Data Compression in Digitized Lines

Khagendra Thapa
Surveying and Mapping Program, Construction Department, School of Technology, Ferris State University, Big Rapids, MI 49307

ABSTRACT: The problem of data compression is very important in digital photogrammetry, computer assisted cartography, and GIS/LIS. In addition, it is also applicable in many other fields such as computer vision, image processing, pattern recognition, and artificial intelligence. Consequently, there are many algorithms available to solve this problem but none of them are considered to be satisfactory. In this paper, a new method of finding critical points in a digitized curve is explained. This technique, based on the normalized symmetric scattered matrix, is good for both critical points detection and data compression. In addition, the critical points detected by this algorithm are compared with those by zero-crossings.

INTRODUCTION

THE ADVENT OF COMPUTERS has had a great impact on mapping sciences in general and digital photogrammetry and cartography in particular. Nowadays more and more existing maps are being digitized, and attempts have been made to make maps automatically using computers. Moreover, once we have the map data in digital form, we can make maps for different purposes very quickly and easily. Usually, the digitizers tend to digitize more data than is required to adequately represent the feature. Therefore, there is a need for data compression without destroying the character of the feature. This can be achieved by the process of data compression. In this paper, a new method of data compression is described which is efficient and has a sound theoretical basis as it uses the eigenvalues of the Normalized Symmetric Scattered (NSS) matrix derived from the digitized data.

THE NATURE OF SCATTER MATRICES AND THEIR EIGENVALUES

Consider the geometry of the quadratic form associated with a sample covariance matrix. Suppose $P = (p_1, p_2, \ldots, p_n)$ be a finite data set in R^2 and P is a sample of n independently and identically distributed observations drawn from real two dimensional population.

Let (μ, Σ) denote the population mean vector and variance matrix and let $(\mathbf{v}_p, \mathbf{V}_p)$ be the corresponding sample mean vector and sample covariance matrix. These are then given by Uotila (1986) as follows:

$$\mathbf{v}_p = \Sigma\, p_i/n;\ \mathbf{V}_p = \Sigma\, (p_i - \mathbf{v}_p)\,(p_i - \mathbf{v}_p). \tag{1}$$

Multiply both sides of the equation for V_p by $(n - 1)$ and denote the right hand side by S_p, i.e.,

$$(n - 1)\, \mathbf{V}_p = \mathbf{S}_p \quad \text{or} \quad \mathbf{S}_p = (n - 1)\, \mathbf{V}_p. \tag{2}$$

The matrices $\mathbf{S}_p$ and $\mathbf{V}_p$ are both 2×2 symmetric and positive semi-definite. Because these matrices are multiples of each other, they share identical eigen-spaces.

According to Anderson and Bezdek (1983) one can use the eigenvalue and eigenvector structure of S_p to extract the shape information of the data set it represents. This is because the shape of the data set is supposed to mimic the level shape of the probability density function $f(x)$ of x. For example, if the data set is bivariate normal, S_p has two real, non-negative eigenvalues. Let these eigenvalues be λ_1 and λ_2. Then the following possibilities exist (Anderson and Bezdek, 1983):

- If both λ_1 and $\lambda_2 = 0$, then the data set P is degenerate, and $\mathbf{S}_p$ is invertible and there exist with probability 1, constants a, b, and c such the $ax + by + c = 0$. In this case the sample data in P lie on a straight line.
- If $\lambda_1 > \lambda_2 > 0$, then the data set represents an elliptical shape.
- If $\lambda_1 = \lambda_2 > 0$, then the sample data set in P represents a circle.

EIGENVALUES OF THE NORMALIZED SYMMETRIC SCATTER MATRIX (NSS)

Supposing that one has the following data:

$$P = (P_1, P_2 \ldots, P_n)$$

where $P_i = (x_i, y_i)$.

Then the normalized scattered matrix A is defined as

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \mathbf{S}_p/\text{trace}\ (\mathbf{S}_p). \tag{3}$$

For the above data set, A is given by the following:

$$\text{Let Deno} = \Sigma\, ((x_i - x_m)^2 + (y_i - y_m)^2), \tag{4}$$

$$\begin{aligned} a_{11} &= 1/\text{Deno}\ \Sigma\, (x_i - x_m)^2 \\ a_{12} &= 1/\text{Deno}\ \Sigma\, (x_i - x_m)\,(y_i - y_m) \\ a_{21} &= 1/\text{Deno}\ \Sigma\, (x_i - x_m)\,(y_i - y_m) \\ a_{22} &= 1/\text{Deno}\ \Sigma\, (y_i - y_m)^2 \end{aligned} \tag{5}$$

where $v_x = (x_m, y_m)$ is the mean vector defined as

$$x_m = \Sigma\, x_i/n, \text{ and } y_m = \Sigma\, y_i/n. \tag{6}$$

Note that the denominator in Equation 3 will vanish only when all the points under consideration are identical. Clearly matrix 'A' is symmetric because matrix S_p is symmetric. The characteristic equation of A is given by

$$|\,\mathbf{A} - \lambda\mathbf{I}| = 0 \tag{7}$$

which may be written as (for 2×2 matrix)

$$\lambda^2 - \text{trace}(\mathbf{A})\, \lambda + \text{Det}\ (\mathbf{A}) = 0 \tag{8}$$

where Det(A) = Determinant of A. It follows from Equations 4 and 5 that the trace of the matrix A, trace(A), = 1, because trace(A) $= a_{11} + a_{22} = 1$. Hence, the characteristics equation of A reduces to

$$\lambda^2 - \lambda + \text{Det}(\mathbf{A}) = 0 \tag{9}$$

The roots of this equation are the eigenvalues and are given by

$$\lambda_1 = [1 + \sqrt{1 - 4^*\text{Det}(\mathbf{A})}]/2 \text{ and} \tag{10}$$

$$\lambda_2 = [1 - \sqrt{1 - 4^*\text{Det}(\mathbf{A})}]/2$$

For convenience put $D_x = \sqrt{1 - 4^*\ \text{Det}(\mathbf{A})}$; then

$$\lambda_1 = (1 + D_x)/2 \tag{11}$$

$$\lambda_2 = (1 - D_x)/2 \tag{12}$$

Now λ_1 and λ_2 satisfy the following two conditions:

$$\lambda_1 + \lambda_2 = 1 \tag{13}$$

PHOTOGRAMMETRIC ENGINEERING AND REMOTE SENSING,
Vol. 56, No. 4, April 1990, pp. 517–518.

0099-1112/90/5604–517$02.25/0

Because the sum of the roots of an equation of the form $ax^2 + bx + c = 0$ are $\lambda_1 + \lambda_2 = -b/a$.

Subtracting Equation 12 from Equation 11, one obtains

$$\lambda_1 - \lambda_2 = D_x \tag{14}$$

Because the eigenvalues λ_1 and λ_2 satisfy Equations 13 and 14, the three cases discussed previously reduce to the following form (Anderson and Bezdek, 1983):

- The data set represents a straight line if and only if $D_x = 1$.
- The data set represents an elliptical shape if and only if $0 \leq D_x \leq 1$.
- The data set represents a circular shape if $D_x = 0$.

ALGORITHM TO DETECT CRITICAL POINTS USING NSS MATRIX

The analysis of the eigenvalues of the NSS matrix can be used to extract the shape of the curve represented by the data set. Hence, this characteristic may be exploited to detect critical points in the digital curve. Assuming that the data are gross error free, and devoid of excessive noise, one can outline the algorithm to detect critical points in the following steps:

(1) First, take three points from the data set.
(2) Compute the NSS matrix and, hence, its eigenvalues.
(3) If D_x is greater than a certain tolerance (e.g., 0.95), add one more point to the data and repeat step 2. If D_x is less than the tolerance, point 2 is a critical point.
(4) Retain point 2 and repeat the process from step 1 with point two as the new starting point.
(5) Repeat the process until the end of the data set is reached.

The algorithm discussed in the previous section is useful in detecting the critical points in vector data. The only parameter involved in this technique is D_x which was defined earlier. By varying the value of D_x between say 0.9 to 1.05, one can get a varying amount of detail in a curve. Figure 1 shows the selected critical points for $D_x = 0.96$.

There are 50 points selected in Figure 1. It is clear from Figure 1 that this method will be very useful for compression of digitized data because it retains the overall shape of the curve without retaining the unnecessary points.

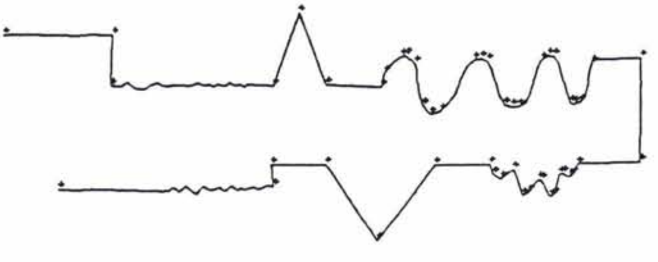

FIG. 1. Results of critical points selection by NSS Matrix. There were 50 points selected.

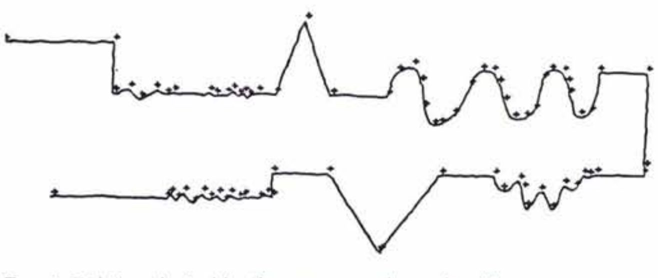

FIG. 2. Points selected by the zero-crossings algorithm.

COMPARISON OF RESULTS OBTAINED BY ZERO-CROSSINGS ALGORITHM AND NSS MATRIX METHOD

The zero-crossings algorithm for line generalization and data compression in raster data was devised by Thapa (1987, 1988a). In this approach, the raster data are first converted to chain codes. These chain codes (Freeman, 1978) are then modified to take care of the discontinuity and scale problem (Eccles *et al.*, 1977; Pavlidis, 1978). The modified chain codes are then convoluted with the mask obtained from the second derivative of the Gaussian for a particular value of sigma (sigma is the parameter of the Gaussian function). The zero-crossings in the convoluted values are then identified. The raster data points which correspond with the zero-crossings provide the points which are sufficient to preserve the shape of the line.

The NSS matrix method of data compression also has only one parameter D_x. By varying the value of D_x between 0.9 and 1.05, one can achieve various levels of data compression.

Note that there were 1324 pixels of size 0.25 mm in the curve which have been reduced to 78 points. A comparison of Figures 1 and 2 shows that for sigma = 4 the zero-crossings algorithm retained more detail in the line than the NSS matrix method. However, as shown in Thapa (1988b), one can vary the number of points retained by zero-crossings algorithm by changing the value of sigma. The lower the value of sigma, the more detail of the line will be retained.

CONCLUSIONS

- The analysis of the eigenvalues of the normalized symmetric scattered matrix provides a useful way of detecting critical points in digitized curves. Consequently, this technique may be used for data compression for digitized curves.
- The NSS matrix algorithm for data compression can retain the same points as the zero-crossings algorithm.

ACKNOWLEDGMENT

I express my gratitude to Dr. Toni Schenk for providing partial funding for this research under seed grant 22214 and NASA project 719035. Computer resources for the project were provided by the Instruction and Research Computer Center of The Ohio State University.

REFERENCES

Anderson, I. M., and J. C. Bezdek, 1984. Curvature and Tangential Deflection of Discrete Arcs: A Theory Based on the Commutator of Scatter Matrix Pairs and Its Application to Vertex Detection in Planer Shape Data, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-6, No. 1, pp. 27–40.

Eccles, M. J., M. P. C. McQueen, and D. Rosen, 1977. Analysis of the Digitized Boundaries of Planer Objects, *Pattern Recognition*, Vol. 9, No. 1, pp. 31–41.

Freeman, H., 1978. Shape Description Via the Use of Critical Points, *Pattern Recognition*, Vol. 10, pp. 159–166.

Pavlidis, T., 1978. A Review of Algorithms for Shape Analysis, *Computer Graphics and Image Processing*, Vol. 7, pp. 243–258.

Thapa, K., 1988a. Automatic Line Generalization Using Zero-Crossings, *Photogrammetric Engineering and Remote Sensing*, Vol. 54, No. 4, pp. xxx–xxx.

———, 1988b. Critical Points Detection and Automatic Line Generalization in Raster Data Using Zero-Crossings, *The Cartographic Journal*, Vol. 25, No. 2, pp. xxx–xxx.

———, 1987. *Critical Points Detection: The First Step To Automatic Line Generalization*. Report Number 379, The Department of Geodetic Science and Surveying, The Ohio State University, Columbus, Ohio.

Uotila, U. A., 1986. *Adjustment Computation Notes*, Dept. of Geodetic Science and Surveying, The Ohio State University.

(Received 1 June 1989; accepted 28 July 1989; revised 16 August 1989)