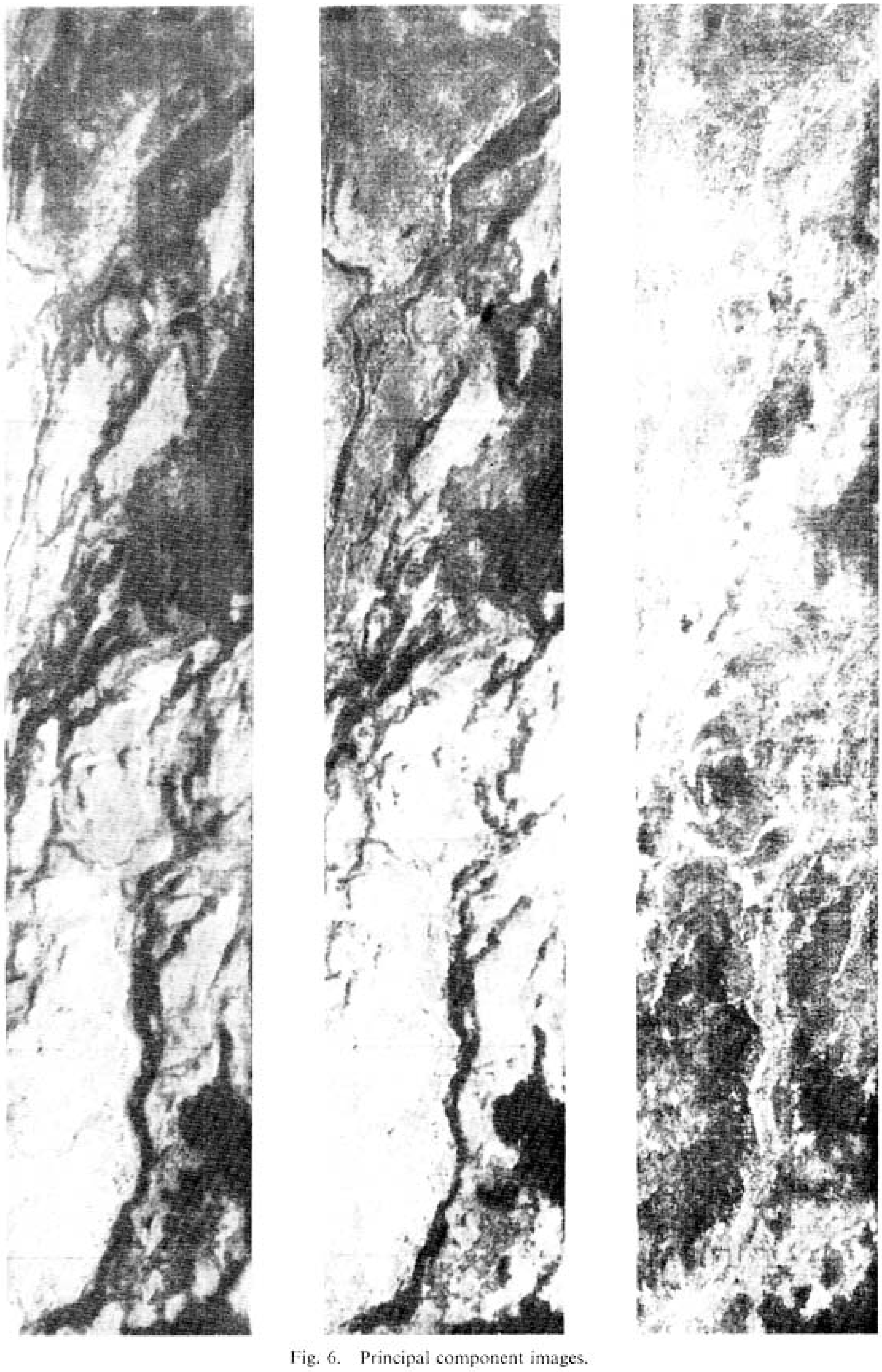

Fig. 6. Principal component images.

are one if and only if the first principal component lies in the kth quantized grey tone class. The first k rightmost components of the middle 7 components are one if and only if the second principal component lies in the kth quantized grey tone class. The first k rightmost components of the leftmost 6 components are one if and only if the third principal component lies in the kth quantized grey tone class. All components which are not one are minus one.

Of the 30 000 resolution cells in the multi-image, a random sample of 1908 was chosen. From this sample data sequence $S_D = \langle d_1, d_2, \cdots, d_{1908} \rangle$ the empirical probability distribution P was defined on $D = \{-1,+1\}^{25}$ by $P(d) = \#\{i \mid d = d_i\}/1908$, where $P(d)$ is the relative frequency of the element d in the sequence S_D. The set C was defined by $C = \{-1,+1\}^3$ so that a maximum of eight clusters were possible. The sample was then clustered by the iterative clustering algorithm. The initial Q and T matrices were chosen by considering each row of the Q and T matrices as a vector and sampling each row from a probability distribution uniform for any direction and dependent only on direction. The final resulting Q matrix was used to define the cluster assignment function $f(d) = \text{sgn}\ (T\ \text{sgn}\ (Qd))$.

Each data element in the sample had 25 components, so the total number of sample components was $1908 \cdot 25 = 47\,700$. The initial Q and T matrices produced 21 708 reconstructed components which did not agree with the original ones. Then the signs of the elements in the T matrix were changed in every row i for which the number of elements in the sample in which $\delta_n \neq (\text{sgn}\ (T\ \text{sgn}\ (Qd)))_n$ was more than half of the elements in the sample. At that point, the number of disagreeing components was reduced to 12 754. Then, by the iterative process, the number was reduced to 8486 after 103 iterations. This means that 83 percent of the components of the training set were correctly reconstructed by the final Q and T matrices through the transformation sgn $(T\ \text{sgn}\ (Qd))$. As a check on this 83 percent estimate, the full data set of 30 044 data elements was applied to transformation sgn $(T\ \text{sgn}\ (Qd))$. A total of 595 512 of the reconstructed components out of 751 100 possible agreed for a result 79.3-percent correct.

In another experiment the initial Q and T matrices were constructed such that the rows of the Q matrix and the columns of the T matrix consisted of the three principal components of the binary coded data. The initial Q and T matrices produced 9035 reconstructed components which did not agree with the original ones. Then, by the iterative process, the number was reduced to 6108 after 59 iterations. This means that 41 592 components, which are 87.3 percent of the 47 700 sampled components, were reconstructed correctly. As a check on this 87.3 percent estimate, the full set of 30 044 data elements was applied to transformation sgn $(T\ \text{sgn}\ (Q))$. A total of 602 933 of the reconstructed components out of 751 100 possible agreed for a result 80-percent correct.

Fig. 7 illustrates a map of the clusters obtained by the assignment of the measurement in each resolution cell to a cluster by the cluster assignment function f. Of the eight clusters possible, only four major clusters were determined;

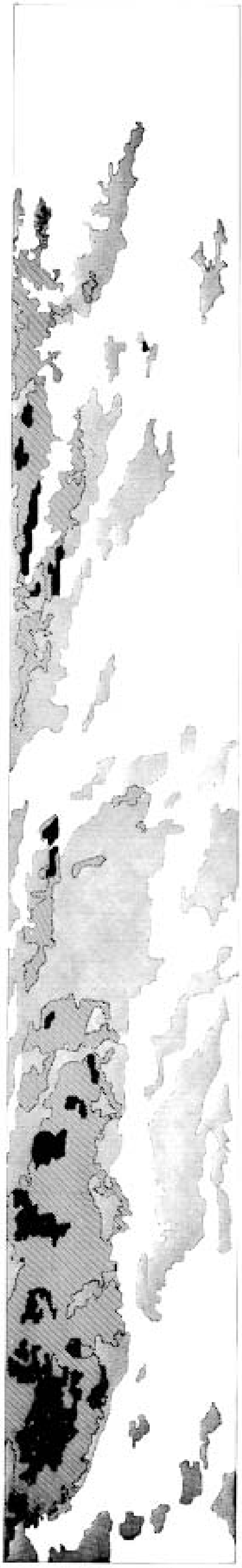

Fig. 7. Map of clusters.

the remaining clusters occupied such a small percentage of the image that they need not be considered. Using the ground truth supplied by Smedes *et al.* [47], a correspondence was made between the clusters and the true ground categories. The white area corresponds to bog, forest, or cloud shadow. The bogs are moist areas with a lush growth of sedges and grasses. The forests are lodgepole forests and Douglas fir. The cross-hatched area corresponds to glacial kame. These are meadows underlain by sand and gravel and vegetated by grass and sagebrush. The diagonally striped area is glacial till. These are grassland and sagebrush meadows underlain by glacial till. Mixtures of silty to bouldery mineral soil is exposed over one-fifth of the area. The black area is bedrock exposure. It consists of mainly unvegetated bare bedrock exposed by glacial and steam erosion.

B. Lightning Data Set

The lightning discharge data were taken during two thunderstorms in Oklahoma during June 1968. A discrimination procedure based on frequency content of the measurements was developed by Shanmugam and Breipohl [44] for classifying the discharges into cloud-to-ground or cloud-to-cloud type. The measurements were taken by airborne instrumentation, and the ground truth was observed and recorded.

Preprocessing of the measurements done by Shanmugam and Breipohl [44] yielded 8-dimensional data elements consisting of the following components.

x_1 Relative value of signal in 10-kHz horizontal.
x_2 Relative value of signal in 50-kHz horizontal.
x_3 Relative value of signal in 150-kHz horizontal.
x_4 Relative value of signal in 250-kHz horizontal.
x_5 Relative value of signal in 10-kHz vertical.
x_6 Relative value of signal in 50-kHz vertical.
x_7 Relative value of signal in 250-kHz vertical.
x_8 Relative value of signal in 250-kHz vertical.

The data set consisted of 134 8-dimensional measurements. The components of the measurements were quantized into five equally probable levels. Each component was then coded into a binary quadruplet in a way that preserved distances. The codes are shown in the following.

Level	Code
1	0000
2	0001
3	0011
4	0111
5	1111

Hence each 8-dimensional data element was coded into a binary vector of 32 components. The clustering procedure was applied to the 134 data elements. The set C was defined by $C = \{-1, +1\}^2$, and four clusters were formed. The initial Q and T matrices were constructed by principal component analysis. The rows of the initial Q matrix consisted of the eigenvectors corresponding to the two largest eigenvalues of the covariance matrix of the binary vectors. The initial T matrix was set by the initial Q transposed. The total number of components was $32 \cdot 134 = 4288$. The initial Q and T matrices produced 1289 reconstructed components which did not agree with the original ones. Then, by the iterative process, the number of disagreeing components was reduced to 1239. Hence 3049 components, which are 71 percent of the components, were reconstructed correctly. The codes of the four clusters were $(-1, -1)$, $(-1, +1)$, $(+1, -1)$, and $(+1, +1)$. Associating $(-1, -1)$ and $(+1, -1)$ with cloud-to-ground discharge and $(-1, +1)$ and $(+1, +1)$ with cloud-to-cloud discharge showed that 100 of the discharges (74.5 percent) were clustered into groups corresponding to the given ground truth. Shannugam and Breipohl [44] reported that a trained classifier provided 82-percent correct identification.

We should note that in the lightning discharges data set the iterative procedure did not significantly reduce the number of disagreeing components from the initial principal components starting point. This raises the interesting possibility of perhaps concentrating effort on the most appropriate binary coding technique and letting the principal components do the clustering.

IV. Conclusions

In conclusion we have developed an iterative clustering technique using two parametric functions, the clustering function f, and the inverse clustering function g, which were defined by

$$f(d) = \operatorname{sgn}(Qd) = c, \qquad g(c) = \operatorname{sgn}(Tc) = d$$

where Q is a $K \times N$ real matrix and T is a $N \times K$ real matrix, where N and K are the dimensions of the binary data elements and the cluster codes, respectively. The clusters were constructed by assigning a cluster code to each measurement by the clustering function f. The data was reconstructed by applying the inverse clustering function g to the cluster codes. The rows of the Q matrix and the columns of the T matrix consisted of eigenvectors corresponding to the largest eigenvalues of the covariance matrix of the binary coded data. The iterative process took place by perturbing elements of the Q and T matrices in a way that suboptimally minimized the differences between the components of the reconstructed data elements and the original ones.

The technique was programmed for the GE 635 and tested on two data sets. The first was a multi-image set of data. For the training set of the multi-image set 1908 data elements of 25 binary components each were sampled from the multi-image and clustered into four major clusters. Of the reconstructed data components, 87.3 percent agreed with the original ones. For the prediction set 30 044 data elements of 25 binary components were used. Of the reconstructed data components, 80 percent agreed with the original ones. It is possible to construct the rows of the Q matrix and the columns of the T matrix by considering them as vectors sampled from a distribution, the directions of which are uniformally distributed. When starting this way, it takes more iterations to get results similar to those obtained by the principal component start (103 iterations compared to 59 iterations). The second data set was a lightning discharge data set. One hundred and thirty-four measurements were clustered into four clusters. Seventy-one percent of the 4288 reconstructed components agreed with the original ones. Associating the cluster codes $(-1,-1)$ and $(+1,-1)$ with cloud-to-ground discharges and $(-1,+1)$ and $(+1,+1)$ with cloud-to-cloud discharges showed that 100 measurements (74.5 percent) were correctly clustered according to available ground truth.

The success of this initial application of clustering to remote sensing data indicates that it might be possible to use clustering techniques to either screen imagery or preprocess imagery. In the screening function images which have many clusters would be transmitted back from distant satellite or spacecraft, while images with few clusters never have to be transmitted. In the preprocessing function only the maps resulting from clustering the multi-images can be transmitted back. In either use substantial data reduction is possible.

Although in this experiment the clusters seem to match the ground truth quite well, it should be possible to refine the technique and obtain more major clusters which would correspond to more specific types of ground categories by using multilayer linear threshold functions. We anticipate doing further work in this direction.

Acknowledgment

The authors would like to thank Dr. H. Smedes of the U.S. Geological Survey for providing the images taken by the Michigan scanner system and Dr. A. M. Breipohl of the University of Kansas for providing the lightning discharge data. All computer work was done at the University of Kansas Computer Center.

References

[1] A. Albert, "A mathematical theory of pattern recognition," *Ann. Math. Statist.*, vol. 34, 1963, pp. 284–299.

[2] G. H. Ball and D. J. Hall, "ISODATA, a novel method of data analysis and pattern classification," Stanford Res. Inst., Menlo Park, Calif., Tech. Rep., 1965.

[3] H. D. Block, B. W. Knight, and F. Rosenblatt, "Analysis of a four layer series coupled perceptron," *Rev. Mod. Phys.*, vol. 34, Jan. 1962, pp. 135–142.

[4] R. E. Bonner, "On some clustering techniques," *IBM J.*, vol. 8, Jan. 1969, pp. 22–32.

[5] L. Breiman and Z. Wurtele, "Convergence properties of a learning algorithm," *Ann. Math. Statist.*, vol. 35, 1964, pp. 1819–1822.

[6] R. O. Duda and H. Fossum, "Pattern classification by iteratively determined linear and piecewise linear discriminant functions," *IEEE Trans. Electron. Comput.*, vol. EC-15, Apr. 1966, pp. 220–232.

[7] A. W. F. Edwards and L. T. Cavalli-Sforza, "A method for cluster analysis," *Biometrics*, vol. 21, June 1965, pp. 362–375.

[8] W. D. Fisher, "On grouping for maximum homogeneity," *J. Amer. Statist. Ass.*, vol. 53, 1958, pp. 789–798.

[9] S. C. Fralick, "Learning to recognize patterns without a teacher," *IEEE Trans. Inform. Theory*, vol. IT-13, Jan. 1967, pp. 57–64.

[10] H. P. Friedman and J. Rubin, "On some invariant criteria for grouping data," *J. Amer. Statist. Ass.*, vol. 62, Dec. 1967, pp. 1159–1179.

[11] D. Gasking, "Clusters," *Australas. J. Phil.*, vol. 38, May 1960, pp. 2–35.

[12] H. Gluckesman, "On the improvement of a linear separation by extending the adaptive process with a stricter criterion," *IEEE Trans. Electron. Comput.* (Short Notes), vol. EC-15, Dec. 1966, pp. 941–944.

[13] H. J. Greenberg and A. G. Konheim, "Linear and nonlinear methods in pattern classification," *IBM J.*, vol. 8, July 1964, pp. 299–307.

[14] R. M. Haralick, "Adaptive pattern recognition of agriculture in western Kansas by using a predictive model in construction of similarity sets," in *Proc. 5th Symp. Remote Sensing of Environment*, Apr. 1967.

[15] R. M. Haralick and G. Darling, "Non-parametric unsupervised learning: ideas and results," in *2nd Hawaii Int. Conf. Systems Sciences*, Jan. 1968.

[16] R. M. Haralick, "Multi-image pattern recognition: ideas and results," Univ. of Kansas, Lawrence, CRES Tech. Rep. 133-11, Sept. 1969.

[17] C. G. Hilborn, Jr., and D. G. Lainiotis, "Optimal unsupervised learning multicategory dependent hypotheses pattern recognition," *IEEE Trans. Inform. Theory*, vol. IT-14, May 1968, pp. 468–470.

[18] J. H. Holland, "Outline for a logical theory of adaptive systems," *Ass. Comput. Mach. J.*, vol. 9, July 1962, pp. 297–314.

[19] W. S. Holmes and C. E. Phillips, "Experimental investigation of large multilayer linear discriminators," in *1967 Spring Joint Computer Conf., AFIPS Conf. Proc.*, vol. 30, pp. 265–271.

[20] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psych.*, vol. 24, 1933, pp. 417–441.

[21] E. R. Ide, C. E. Kiessling, and C. J. Tunis, "Some conclusions on the use of adaptive linear decision functions," in *1968 Fall Joint Computer Conf., AFIPS Conf. Proc.*, vol. 33, pp. 1117–1124.

[22] K. S. Jones and D. Jackson, "Current approaches to classification and clump-finding at the Cambridge Language Research Unit," *Comput. J.*, vol. 10, 1967, no. 1, pp. 29–37.

[23] J. S. Koford and G. F. Groner, "The use of an adaptive threshold element to design a linear optimal pattern classifier," *IEEE Trans. Inform. Theory*, vol. IT-12, Jan. 1966, pp. 42-50.

[24] H. Kazmierczak and K. Steinbuch, "Adaptive systems in pattern recognition," *IEEE Trans. Electron. Comput.*, vol. EC-12, Dec. 1963, pp. 822–835.

[25] G. N. Lance and W. T. Williams, "Computer programs for monothetic classification (association analysis)," *Comput. J.*, vol. 8, Oct. 1965, pp. 246–249.

[26] ——, "A general theory of classificatory sorting strategies, pt. 1: hierarchical systems," *Comput. J.*, vol. 9, Feb. 1967, pp. 373–380.

[27] ——, "Mixed-data classificating program II: divisive systems," *Aust. Comput. J.*, vol. 1, May 1968, pp. 82–85.

[28] A. H. Lochlan, "A note on the elementary α-perceptron," *Bull. Math. Biophys.*, vol. 26, 1964, pp. 45–57.

[29] L. L. McQuitty, "Hierarchical linkage analysis for the isolation of types," *Educ. Psychol. Measurement*, vol. 20, 1960, no. 1, pp. 55–67.

[30] R. L. Mattson and J. E. Dammahn, "A technique for determining and coding subclasses in pattern recognition problems," *IBM J. Res. Develop.*, July 1965, pp. 294–302.

[31] H. M. Martinez, "An introduction to the theory of adaptive pattern recognition," Office of Aerospace Research, U.S. Air Force Office of Scientific Research, Washington, D.C., AFOSR Sci. Rep. AF-AFOSR 370-63, AD 608 157, Oct. 1964, p. 62.

[32] ——, "A convergence theorem for linear threshold elements," *Bull. Math. Biophys.*, vol. 27, 1965, pp. 153–159.

[33] C. D. Michener and R. R. Sokal, "A quantitative approach to a problem in classification," *Evolution*, vol. 11, 1957, pp. 130–162.

[34] N. J. Nilsson, *Learning Machines: Foundations of Trainable Pattern Classifying Systems.* New York: McGraw-Hill, 1965.

[35] A. Novikoff, "On convergence proofs for perceptions," Stanford Res. Inst., Menlo Park, Calif., Tech. Rep., Jan. 1963, p. 12.

[36] E. A. Patrick and J. C. Hancock, "Nonsupervised sequential classification and recognition of patterns," *IEEE Trans. Inform. Theory*, vol. IT-12, July 1966, pp. 362–372.

[37] E. A. Patrick, "On a class of unsupervised estimation problems," *IEEE Trans. Inform. Theory*, vol. IT-14, May 1968, pp. 407–415.

[38] E. A. Patrick and J. P. Costello, "Unsupervised estimation and processing of unknown signals," Rome Air Development Center, Rome, N.Y., Tech. Rep. TR-69-430, Feb. 1970.

[39] J. D. Patterson and B. F. Womack, "An adaptive pattern classification system," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-2, Aug. 1966, pp. 62–67.

[40] L. Roberts, "Pattern recognition with an adaptive network," M.I.T. Lincoln Lab., Cambridge, Mass., Rep. 51-G-0013, AD 236 325, Apr. 1960.

[41] F. Rosenblatt, *Principles of Neurodynamics.* Spartan, Washington, D.C.: 1962.

[42] J. Rubin, "An approach to organizing data into homogeneous groups," *Syst. Zool.*, vol. 15, Sept. 1966, pp. 169–182.

[43] E. H. Ruspini, "A new approach to clustering," *Inform. Contr.*, vol. 15, 1969, pp. 22–32.

[44] K. Shanmugam and A. M. Breipohl, "Discriminating between cloud-to-ground and cloud-to-cloud lightning discharges," *J. Geophys. Res.*, (to be published).

[45] R. R. Sokal and P. H. A. Sneath, *Principles of Numerical Taxonomy*. San Francisco, Calif.: Freeman, 1963.

[46] H. W. Smedes, K. L. Pierce, M. G. Tannguay, and R. M. Hoffer, "Digital computer terrain mapping from multi-spectral data and evaluation of proposed earth resources technology satellite (ERTS) data channels, Yellowstone National Park," in *AIAA Earth Resources Observations and Information System Meeting*, Paper 70-309, Mar. 1970.

[47] J. Spragins, "Learning without a teacher," *IEEE Trans. Inform. Theory*," vol. IT-12, Apr. 1966, pp. 223–230.

[48] D. F. Stanat, "Unsupervised learning of mixtures of probability functions," in *Pattern Recognition*, L. Kanal, Ed. Washington, D.C.: Thompson, 1966, pp. 357–389.

[49] K. Steinbuch and U. A. W. Piske, "Learning matrices and their applications," *IEEE Trans. Electron. Comput.*, vol. EC-12, Dec. 1963, pp. 846–862.

[50] H. Teicher, "Identifiability of mixtures of product measures," *Ann. Math. Statist.*, vol. 38, 1967, pp. 1300–1302.

[51] ——, "Identifiability of finite mixtures," *Ann. Math. Statist.*, vol. 34, 1963, pp. 1265–1269.

[52] ——, "Identifiability of mixtures," *Ann. Math. Statist.*, vol. 32, 1961, pp. 244–248.

[53] R. C. Tryon, *Cluster Analysis*. Ann Arbor, Mich.: Edwards, 1939.

[54] J. H. Ward, "Hierarchical grouping to optimize an objective function," *J. Amer. Statist. Ass.*, vol. 58, Mar. 1963, pp. 236–245.

[55] M. Wirth, G. Estabrook, and D. Rogers, "A group theory model for systematic biology, with an example for the oncidunea," *Systematic Zoology*, vol. 15, no. 1, Mar. 1966, pp. 59–69.

[56] Z. Wurtele, "A problem in pattern recognition," *J. Soc. Ind. Appl. Math.*, vol. 13, Mar. 1965, pp. 60–67.

[57] S. J. Yakowitz, "Unsupervised learning and the identification of finite mixtures," *IEEE Trans. Inform. Theory*, vol. IT-16, May 1970, pp. 330–338.

[58] S. J. Yakowitz and J. D. Sprogins, "On the identifiability of finite mixtures," *Ann. Math. Statist.*, vol. 39, 1968, no. 1, pp. 209–214.

[59] S. J. Yakowitz, "A consistent estimator for the identification of finite mixtures," *Ann. Math. Statist.*, vol. 40, 1969, no. 5, pp. 1728–1735.

Robert M. Haralick (S'62–M'69) was born in Brooklyn, N. Y., on September 30, 1943. He received the B.A. degree in mathematics and the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Kansas, Lawrence, in 1964, 1966, 1967, and 1969, respectively.

He has worked for Autonetics and the IBM Corporation. In 1965 he joined the Center for Research, Inc., Engineering Science Division, University of Kansas. He is also an Assistant Professor of Electrical Engineering at the University of Kansas. He has done research in pattern recognition, adaptive self-organizing systems, multi-image processing techniques, texture analysis, clustering, and general systems theory.

Dr. Haralick is a member of Sigma Xi, Eta Kappa Nu, the Pattern Recognition Society, and the Society for General Systems Research.

Its'hak Dinstein (S'70) was born in Haifa, Israel, on January 15, 1939. He received the B.S.E.E. degree from the Technion, Israel Institute of Technology, Haifa, in 1965, and the M.S.E.E. degree from the University of Kansas, Lawrence, in 1969. He is presently a candidate for the Ph.D. degree at the University of Kansas.

From 1965 to 1967 he was employed by Elron-Elbit, Israel, where he was engaged in the development of digital instrumentation. In 1967–1968 he was a Teaching Assistant at the University of Kansas. He is currently with the Center for Research, Inc., University of Kansas, where his research is in the area of pattern recognition.