JOURNAL OF VISUAL COMMUNICATION AND IMAGE REPRESENTATION
Vol. 5, No. 2, June, pp. 127–138, 1994

Nonlinear Mapping Algorithm and Applications for Multidimensional Data Analysis

MARK J. CARLOTTO

TASC, 55 Walkers Brook Dr., Reading, Massachusetts 01867

Received January 10, 1992; accepted April 9, 1993

A new algorithm to assist in the analysis of data sets of very high dimensionality (from 10 to over 1000 dimensions) is described. The algorithm is based on a nonlinear mapping (NLM) algorithm developed by Sammon which maps a configuration of points in one space to a configuration in another such that the distances between points in the two spaces are approximately preserved. Sammon's algorithm is initially used to analyze multidimensional data from a brain mapping experiment. Because the complexity of his algorithm grows quadratically with the number of points, it is limited to relatively small data sets. An extended NLM algorithm is then described that is capable of handling large data sets (e.g., images) by using a multidimensional interpolation approach. A method for interpreting hyperspectral imagery data based on this extended algorithm is illustrated. Finally, the analysis of the structure and content of a collection of text documents using NLM is considered that involves the use of alternative distance measures and binary vectors of extremely high dimensionality (>1000). © 1994 Academic Press, Inc.

1. INTRODUCTION

We experience the world in three spatial dimensions plus time. However, in a number of scientific visualization and pattern recognition applications, data sets involving hundreds and perhaps thousands of dimensions are involved. In the pattern recognition community, linear and nonlinear transformations have been used to map multidimensional data into fewer dimensions for display and analysis (Andrews 1972). Meanwhile, scientific data visualization techniques have been developed to augment conventional spatial representations with icons and texture (Grinstein and Smith 1990), color and motion (Young and Rheingans 1990), sound (Coughran and Grosse 1990), and other perceptual modalities.

This paper explores the use of *nonlinear mapping* techniques in conjunction with conventional 2-D graphical representations for visualizing feature spaces of very high dimensionality, from tens to thousands of dimensions. As an example, consider two hyperspherical distributions in a five-dimensional space: (a) one hypersphere inside another, and (b) two hyperspheres separated in space. In Fig. 1 the two sets of points have been mapped into two-space using a nonlinear transformation. Even though the data are five-dimensional, the inherent two-dimensional structure of the configurations are apparent.

Our approach is based on the nonlinear mapping (NLM) algorithm developed by Sammon (1969) which transforms N-dimensional data into M-dimensional spaces. Sammon's algorithm maps a configuration of points from one space to another so that the differences in distance between all pairs of points in the two spaces are minimized. The effectiveness of the algorithm in attempting to preserve local and global relationships in multivariate data has been demonstrated extensively in the literature. Unfortunately, the complexity of the algorithm is quadratic in the number of points $O(K^2)$ and it is thus impractical for large data sets.

After other related nonlinear techniques are briefly summarized in Section 2, the key elements of Sammon's algorithm are reviewed in Section 3. A case study illustrating the use of NLM in visualizing the relationships between cross-sectional measurements derived from digitized brain slices is presented in Section 4. In Section 5, we then describe an extended NLM algorithm that involves mapping a subset ($K' \ll K$) of points using Sammon's original algorithm and interpolating the remaining ($K - K'$) points. The resultant complexity is almost linear in K for ($K' \ll K$) but is not limited to two dimensions as are previous algorithms. A second case study illustrating the use of our extended NLM for interpreting hyperspectral imagery data is presented in Section 6. In Section 7 alternative distance measures are explored in a third case study involving the visualization of text databases where texts are represented by binary vectors in excess of 1000 dimensions.

2. PREVIOUS WORK

Shepard (1962a, b) considered an early nonlinear mapping problem that involved determining the multidimensional structure of results from psychological experiments. Observations were based on a subjective ranking

1047-3203/94 $6.00

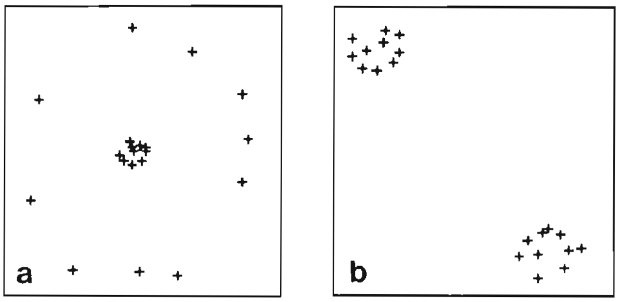


FIG. 1. Five-dimensional hyperspherical distributions.

of stimuli. Subjects were asked to rank the similarity between facial expressions, colors, and other perceptual stimuli (e.g., Face A is more like Face B than Face C, etc.). Shepard's approach was based on the concept of multidimensional scaling (Shepard 1962a, b). Initially K points, each representing an observation, are placed at the vertices of a regular simplex in a space of $K - 1$ dimensions (e.g., for $K = 3$ observations, three points are placed at the vertices of an equilateral triangle in 2-space). Next the points are moved in such a way as to reduce the dimensionality of the configuration while preserving the ordering relations between the corresponding stimuli. Earlier, Hammersley (1950) had observed that for points distributed within a hypersphere of fixed radius, as the number of dimensions is increased the variance of the interpoint distance approaches zero. Shepard therefore reasoned that it should be possible to reduce the dimensionality of a configuration by increasing the variance of the interpoint distances. This was done by increasing the distance between points that are far apart and decreasing the distance between points that are close together.

Bennett (1969) later applied the same approach to estimate the *intrinsic dimensionality* of a collection of signals where each signal is represented by a point in N-space. Bennett's algorithm increases the variance of the configuration as was done above but only preserves the ranking between neighboring points in the configuration where the size of the neighborhood is a parameter. The intrinsic dimensionality was then determined by the method of principal components.

In the above algorithms, the dimensionality of the output space depends on the intrinsic dimensionality of the data. Sammon (1969) developed an algorithm for mapping points into output spaces of arbitrary dimensionality. The mapping is accomplished iteratively and attempts to preserve the distances between all points in the configuration. Although the effectiveness of his NLM algorithm has been demonstrated extensively in the literature, the complexity of the algorithm is quadratic in the number of points $O(K^2)$ and is thus impractical for large data sets.

Alternatives to and extensions of NLM which reduce computation by preserving fewer interpoint distances (e.g., between nearest neighbors and one global reference point only) and/or by restricting the mapping to two dimensions (e.g., by triangulation) have been developed. Chang and Lee (1973) developed a modified NLM for mapping points into a plane which reduces computation by preserving fewer interpoint distances. Lee *et al.* (1977) describe a sequential $O(K)$ triangulation algorithm that maps points into 2-space preserving only $(2K - 3)$ distances. Biswas *et al.* (1981) describe a method that combines Sammon's algorithm with that of Lee's. In Section 5, we describe an extension of NLM that involves mapping a subset $(K' \ll K)$ of points using Sammon's original algorithm and interpolating the remaining $(K - K')$ points. The resultant complexity is almost linear for $(K' \ll K)$ but is not limited to two dimensions as are the above algorithms.

Kohonen's topology-preserving mapping (e.g., Kohonen 1988) is another method for mapping multidimensional data into spaces of given dimensionality. It is based on an array of P locally interconnected processing units, where M is the dimensionality of the array. The input consists of a set of K vectors, $\mathbf{x}_k$. The weight vector of the pth unit at time t is $\mathbf{w}_p(t)$. Input and weight vectors are N-dimensional. After the weight vectors are randomly initialized, for each input vector the closest weight vector is determined, and its value, along with those of neighboring units, is updated iteratively:

$$\mathbf{w}_p(t + 1) = \mathbf{w}_p(t) + \alpha[\mathbf{x}_k - \mathbf{w}_p(t)].$$

The array "self-organizes" in the sense that $\mathbf{x}_k$ vectors that are close to one another in N-space map to nearby processing units. A disadvantage of Kohonen's approach is that the computational complexity is on the order of the number of processing units, which may be much greater than the number of points to be mapped.

3. NONLINEAR MAPPING ALGORITHM

Given a configuration of points in $\mathfrak{R}^N$, Sammon's algorithm computes a configuration of points in $\mathfrak{R}^M$ such that the distances between points are approximately preserved. Let $\mathbf{x}_k = \{x_{kn}\}$ be the position of the kth point in N-space and let $\mathbf{y}_k = \{y_{km}\}$ be its position in M-space. The distances between points in the two spaces are $\nu_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$ and $\mu_{ij} = \|\mathbf{y}_i - \mathbf{y}_j\|$, respectively, where $\| \ \|$ is the Euclidean norm. The *mapping error* is a measure of how well a configuration of points in M-space matches the original configuration in N-space in terms of the differences between interpoint distances:

$$E = \frac{\sum_{i<j}^{K} \frac{[\nu_{ij} - \mu_{ij}]^2}{\nu_{ij}}}{\sum_{i<j}^{K} \nu_{ij}}.$$

The mapping error is minimized iteratively by adjusting the KM variables $\{y_{km}\}$ with a gradient descent procedure,

$$y_{km}(t+1) = y_{km}(t) - \alpha_{km}(t) \frac{\partial E(t)}{\partial y_{km}(t)},$$

where $\alpha_{km}(t)$ is a variable gain term that controls the rate of convergence. Sammon uses

$$\alpha_{km}(t) = \alpha_0 \left| \frac{\partial^2 E(t)}{\partial y_{km}(t)^2} \right|,$$

with $0.3 < \alpha_0 < 0.4$ found empirically to lead to satisfactory convergence. The mapping error is recomputed at each iteration and is a function of the original distances in N-space $\{\nu_{ij}\}$ and the computed distances $\{\mu_{ij}\}$ in M-space. Since there are $K(K-1)/2$ unique distances involved the algorithm requires on the order of K^2 operations per iteration.

Although there is no restriction on the dimensionality of the output space we use $M = 2$ here for convenience. To illustrate the behavior of this algorithm consider first mapping the vertices of a square ($N = 2$). The four points in $\Re^2$ are $\{\mathbf{x}_k\} = (0\ 0), (0\ 1), (1\ 0), (1\ 1)$. Two solutions are shown in Fig. 2 after 20 iterations. The first configuration is the correct solution ($E = 0$); the other solution occurs for a different initialization because the algorithm gets trapped in local minimum ($E = 0.069$). The four vertices are marked 0–3; the lines are drawn for reference. In three dimensions, one can only approximately map the vertices of a cube (0–7) into a 2-D map. Two solutions after 50 iterations each are shown in Fig. 3 ($E = 0.062$).

As pointed out by Bennett (1969) the intrinsic dimensionality of a collection of signals (configuration of points in N-space) depends only on the signals and not on the space itself. For example, consider a kind of random walk in a 10-D space where one takes a unit step forward in a random direction. The 2-D map of a 20 step "random walk" (50 iterations) resembles a Peano curve (Fig. 4).

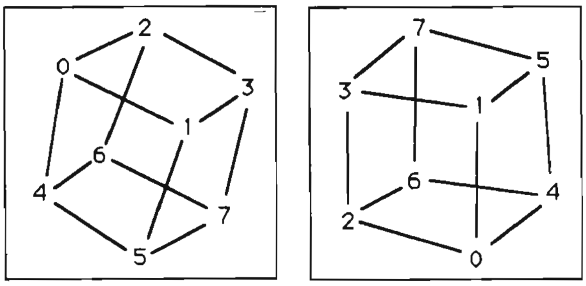


FIG. 3. Two mappings of the vertices of a cube.

In general, as the number of dimensions increases, the mapping error increases, as the interpoint distances can be only approximately preserved. Nevertheless, as we demonstrate in the remainder of the paper, the resultant mappings can still provide a great deal of insight into the global and local structure of the data.

4. CASE STUDY: BRAIN MAPPING

Several standard data sets have been used in the literature to demonstrate NLM and related algorithms. Here we present an initial case study illustrating the utility of NLM in an on-going effort to map the major cellular areas of the brain (Armstrong *et al.*, 1991).

Standard maps which divide the brain up into cytoarchitectonical areas based on structural features such as the size and orientation of cells, the thickness of individual layers of cells, and the homogeneity of cells within a layer have been in use for almost a century. These maps, originally developed by Brodmann (Damasio and Damasio 1989), have been the standard used by neuroanatomists for comparing brains and are often used as a guide for interpreting imagery such as that collected by NMI and PET scanners.

Our goal is to develop texture signatures of the major cellular regions of the brain areas in order to assess quan-

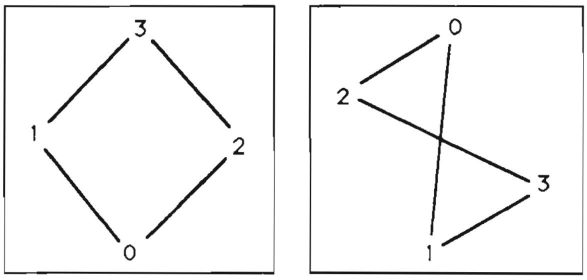


FIG. 2. Two mappings of the vertices of a square.

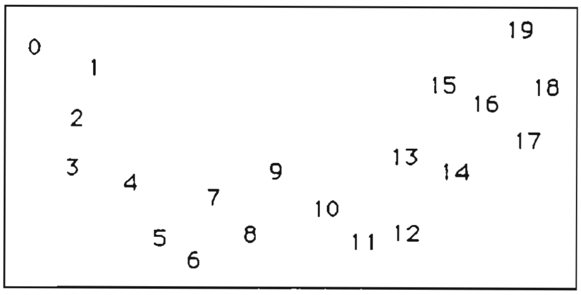


FIG. 4. Random walk in 10-dimensional space.

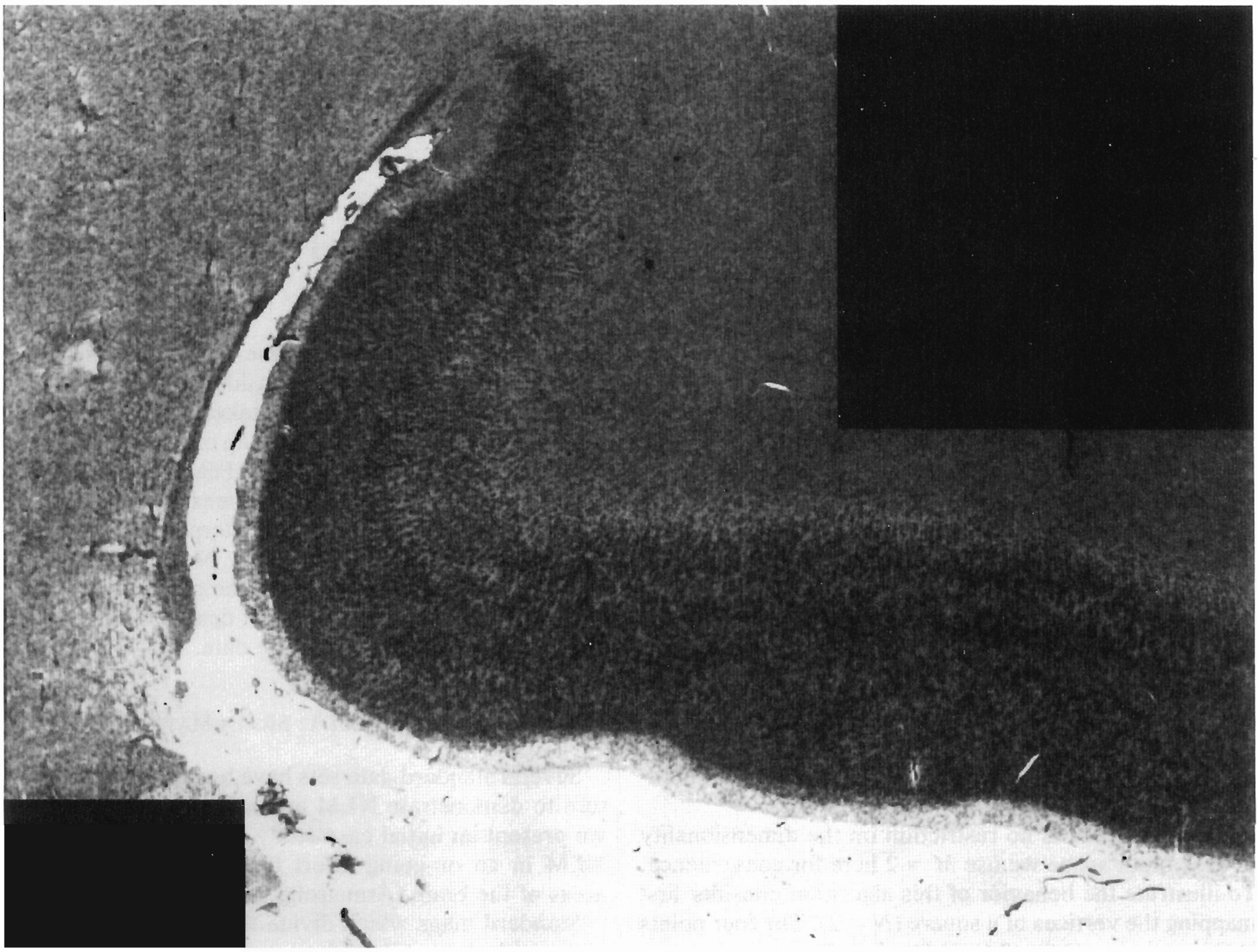


FIG. 5. Image of a portion of the posterior cingulate gyrus.

titative differences between individuals. We are currently examining differences in the posterior cingulate gyrus (Fig. 5), which is a part of the limbic system responsible for attentional mechanisms and emotions. NLM is used to visualize the relationship between texture measurements for different parts of the brain.

Our texture measurement is simply the cross-sectional optical density profile normal to the surface of the cortex.[1] Because of the complex structure of folds (gyri and sulci) we first identify those regions in the digitized image where the slice is normal to the surface of the brain. Next the brain is "straightened" by computing the medial axis of the cortex in these regions and resampling the optical densities along profiles perpendicular to the medial axis. A section of the cortex in Fig. 5 after it has been straightened is shown in Fig. 6a. The image has also been proportionally scaled perpendicular to the medial axis. The straightened image is smoothed and sub-sampled to produce a series of vectors through the cortex at fixed intervals (Fig. 6b).

An initial experiment (Armstrong *et al.*, 1991) was performed to attempt to (1) identify representative vectors (signatures) by clustering, (2) associate each of these vectors with one of the Brodmann regions, and (3) classify the remaining vectors.

A series of 32, 32-element vectors were extracted from each slice of the straightened image in Fig. 6b. An iterative clustering procedure was used to find representative vectors or clusters. The procedure consisted of finding the two closest vectors (in 32-space), randomly eliminating one of the vectors, and repeating until only one vector was left. A heuristic based on the principle of minimum description length was used to find the optimal number of

[1] The data are digitized sections of actual brains that have been sliced, mounted on glass slides, and stained to define cellular structures.

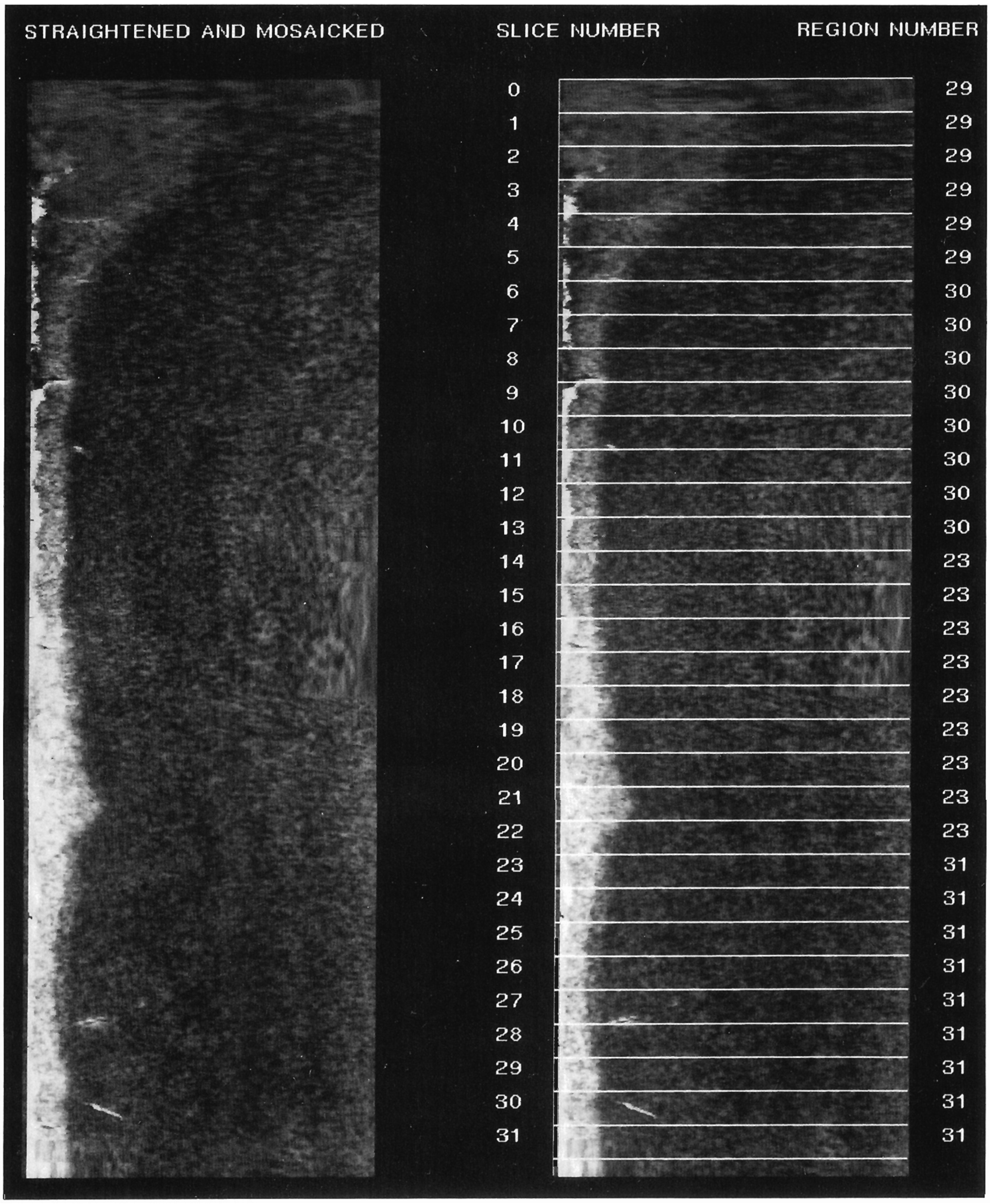


FIG. 6. Straightened and proportionally rescaled image (left) showing slice and region numbers (right).

clusters. At each step, an objective function proportional to the number of vectors remaining (i.e., the number of clusters) and the total error in approximating all 32 vectors by the clusters was computed. The minimum value of the objective function was used to find the best clustering. This occurred for four vectors corresponding to slices 0, 9, 18, and 25. One of four Brodmann regions (23, 29, 30, and 31) was then assigned, by a neuroanatomist, to each cluster. The remaining clusters were classified into one of the Brodmann regions using a minimum Euclidean distance classifier. The final assignment of Brodmann regions to slices is shown in Fig. 6b.

The results are depicted in a 2-D map (Fig. 7) obtained by mapping the set of 32, 32-D vectors using the NLM algorithm. The vectors extracted from each slice are shown in (a), the clusters in (b), and the classification

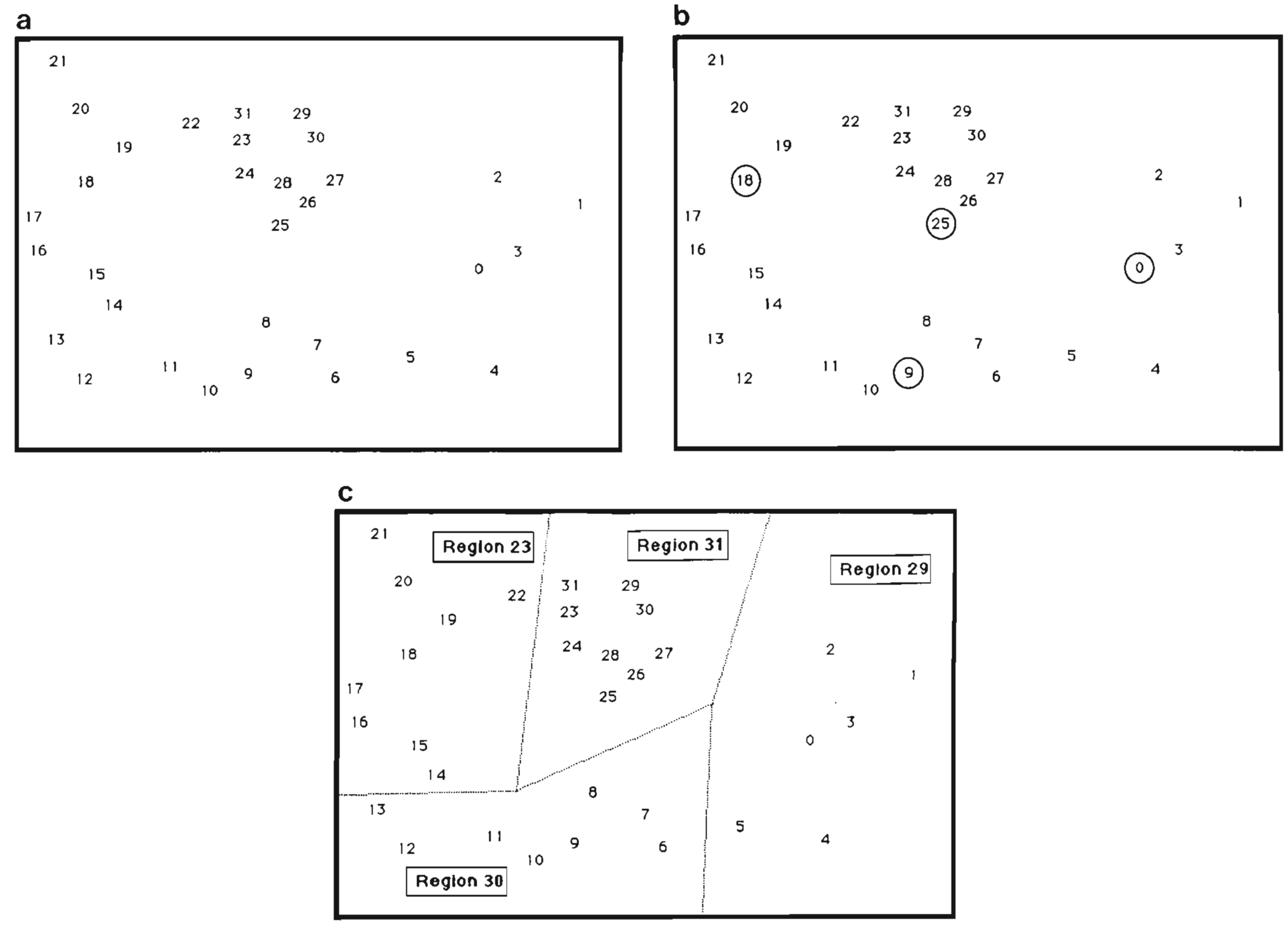


FIG. 7. Two-dimensional map and classification results. Numbers 0–31 refer to the position along the cortex. (a) Two-dimensional map of brain data; (b) clusters; (c) classification results.

results in (c). The decision regions in (c) were hand drawn based on the classification results. With a few exceptions the points tend to follow a 1-D trajectory in the 2-d map with distinct transitions between cellular regions (5–6 transition between regions 29 and 30, 13–14 transition between regions 30 and 23, and 22–23 transition between regions 23 and 31). We have noted similar trajectories in other brain slices analyzed and are currently attempting to develop image normalization techniques based on these results for matching textures between brains for classification.

5. EXTENDED NONLINEAR MAPPING ALGORITHM

As noted earlier the main limitation of the basic NLM algorithm is that the computation grows quadratically with the number of points. This section describes an extension to NLM based on mapping a subset of points using Sammon's original algorithm and mapping the remaining points by multidimensional interpolation.

Recall that $\{\mathbf{x}_k: k = 1, 2, \ldots, K\}$ is the original set of points in N-space, and $\{\mathbf{y}_k: k = 1, 2, \ldots, K\}$ is the set of points in M-space computed by NLM. Let $\mathbf{z} = \mathbf{x} @ \mathbf{y} = [x_1 x_2 \cdots x_N y_1 y_2 \cdots y_M]$ be a point in $\mathfrak{R}^{N+M}$, i.e., in the space formed by adjoining $\mathfrak{R}^N$ and $\mathfrak{R}^M$. It can be shown (Carlotto and Izraelevitz 1989) that a new point $\mathbf{y}^*$ in $\mathfrak{R}^M$ can be computed as a function of a new point $\mathbf{x}^* \in \mathfrak{R}^N$ and $\{\mathbf{z}_k: k = 1, 2, \ldots, K\}$ by weighting the $\mathbf{y}_k$ by an amount inversely related to the distance between $\mathbf{x}^*$ and the corresponding $\mathbf{x}_k$; e.g.,

$$\mathbf{y}^* = \frac{\sum_k \mathbf{y}_k a_k}{\sum_k a_k} = \frac{\sum_k \mathbf{y}_k \exp - [\mathbf{x}_k - \mathbf{x}^*]^{\mathrm{T}}[\mathbf{x}_k - \mathbf{x}^*]/2\sigma^2}{\sum_k \exp - [\mathbf{x}_k - \mathbf{x}^*]^{\mathrm{T}}[\mathbf{x}_k - \mathbf{x}^*]/2\sigma^2},$$

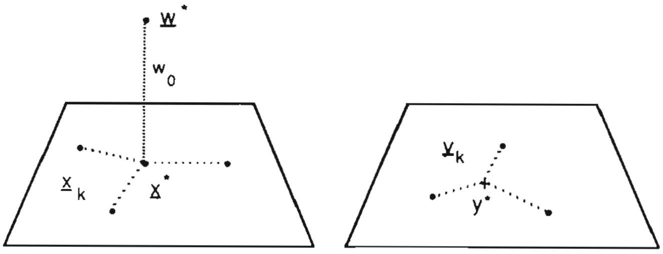


FIG. 8. Interpolation of new points may not be unique.

where σ is a parameter that controls the amount of smoothing. For small σ the output $\mathbf{y}^*$ jumps to the value of the nearest neighbor $\mathbf{y}_k$; at the other extreme as σ increases, $\mathbf{y}^*$ approaches the average of the $\mathbf{y}_k$. The smoothing factor is automatically determined using a leave-one-out procedure described in (Carlotto and Izraelevitz 1989).

A problem not previously identified with earlier approaches that are based on mapping new points by way of a smaller set of reference points (e.g., Biswas *et al.*, 1981) is that in certain cases the mapping may not be unique. In particular, consider the case where a new point $\mathbf{w}^*$ is outside of the space spanned by the $\{\mathbf{x}_k\}$, as depicted in Fig. 8. Let w_0 be its distance in the orthogonal subspace. The weights a_k are given by

$$\exp - \left[\frac{1}{2\sigma^2} \sum_n^N (x_n^* - x_{kn})^2 + w_0^2\right]$$

$$= \exp - \left[\frac{w_0^2}{2\sigma^2}\right] \prod_n^N \exp - \left[\frac{(x_n^* - x_{kn})^2}{2\sigma^2}\right].$$

The orthogonal term appears in the numerator and denominator and so cancels. The response to $\mathbf{w}^*$ thus cannot be distinguished from that of $\mathbf{x}^*$ (Fig. 8). Determining when such cases occur can be accomplished by feeding back $\mathbf{y}^*$ and using it to compute an estimate $\mathbf{u}^*$ of $\mathbf{w}^*$. As $\mathbf{w}^*$ moves outside the original data space, the distance w_0 increases, and the correlation between $\mathbf{w}^*$ and $\mathbf{u}^*$ decreases. In practice, if the correlation falls below an established threshold, one may decide not to map the point, as the result may be misleading. The correlation may also be used to detect outliers or anomalies in the data.

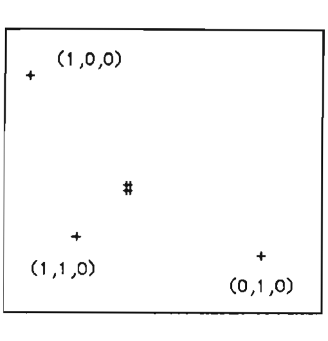


FIG. 9. Two-dimensional map for original three reference points + and new points #.

TABLE 1
Correlation Decreases as Point Moves outside of Original Data Space

w_0	w_1	w_2	u_0	u_1	u_2	y_0	y_1	ρ
0.5	0.5	0	0.5	0.5	0	0.58	0.52	1
0.5	0.5	0.5	0.5	0.5	0	0.58	0.52	0
0.5	0.5	1.0	0.5	0.5	0	0.58	0.52	−1

To illustrate this phenomenon, consider three points located at (1 0 0), (1 1 0), and (0 1 0). A 2-D mapping of those points is shown in Fig. 9. Three new points, (.5 .5 0), (.5 .5 .5), and (.5 .5 1), were then mapped as described above. The results are summarized in Table 1 where the new points $\mathbf{w}^* = (w_0\ w_1\ w_2)$, interpolated coordinates $\mathbf{y}^* = (y_0\ y_1)$, back interpolated points $\mathbf{u}^* = (u_0\ u_1\ u_2)$, and ρ is the normalized correlation coefficient. All three $\mathbf{w}^*$ are mapped to the same $\mathbf{y}^*$ (denoted # in the 2-D map). The decreasing correlation for the last two points indicates that they are increasingly outside the original data space and thus cannot be uniquely interpolated.

6. CASE STUDY: HYPERSPECTRAL DATA ANALYSIS

Imaging spectrometers and hyperspectral imagery are becoming increasing important in geological and environmental remote sensing (Goetz *et al.*, 1985). The narrow spectral bandwidth of these sensors permits the accurate classification of minerals, the identification of chemical plumes, and the detection of stressed vegetation to name just a few applications. A color image is composed of three components or bands: red, green, and blue. A hyperspectral image consists of hundreds of bands where each band is an image collected over a very narrow spectral range $\Delta\lambda \sim 10$ nm. Because of the large number of bands, one can think of hyperspectral imagery as a data volume (x, y, λ) or as an image of vectors as shown in Fig. 10.

The interpretation of hyperspectral imagery generally involves comparing spectral signatures of unknown materials to those of known materials (e.g., from a spectral signature library or previously identified within the image). We have explored the use of the extended NLM algorithm to compute material maps from a set of reference spectra of known surface materials. These material maps provide a context for identifying unknown mate-

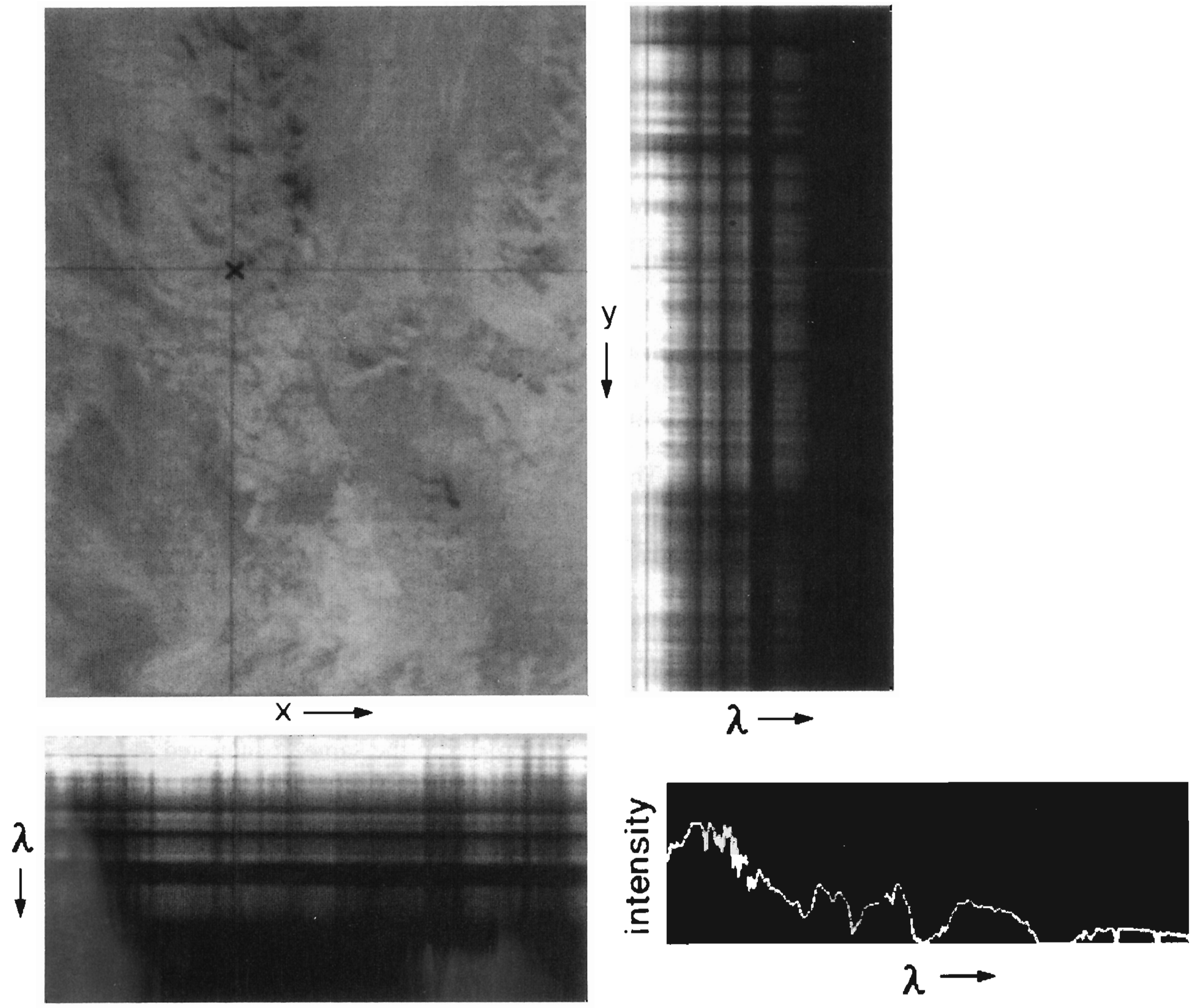


FIG. 10. AVIRIS data set (Drum Mountain, Utah).

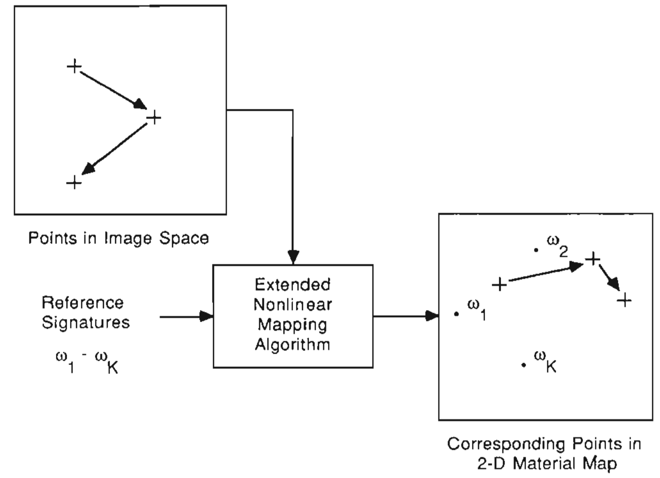


FIG. 11. Hyperspectral data analysis tool based on the extended NLM algorithm.

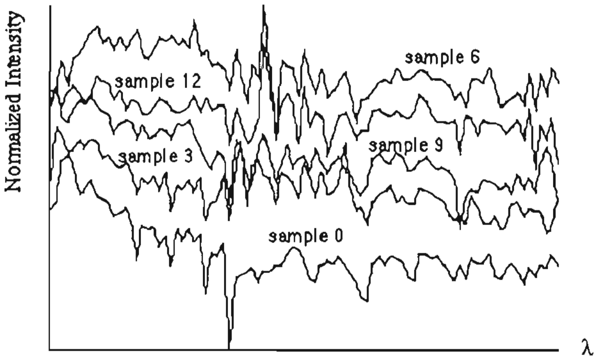


FIG. 12. Spectral response curves for selected samples of hyperspectral data.

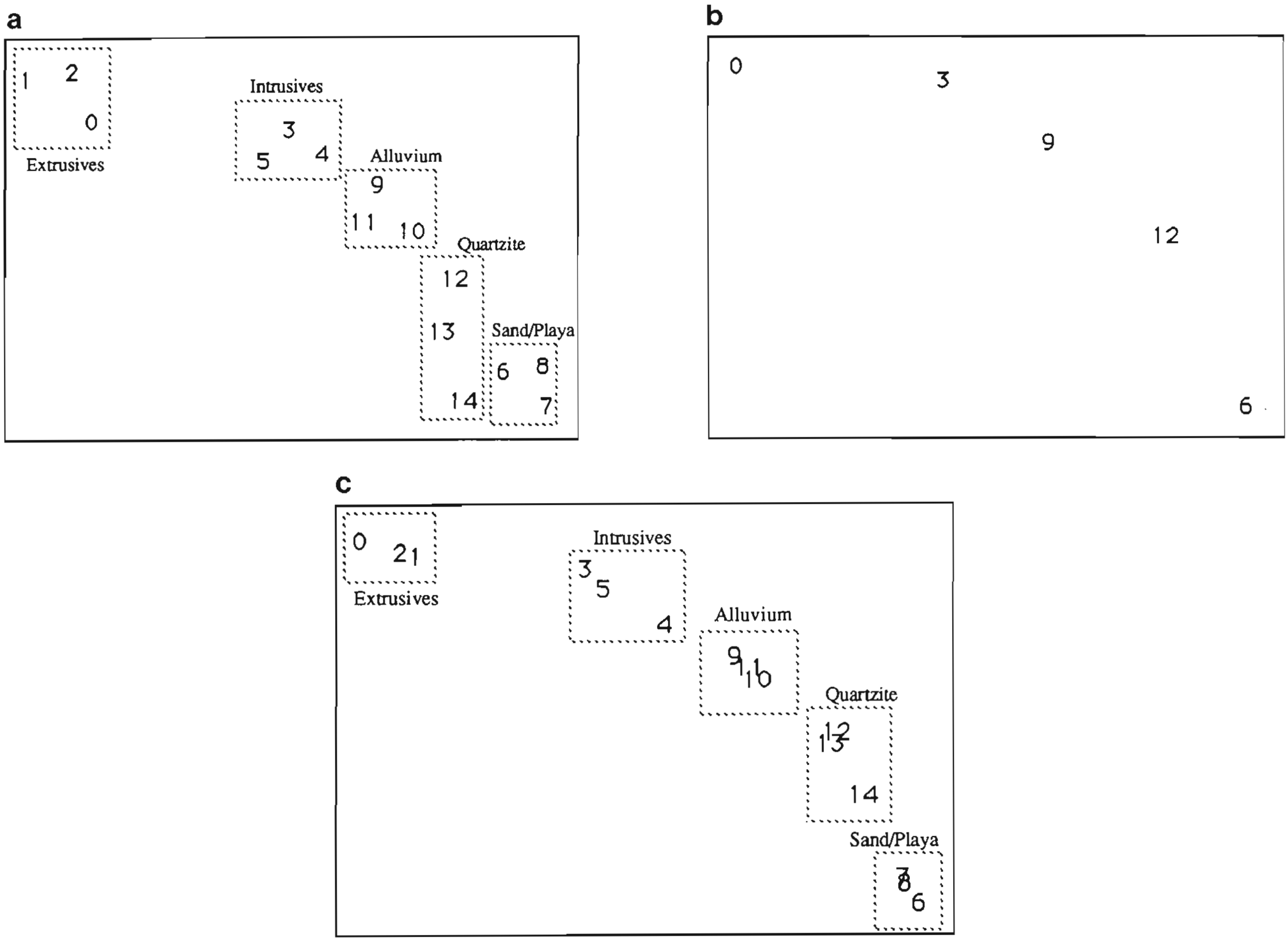


FIG. 13. Two-dimensional maps for hyperspectral visualization: (a) full data set mapped by NLM algorithm; (b) selected samples mapped by NLM algorithm; (c) interpolated results for full data set.

rials in the image. The concept is illustrated conceptually in Fig. 11. A reference map is initially computed for a set of spectra of known materials using the conventional NLM algorithm. The map assigns each material to a point in space. As the imagery analyst moves a cursor to different points in the image, the spectral vector at each point is mapped into the 2-D map by interpolation, using the extended NLM algorithm described in the previous section. The material properties at a given point in the image can be assessed by comparing its position in the 2-D map to that of known materials. For example, if an image point falls in between two reference spectra it is likely that the material in the image is a mixture of the corresponding materials; if the correlation falls below a specified threshold, the spectrum may correspond to a new material not in the reference library. In principle, the entire image can be systematically mapped by this process.

An experiment was performed using the first 90 bands of NASA's AVIRIS sensor over Drum Mountain, Utah. Three samples from five surface categories were identified in the image by a geologist: extrusives (0, 1, 2), intrusives (3, 4, 5), sand/playa (6, 7, 8), alluvium (9, 10, 11), and quartzite (12, 13, 14). Selected signatures are plotted in Fig. 12. Fig. 13 summarizes the results from the first part of the experiment. First we mapped all 15 samples using NLM (a). Next, one sample from each of the categories was mapped (b) and used to interpolate the remaining 10 samples by the extended NLM (c). Samples of similar materials appear near one another both in (a) and in (c) as expected. In the second part of the experiment, we mapped samples 0, 3, and 9 (extrusives, intrusives, and alluvium only), and attempted to interpolate values for samples 1, 4, 10, 7, and 13 (the last two being sand/playa and quartzite). The computed correlation values were 0.9, 0.84, 0.83, 0.48, and 0.57, respectively. The correlation of the last two materials not in the reference map were significantly lower, as expected.

7. CASE STUDY: TEXT VISUALIZATION

Advances in document retrieval systems and network-based information services are providing increased access to large distributed text databases. The focus to date has been on the development of document retrieval and natural language understanding systems (e.g., Salton 1986, Stanfill and Kahle 1986, and Jacobs and Rau 1990) which are designed to retrieve and, to a limited extent, understand freetext pertaining to particular topics of interest. Less attention has been devoted to text visualization, e.g., for clustering and visualizing the contents of an entire database of freetext.

This final case study describes a graphical approach for visually summarizing text databases which uses NLM to map texts to points in a 2-D space. The approach is based on converting texts to binary N-vectors known as *surrogate codes*. The method, which is similar to hashing, converts words into random P-bit codes which are stored in a dictionary (Knuth 1973). Each code is a list of P integers (bit positions) selected at random between 0 and $N - 1$. A word is encoded by setting its P bit positions to 1 in the binary N-vector. The surrogate code for a piece of text is thus computed by simply "or-ing" together the codes for each word in the text. The performance of the surrogate coding method depends on the size of the vector N, the number of code bits P per word, and the number of words R encoded per text (Stanfill and Kahle 1986).

We restrict ourselves to 2-D maps ($M = 2$) here for convenience but are exploring volumetric (3-D) rendering on high-end graphics workstations at present. By mapping surrogate codes to points in a 2-D map, we expect that texts which deal with similar subject matter will tend to cluster. The assumption is that if the texts are long enough, similar texts will tend to have more words in common than those that are dissimilar. The similarity between texts is measured by correlating the corresponding binary vectors. For two texts, A and B, the correlation between their corresponding surrogate codes $\mathbf{a}$ and $\mathbf{b}$ is $C(\mathbf{a}, \mathbf{b}) = \Sigma a_n b_n$. The correlation is roughly proportional to the number of words common to the two texts.

The text encoding and mapping process may be summarized as follows.

Build Word Frequency Table. Over all text files a table $\mathbf{F} = \{f_i\}$ is computed that lists for each unique word that is encountered the number of text files in which it appears. For K text files, the frequency of the ith word f_i is $0 < f_i \leq K$.

Encode Text. For each text file, the words within it are sorted in terms of their frequencies in the previous table. It has been observed that words that occur neither too frequently nor too infrequently tend to contain most of the content of the text (Salton 1975). Here we eliminate words that occur in only one text ($f_i = 1$) since they do not contribute to the clustering, and sort the remaining words in order of increasing frequency. The first R words are selected and encoded as described earlier.

Compute Distance Table. First the $K(K - 1)/2$ unique correlations between all pairs of text vectors are computed. Next, the correlations are converted to the $\{v_{ij}\}$ distances required by the NLM algorithm (Section 3) according to

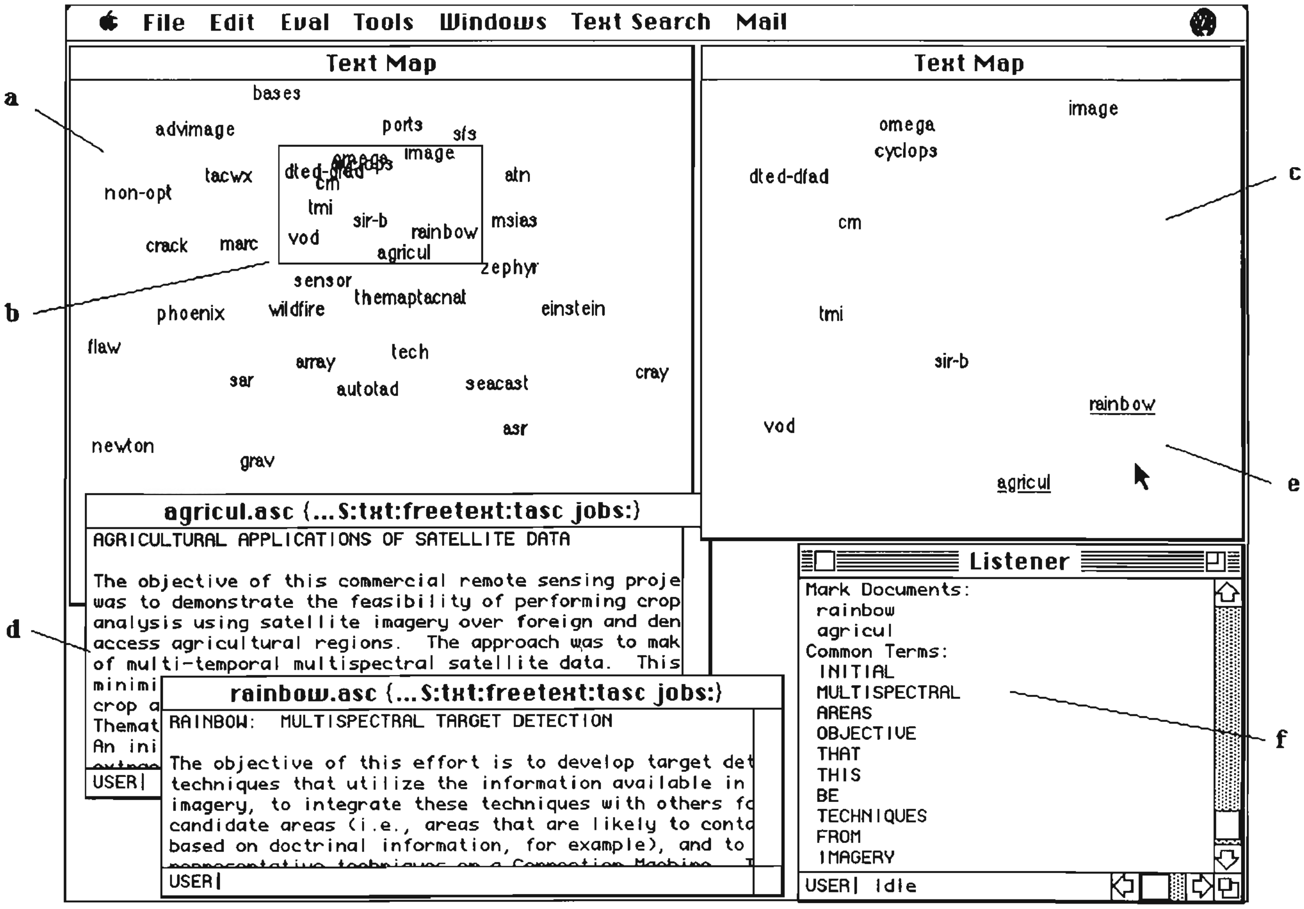
$$D(\mathbf{a}, \mathbf{b}) = C_{\max} - C(\mathbf{a}, \mathbf{b}) + \varepsilon,$$

where $D(\mathbf{a}, \mathbf{b})$ is the distance between texts A and B, $C_{\max}$ is maximum correlation computed over all texts in the database, and ε is a small positive number that keeps the mapping error finite.

As an example of the above application, a database of TASC project summaries was examined. A 2-D map (Fig. 14) was computed using $R = 128$ words/text file, $P = 4$ bits/word, and $N = 1024$ bit vectors. As shown in the figure, the 2-D text map depicts each project summary as a point in 2-space (a). For large databases and in regions of the map where a large number of documents have clustered (b), a tool is provided to expand selected portions of the map (c). Documents may be selected with the mouse from the 2-D map and displayed (d). Finally, the content of the database can be locally examined by selecting one or more documents (shown underlined) with the mouse (e) and listing in a separate window (f), in order of their frequency of occurrence, the words common to the selected documents. Common words are determined by "and-ing" the surrogate code vectors, and searching the dictionary for all words whose P bit positions are set to one in the resultant vector.

8. SUMMARY

Nonlinear mapping techniques for analyzing multidimensional data were described. An application of Sammon's NLM algorithm for visualizing the relationships between texture measurements derived from digitized brain sections in 32 dimensions was initially presented. An extension to NLM for mapping arbitrarily large data sets using a novel multidimensional interpolation approach was then described. Its use in identifying surface materials in hyperspectral imagery on the basis of their spectral signatures was demonstrated on an AVIRIS data set (~100 dimensions). Finally, a text visualization application was considered that involved the use of a non-Euclidean distance measure and binary vectors of extremely high dimensionality (>1000) for analyzing the structure and content of databases of freetext.



agricul.asc {...S:txt:freetext:tasc jobs:}

```
AGRICULTURAL APPLICATIONS OF SATELLITE DATA

The objective of this commercial remote sensing proje
was to demonstrate the feasibility of performing crop
analysis using satellite imagery over foreign and den
access agricultural regions.  The approach was to mak
of multi-temporal multispectral satellite data.  This
minimi
crop a
Themat
An ini
USER|
```

rainbow.asc {...S:txt:freetext:tasc jobs:}

```
RAINBOW:  MULTISPECTRAL TARGET DETECTION

The objective of this effort is to develop target det
techniques that utilize the information available in
imagery, to integrate these techniques with others fo
candidate areas (i.e., areas that are likely to conta
based on doctrinal information, for example), and to
USER|
```

```
Mark Documents:
  rainbow
  agricul
Common Terms:
  INITIAL
  MULTISPECTRAL
  AREAS
  OBJECTIVE
  THAT
  THIS
  BE
  TECHNIQUES
  FROM
  IMAGERY
USER| Idle
```

FIG. 14. Prototype text visualization system.

One area of future work is to combine nonlinear mapping techniques for dimensionality reduction with scientific visualization techniques; e.g., to visualize a collection of text documents in an animated and color-coded 3-space.

In addition to extending Sammon's original technique to arbitrarily large data sets, this paper has demonstrated that NLM can be used to assist in the interpretation of data sets of very high dimensionality across a wide range of applications. These results suggest that one can effectively interpret such data and need not suffer Bellman's "curse of dimensionality" (Duda and Hart 1973) in the process.

REFERENCES

H. Andrews, *Introduction to Mathematical Techniques in Pattern Recognition,* Wiley–Interscience, New York, 1972.

E. Armstrong, R. Becker, and M. Carlotto, Feature-based mapping of major cellular regions in the posterior cingulate gyrus, in *Abstracts 21st Annual Meeting, Society for Neuroscience, New Orleans, 1991.*

R. Bennett, The intrinsic dimensionality of signal collections, *IEEE Trans. Inform. Theory* **IT-15** (5), 1969.

G. Biswas, A. Jain, and R. Dubes, Evaluation of projection algorithms, *IEEE Trans. Pattern Anal. Mach. Intelligence* **PAMI-3** (6), 1981.

M. Carlotto and D. Izraelevitz, System realization using associative memory building blocks, in *Proceedings of SPIE Symposium on Advances in Intelligent Robotics Systems, Philadelphia, 1989,* Vol. 1192.

C. Chang and R. Lee, A heuristic relaxation method for nonlinear mapping in cluster analysis, *IEEE Trans. Systems Man Cybernet.* **SMC-3,** 1973, 197–200.

W. M. Coughran, Jr., and E. Grosse, Techniques for scientific animation, in *Extracting Meaning from Complex Data: Processing, Display, Interaction,* Proceedings of the SPIE, Vol. 1259, 1990.

H. Damasio and A. Damasio, Road maps to neuroanatomy, in *Lesion Analysis in Neuropsychology,* Oxford Univ. Press, London, 1989.

R. Duda and P. Hart, *Pattern Classification and Scene Analysis,* Wiley, New York, 1973.

G. Grinstein and S. Smith, The perceptualization of scientific data, in *Extracting Meaning from Complex Data: Processing, Display, Interaction,* Proceedings of the SPIE, Vol. 1259, 1990.

A. Goetz, G. Vane, J. Solomon, and B. Rock, Imaging spectrometry for earth remote sensing, *Science* **228,** 1987, 1147–1153.

J. Hammersley, The distribution of distances in a hypersphere, *Ann. of Math. Statist.* **25,** 1950, 447–452.

P. Jacobs and L. Rau, SCISOR: Extracting information from on-line news, *Comm. ACM* **33** (11), 1990.

D. Knuth, *The Art of Computer Programming: Sorting and Searching,* Addison–Wesley, Reading, MA, 1973.

T. Kohonen, *Self-Organization and Associative Memory,* Springer-Verlag, Berlin/New York, 1988.

R. Lee, J. Slagle, and H. Blum, A triangulation method for the sequential mapping of points from N-space to two-space, *IEEE Trans. Comput.* **C-27,** 1977, 288–292.

G. Salton, *A Theory of Indexing,* Soc. Industrial Appl. Math., Philadelphia, 1975.

G. Salton, Another look at automatic text-retrieval systems, *Comm. ACM,* **29** (7), 1986.

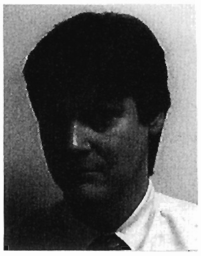
J. Sammon, A nonlinear mapping algorithm for data structure analysis, *IEEE Trans. Comput.* **C-18** (5), 1969.

R. Shepard, The analysis of proximities: Multidimensional scaling with an unknown distance function, I, *Psychometrika* **27** (2), 1962.

R. Shepard, The analysis of proximities: Multidimensional scaling with an unknown distance function, II, *Psychometrika* **27** (3), 1962.

C. Stanfill and B. Kahle, Parallel free-text search on the Connection Machine system, *Comm. ACM* **29** (12), 1986.

F. W. Young and P. Rheingans, Dynamic statistical graphics techniques for exploring the structure of multivariate data, in *Extracting Meaning from Complex Data: Processing, Display, Interaction,* Proceedings of the SPIE, Vol. 1259, 1990.



MARK J. CARLOTTO received B.S., M.S., and Ph.D. degrees in electrical engineering from Carnegie–Mellon University in 1977, 1979, and 1981, respectively. His research at CMU was in hybrid (optical/digital) computing architectures for solving linear matrix equations. From 1981–1983 he was an assistant adjunct professor in the College of Engineering at Boston University. From 1981 to the present he has been employed by TASC. Dr. Carlotto is currently a division staff analyst and is involved in research and development projects related to satellite remote sensing, information visualization, biotechnology, and image databases. His publications include one book and over 50 technical papers in the areas of image processing, pattern recognition, remote sensing, geographic information systems, and optical computing. He recently chaired an SPIE conference devoted to the application of visual computing technologies in the earth and atmospheric sciences. Dr. Carlotto is a senior member of the IEEE.