A THEORY OF THE CARTOGRAPHIC LINE

Thomas K. Peucker
Simon Fraser University

INTRODUCTION

The basic difference between the approach taken in this paper and those in other studies of the line is that it is postulated that a line has always a certain thickness.

The concept of a line having an areal extent is not too obvious to the mathematician who usually thinks of the line as a locus with zero width. The cartographer can, however, feel comfortable with it; in fact, the basis of many manual line géneralizations is the idea of widening the line with decreasing scale. Since the curvature of a line has to be larger than the width of its generalized equivalent in order that it does not disappear within the black area of the line, with decreasing scale, a line a) occupies a larger area, b) the curvature of the line decreases and c) it can be defined by less points. The reduction of points for the definition of the line can also be called an increase of abstraction in the definition of the line.

The presented theory can be associated with an alternative explanation. A line is a combination of a number of frequencies, each of which can be represented by certain band-widths. The break-up of a line into series of bands, therefore, could be equated with the stepwise removal of the high frequencies from the line.

Before going on, however, a few terms should be explained in detail. A line is here defined as a sequence of connected points. A connection between two points in a sequence of points is called a segment. The extent of a line is its number of segments. We thus exclude from this discussion those lines which are defined by one or several mathematical functions since their characteristics and problems are quite different from the discretely identified lines.

To the cartographer, the use of the term "line" for what we intend to discuss might be quite obvious. However, other terms such as chain (Chrisman, 1974), chord (O'Callaghan, 1975), segment, arbitrary line (H. Freeman, 1975), random line, snake

Editor's Note: This paper has been developed with the support of ONR Contract N00014-73-C-0109. The author is thankful to David H. Douglas, University of Ottawa, Peter Frisbee, Harvard University, and James Little, Robert Fowler, David Mark and Christopher Boerner, all Simon Fraser University, for their discussion and programming support.

(Clement, 1974), etc., have all been used. Most of them, however, can lead to misunderstanding if transferred to another discipline. Chain is also used for a type of line encoding (Freeman, 1961); chord is a straight line in photogrammetry, as is segment in several disciplines. The term "arbitrary line" could be misleading in cartography since a cartographic line is never arbitrary, with at most arbitrary deviations from a straight or smoothly curved line, etc. Furthermore, some of the definitions like chain (Chrisman, 1974), and snake (Clement, 1974), imply additional topological information. Therefore, it might be safer to use the general term "line."

In computer graphics, the number of line segments per line is usually relatively small, it is, however, frequently very large in cartography. Lines with more than 1,000 points are not rare in some storage systems (Schmidt, 1969). It is therefore logical that the interest to develop better algorithms for line manipulation is very great in cartography. This paper is a contribution to these efforts through an attempt to formalize some characteristics of lines, after a short discussion of the most frequent encoding systems.

The most straight-forward form of discrete encoding of lines is by the absolute coordinates (Figure 1a) of the points along the line. The order of the point in the array of coordinates gives the sequence of points. If the lines are discontinuous, their connectedness can be given by a continuancy label (the structure of many plotting commands) or a dummy point which indicates the end of one line and the start of the next.

For the incremental mode, an absolute start is given for every line, but the subsequent points are encoded in terms of their distance to the previous points (Figure 1b). This approach sometimes reduces storage requirements since the increments can be given by shorter storage units (e.g., 16 bits); some computations can be faster, like the computation of the length of a line -- but for most line manipulations the absolute coordinates of the points have to be computed.

The third very common encoding mode is called "chain encoding" (Freeman, 1961, Freeman and Shapiro, 1975). A line is broken into straight portions of equal change in the x- and y- direction. This results in eight unique step directions which can be encoded in three bits (Figure 1c). Therefore, a byte of six bits can contain two steps. For bytes of eight bits, the first bit can be a continuation code and the other three are either step-indicators or continuation counters (Tomlinson, 1974).

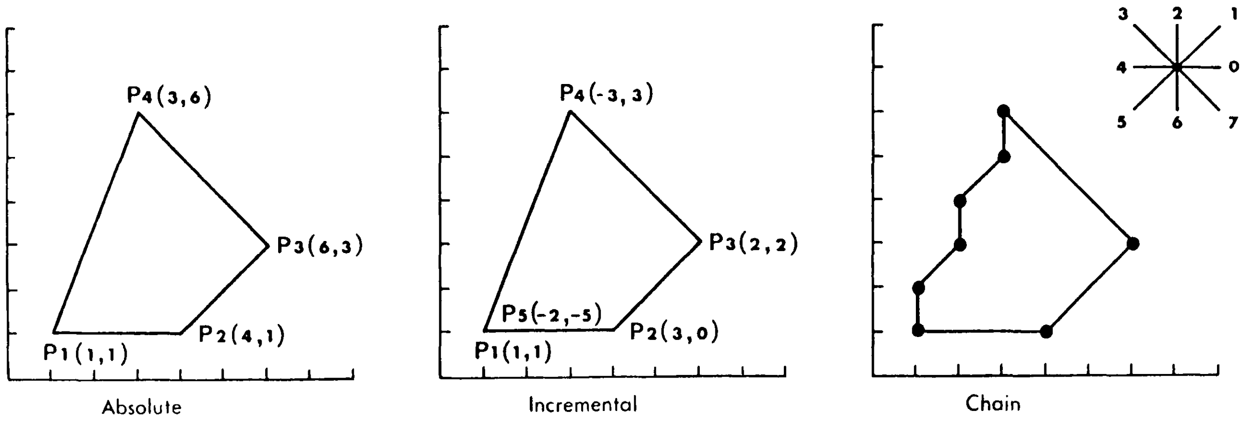

Figure 1

There are other encoding systems (see for example, Pfaltz and Rosenfeld, 1967, for skeleton encoding), but they usually have to be converted into absolute coordinates for most computations.

Although the presented theory is implemented on the basis of the absolute coordinate scheme, it is unimportant which coordinate scheme the reader is used to. Indeed, certain computations are more efficient in some of the other schemes.

THE CONCEPT

A line of any extent can be defined by

1) A general direction and a band with
2) A width, and
3) Its length

Different ways of computing the general direction will be discussed later. At this time it can be said that the general direction of a line can be any direction; the reduction of a problem to a solution is however faster the closer the general direction approaches the direction of the minimum bounding rectangle.

The band is the bounding rectangle, given a certain general direction (Figure 2). In other words, the sides and ends of the band are parallel, and perpendicular, respectively, to the general direction, totally enclosing it. In set-theoretic notation this can be stated as follows:

$$L^1 C\ B^1,\ L^1 = P_1,\ P_2,\ ....\ P_i,\ ....\ P_n \qquad (1)$$

Where L^1 represents the total extent of the line, and B^1 the band of the total line and P_1 to P_n all the points of the line.

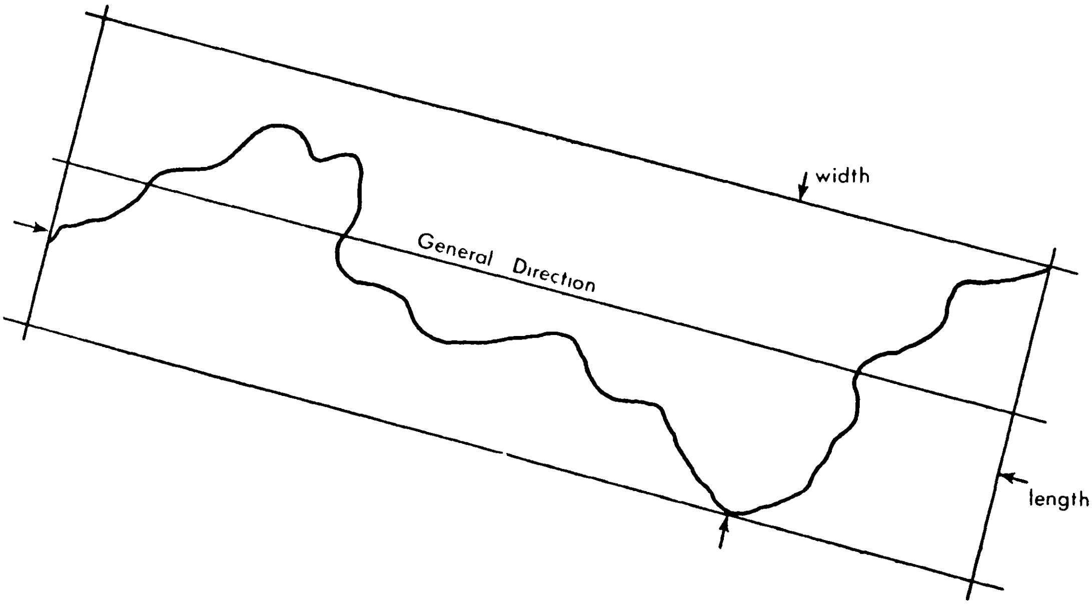

Figure 2

A line of n points can be subdivided into up to n-1 sublines, each with its own characteristic band. At the highest degree of abstraction it is represented by only one band, at the lowest degree it is represented by n bands. At the highest degree of abstraction the band is generally the widest, at the lowest, the bands have a zero width. In other words, with decreasing degrees of abstraction, the area covered by the band decreases (Figure 3) although it can remain the same from one step to the next in exceptional cases.

$$B^1 \geq B_1^2 \cup B_2^2 \geq \dots \bigcup_{i=1}^{k} B_i^k \geq \dots \bigcup_{i=1}^{n} B_i^n \qquad (2)$$

Here, the upper subscript indicates the order of abstraction and, implicitly, the number of sub-bands.

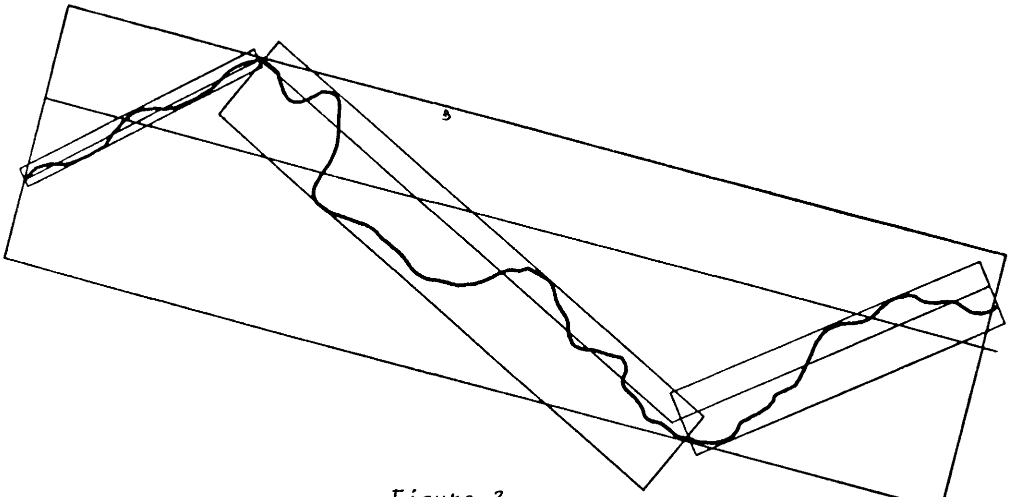

Figure 3

(2) can be proven with one reservation. If one breaks up a subline at step i into two sublines at step i+1, the limiting case is that the two resulting bands have the same width and direction as the previous band in which case the combination of the two bands is identical to the original band. In all other cases the bands have to have a different direction and therefore leave some portion of the original band unoccupied. Although the new bands can have portions outside the original band, these portions are insignificant since they do not contain any sections of the line. The width of a band can actually get larger from one level to the next. These degenerate cases can be easily detected and circumvented in the implementation of the theory by breaking the line up into sections if the width of the band is larger than a given portion of the length.

According to (2) a line with n points has n levels of abstraction with a total of $\frac{(n-1).n}{2}$ bands. It is the objective of the theory to keep the computation at any time on the highest level of abstraction that the particular problem allows. The critical issue of the theory is to find the highest level in any different type of problem. Some applications of this approach will be shown after a discussion of different types of bands.

THE TYPES OF BANDS

A multitude of bands can be constructed from a line. The right type depends upon the frequency at which the characteristics of the band must be computed (the higher the frequency the faster the computation has to be), the particular speed of convergence (the slower the problem converges the smaller the band has to be), and on the special characteristics of some encoding systems.

The band which can be computed most quickly is the bounding rectangle which is orthogonal to the coordinate system. One axis of the coordinate system is the direction of the band and the extreme points in the direction of the other axis mark the width of the band. This method has been used for a long time to test situations like the potential intersection of two segments. The major advantage of this type of band is its computational speed. Since all four sides have to be computed every time, however, it is not always the fastest, because with other types one often has to compute only two sides.

For chain encoding (Freeman, 1961) and skeleton encoding (Pfaltz and Rosenfeld, 1967), a band at an angle of 45 or 135 degrees can be of advantage. The eight steps of the chain encoding can be converted into steps to the left or right of the 45 degree and the 135 degree directions with four counters keeping track of the maxima and minima. The wider margin will then give the length of the band and the smaller one its width. Similarly for skeleton encoding half the block-distance of a rhombus gives its extension in the four 45 degree directions.

An easy compromise which the author has adopted for most of the implementations of the theory is to link the start and the end of the line and take that as its general direction. The advantage of this approach is again speed. Its disadvantage is that the band can become relatively large in cases where the line is closed or nearly closed.

A more involved approach which should bring better results is to find the principle axis of the point set. This method should produce a band which is fairly close to the minimum band.

Freeman and Shapiro (1975) have shown that this minimum band can be produced via the bounding convex polygon. The minimum band and the bounding convex polygon have at least one side in common. The minimum band can therefore be constructed by computing the convex polygon and then testing for each side whether the resulting band is smallest.

LINE GENERALIZATION

The theory of the cartographic line has initially been conceived as an afterthought to the development of a algorithm for line generalization (Douglas and Peucker, 1973). This algorithm was developed at about the same time in at least two other papers (Ramer, 1972; Duda & Hart, 1973). Therefore, the execution of the idea has to be explained only as far as it relates to the theory (Figure 4).

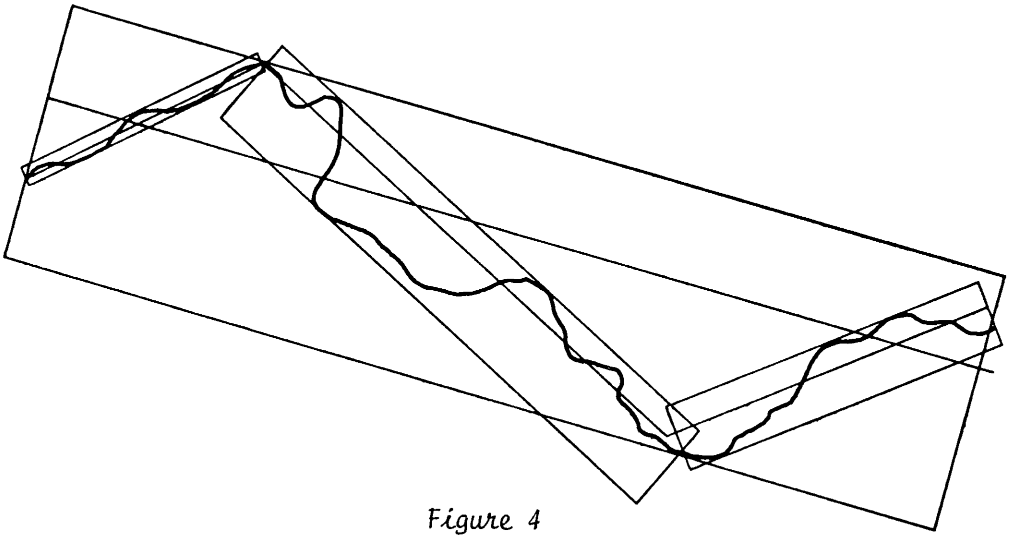

Figure 4

The line is partitioned into subsets until each subset has a band with a width less than a predetermined threshold. At each step, the partitioning process is performed by selecting those points which touch the sides of the bands as starts and ends of the subsets.

The actual implementation of the concept is somewhat different (Douglas and Peucker, 1973). The general direction of the line is given by the link between its start and its end. For every point, the vertical distance from the link is computed and the point with the maximum absolute value is retained as the new point which divides the line in two portions which are subsequently treated independently the same way (Figure 5).

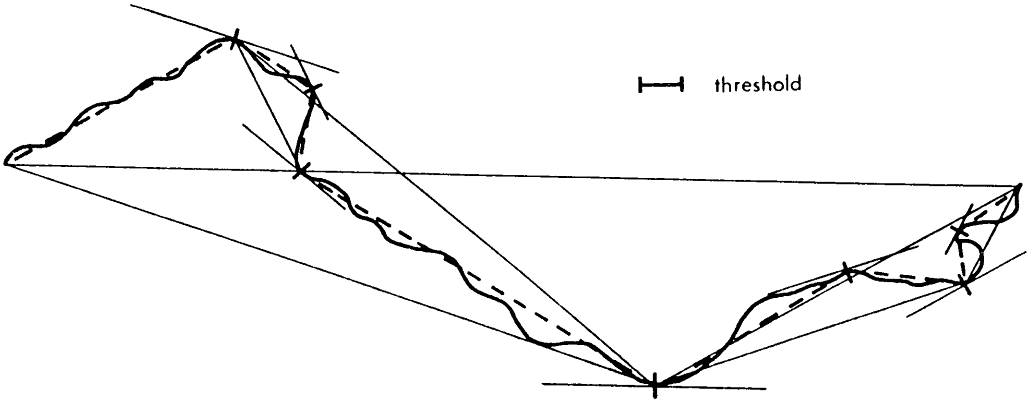

Figure 5

LINE INTERSECTION

In cartography, lines can be very extended, i.e., contain many segments. When the computations involve two or more lines, the speed of traditional algorithms usually develops in proportion to the product of the extents of the two lines. The intersection of two lines is a good example. The traditional approach tests every segment of one line against every segment of the other line. There is usually a pretest involved which checks whether the bounding orthogonal rectangles of the two segments overlap and return with a "false" signal if they do not. But even this pretest can be rather time consuming if it has to be performed at a frequency of the product of the number of segments of the two lines.

The concept of the band of a line can be of substantial help. It is quite obvious and easy to prove that the intersection(s) of two lines have to be located within the parallelogram which is built by the intersection of the two bands (Figure 6). Since for the second step only the portions of the lines within the parallelogram are used, the second pair of bands tend to be much smaller. Therefore, the problem tends to converge to the pair of minimum bands (two single segments) very quickly.

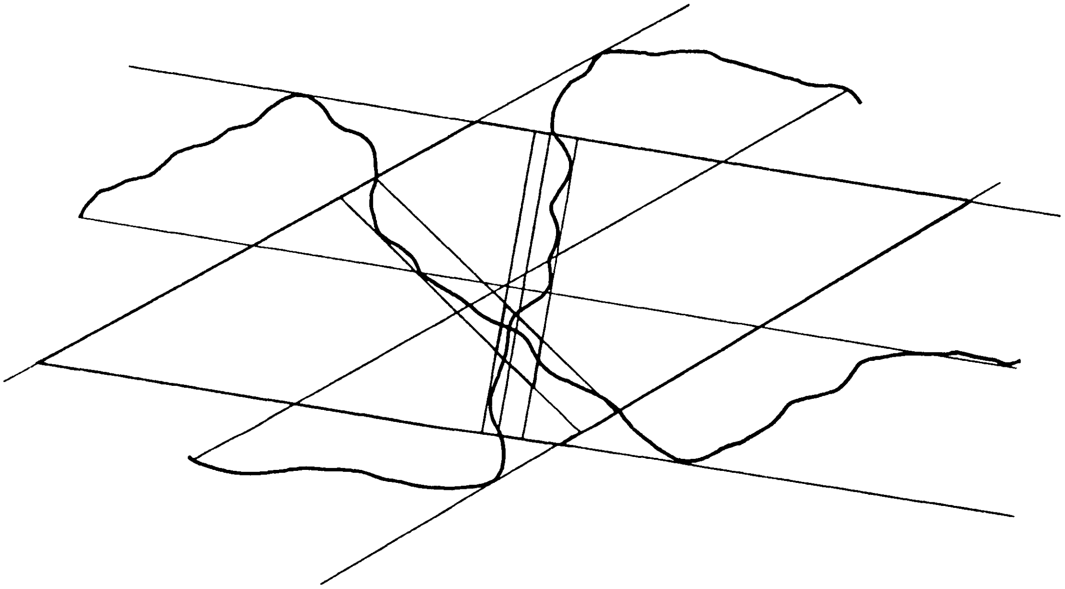

Figure 6

The practical implementation of the approach takes advantage of several simplifications characteristic of the problem. For one, only the width of the band has to be determined. Furthermore, the parallelogram is not derived but the process is shortened to close to half by constructing the band of one line and then immediately finding the portion of the other line within the first band, etc. (Figure 7). Also the process is stopped when the remaining lines consist of a total of six or less segments because the conversion process with less segments can run into complications.

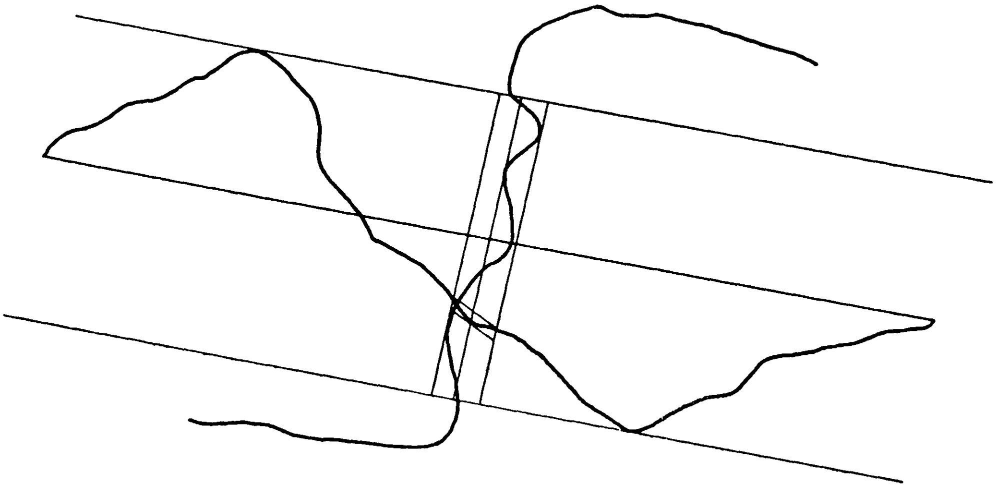

Figure 7

There are other precautions which take action if the procedure is in danger of converging slowly or has reached a dead end loop. If the "raw length" of the line (the distance between start and end of the line) is less than three times its width, the line is partitioned into three portions which are processed one after the other.

The savings in computer time over the traditional method are impressive. For example, in the implemented algorithm, the multiple intersection of two lines with 120 points each was achieved in 1/23 of the time of the traditional method. This would mean a ratio of less than 1/200 with 1,000 points per line.

LINE MATCHING

Another problem which can be solved very elegantly by the band approach, but can be a tedious problem without it, is the test whether or not two lines can be considered to be the independently coded representations of one line. Encoding usually produces two types of factors which will cause two digital representations of a line to differ:

- Encoding noise or digitizing errors
- Differences in the choice of the sampling points

A method which measures the degree of agreement of the two representations is therefore desirable.

This can be done by the requirement that the bands of the two lines overlap at a certain percentage. If, therefore, the error variance of the digitizing is known, one can construct the bands with a width of twice that variance and then test whether or not the bands of one line overlap with the bands of the other at more than a given percentage.

POINT SORT AND POINT-IN-POLYGON SEARCH

The point-in-polygon problem has been attacked many times but to the author's knowledge, only three so far have developed algorithms which employ less than all the segments of the polygon for the test whether a point is inside or outside a polygon (Pfaltz and Rosenfeld, 1967; Loomis, 1968; James, et al., 1973).

The concept of the band can also be used for this purpose if the procedure has to be performed several times for one polygon, or if several points have to be sorted into a number of polygons.

If the data structure is based on "chains" (i.e., portions of lines with topological indicators, see Chrisman and Little, 1974), the chains can be used for the first case (one polygon -- several points). Otherwise, any procedure can be used which partitions the polygon boundary into three or more portions. In the second case (several polygons -- several points), a chain structure has to be produced if it is not available.

The basic idea of the sorting procedure is that the problem space can be divided into three parts -- the area left of a band, the area right of it and the band itself. If one or more points fall into the band, the level of abstraction has to be lowered and bands of several subsections employed. The level is lowered until there are no points left within the bands.

It is quite clear how this approach can be used for the case of one polygon. First, the boundary is broken down into three or more portions. The bands of these portions are constructed and the first point tested. If any of the bands have to be split, every new band is stored with its direction and dimensions (a tree structure would serve the purpose best), to be used for later tests. Of course, this approach is economically feasible only if the number of points to be tested is relatively large.

In the case of several polygons and several points, the method can already be feasible with relatively few points because it avoids a loop in which a point has to be tested through all polygons. For this problem, the chains have to be grouped to lines which divide the problem space in a successive order. The first divide the total area, the next two the two areas left and right of the first, etc. When all chains have been used, all points are sorted in. In other words, the method uses a chain only once, no matter how many points are involved.

DOUBLE LINE AND OTHER LINE SYMBOLS

It is hoped that the theory of the cartographic line will be useful to problems yet to be discovered. It should also serve as a guideline for the solution of problems which seem to be less connected with the theory. As an example, the construction of the double line and other line symbols will be discussed.

Several algorithms are known which construct double lines around a given line or one new line at a given distance from an input line. The procedures usually simply reconstruct every point at the given vertical distance to one or both sides of the line. As Figure 8 shows, the procedure can get totally confused when very small zig zag lines occur.

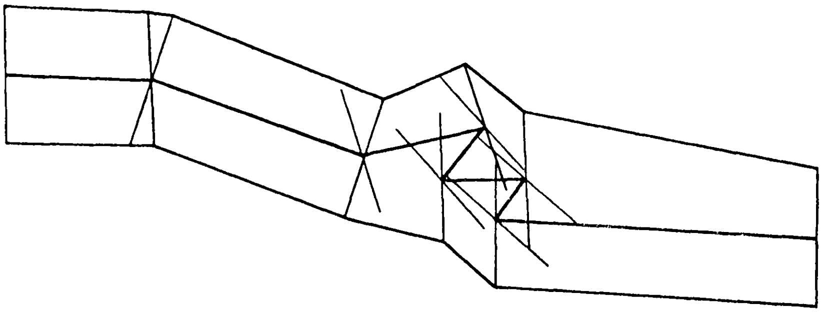

Figure 8

For a solution within the framework of the above concept one has to realize that these zig zag lines would disappear if the double line were filled black. It is therefore advisable to generalize the line with an offset of half the width of the double line before producing the double line. If the resulting corners are undesirable, additional points can be inserted between the points by a smoothing routine.

Dashed lines pose similar problems. If the distance between two starts of dashes is called the unit distance, any segment shorter than that unit will cause problems for this type of line drawing. The line should therefore also be generalized before the construction of the line symbol.

CONCLUSION

This paper presents a theory of the cartographic line which has already resulted in some algorithmic manipulations of lines and promises to serve for more applications. The theory, however, is also helpful for the understanding of the cartographic line itself. Only a slim attempt has been made to exploit this aspect of the theory. It is hoped that others will pick up the idea and expand it as part of a growing body of cartographic theory.

REFERENCES

1. Chrisman, N., and J. Little, 1974: POLYVRT Manual, Laboratory for Computer Graphics and Spatial Analysis, Harvard University.

2. Clement, A., 1973: The INTURMAP System Paper, University of British Columbia.

3. Douglas, D.H. and T.K. Peucker, 1973: "Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its Caricature," The Canadian Cartographer, Vol. 10, No. 2, December 1973, 112-122.

4. Duda, R.O. and P.E. Hart, 1973: Pattern Classification and Scene Analysis, New York, Wiley.

5. Freeman, H., 1961: On the Encoding of Arbitrary Geometric Configurations, Transaction on Electronic Computers, Institute of Radio Engineers, Vol. EC-10, 262-268.

6. Freeman, H. and R. Shapiro, 1975: "Determining the Minimum Area Encasing Rectangle for an Arbitrary Closed Curve," Communications, Association of Computing Machinery, Vol. 18, No. 7, July 1975, 409-413.

7. James, G.D., K.S. Heard and I.R. Suttie, 1973: Point in Polygon Contract -- Stage 1, Report, Nuclear Physics Division, Atomic Energy Research establishment, Harwell, Berkshire, 1973.

8. Loomis, R.G., 1968: "Boundary Networks," Communications, Association of Computing Machinery, Vol. 8, 44-48.

9. O'Callaghan, J.F., 1974: "Recovery of Perceptual Shape Organizations from Simple Closed Boundaries," Computer Graphics and Image Processing, Vol. 3, No. 4, December 1974, 300-312.

10. Pfaltz, J.L. and A. Rosenfeld, 1967: "Computer Representation of Planar Regions by Their Skeletons." Communications, Association of Computing Machinery, Vol. 10, 119-122.

11. Ramer, U., 1972: "An Iterative Procedure for the Polygonal Approximation of Plane Curves," Computer Graphics and Image Processing, Vol. 1, No. 3, November 1972.

12. Schmidt, W.E., 1969: "The Automap System," Surveying and Mapping, March 1969, 101-106.