Automatic Digitization of Large Scale Maps

Andreas Illert

Institute of Cartography
University of Hannover
Appelstrasse 9A
3000 Hannover 1
Germany

Abstract

This paper describes a software system for automatic digitization of large scale maps. The system is capable of converting raster data into structured vector data. Strategy and configuration of the software are explained. Some tests on German maps are introduced in brief.

1. Introduction

A GIS is a computerized database management system used for the capture, storage, retrieval, analysis and display of spatial data. Of these items, especially storage and analysis of mass data are what makes a GIS such a powerful tool. However, the problem remains to get the mass data into the computer.

Digital data flowing from the environment into the computer may be managed solely by remote sensing techniques. Satellite images supply up-to-date information, but such information is restricted by pixel resolution, multispectral classification and visibility of the features. Terrestrial topography produces data which are very accurate, but field work is time-consuming and expensive. For these reasons, digitization of areal photography or existing maps has become the most common method for data capture in large scale mapping.

As mentioned above, data capture is defined as part of a GIS. If you look around at the tools GIS products offer, most of them prefer manual digitization by hand-held cursor. Such conventional systems are easy to handle and adapt well to different tasks. Unfortunately, manual digitization is a laborious procedure, and human labour is costly today.

To overcome this problem some companies have provided semi-automatic systems which support manual digitization by line following algorithms. The operator has just to indicate the beginning of a line, the computer then traces the line until the next node. Line following systems are quite effective with isoline maps, but interaction increases with the number of nodes on complex maps.

Data capture by scanner is very fast and requires a minimum of human interaction. Result of scanning is a raster image. Pixel format performs excellent with two dimensional coverages in small scales, but whenever linear features, topology or non-geometric attributes have to be handled vector data are better suited. Especially in large scale mapping, vector format satisfies the demands much better than raster data. Thus, scanned data needs to be converted into structured vector data.

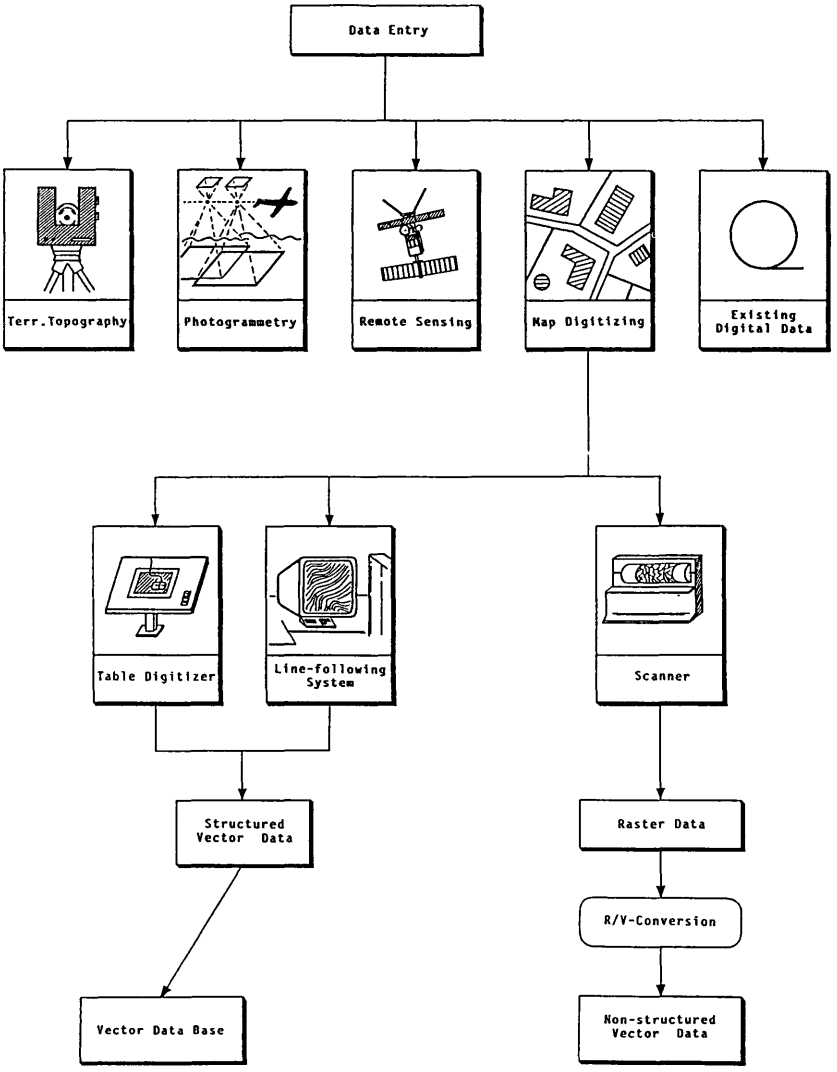

Figure 1 Data Entry to GIS */Lichtner,Illert 1989/*

At present, visual recognition of printed texts with computers seems to have no problem any more. On the other hand, automatic interpretation of drawings is still a matter of research. Nevertheless, some algorithms of text recognition may as well be applied on maps. Scientists at the Institute of Cartography, University of Hannover have developed a software system which combines common methods of Optical Character Recognition with specific algorithms for mapping applications. Concepts and results will be described in the following.

2. Strategy

Input of the recognition system is a raster image, while output is structured vector data. In other words the system has to handle both raster- and vector data, including raster-to-vector-conversion. Pattern recognition methods may be performed using both types of data.

Automatic interpretation of maps splits into two parts: First the contents of a raster image have to be broken down into graphic primitives, such as arcs, letters and symbols. Prior to application of character recognition procedures, the texts and symbols are separated from the rest of data. Next the shape of the features has to be described by numerical characteristics. Finally the features are classified according to these characteristics using methods of statistical pattern recognition. The features are treated one by one without considering any context.

In the second step the recognition system combines primitive elements into objects of a higher level. In contrast to the first step, not isolated elements but relations between features are examined. Classification of relations and structures is performed using rules which build a model of the map. Two approaches are known: the knowledge-based approach and the procedural approach.

The knowledge-based approach takes advantage of Artificial Intelligence tools. A knowledge-based system consists of a knowledge base and an inference engine. The knowledge base holds rules and facts about map features. The inference engine enables classification by matching the rules with the data. Sequence of the rules and facts in the knowledge base should play no role at all. Therefore updating of the system is quite easy. The knowledge-based system does not require information about a certain solution strategy — the inference engine tries to reach a goal without human help so long as the knowledge base supplies sufficient information. Thus knowledge-based systems are very flexible and user-friendly.

Unfortunately, the inference engine slows down rapidly with the increasing number of rules. Efficient applications are limited with nowadays technologies to a number of about 300 rules. If you consider the variety of graphic representations in map design, this is much less than required. Therefore the performance of the inference engine has to be improved by information which defines the combination of rules. As a result, application of rules tends towards a fixed sequence, and the knowledge-based system converts into a procedural system.

A procedural system follows the principles of traditional software engineering. The programmer evaluates a strategy by arranging rules in a fixed sequence, which turns to be an algorithm. Algorithms run very fast and effectively — depending on the skills and experiences of the programmer. Building up a procedural system

for automatic digitization of maps means to develop specific algorithms for different representations of spatial features. The expenditure on software development is very high. Once a procedure is established its application is limited to a specific type of map.

Recently procedural systems seem more suitable to practical applications with large sets of data than knowledge-based systems. In this context, the system developed at Hannover is oriented to the procedural approach. Nevertheless, research on knowledge-based methods continues and may replace the traditional methods in future */Meng,1990/*.

3. The recognition procedure

At the Institute of Cartography, University of Hannover a software system named CAROL (Computer-assisted Recognition of linear features) has been developed during the last five years. Goal of the system is automatic digitization of large scale German maps. As mentioned above, the system follows a procedural strategy.

3.1 Data aquisition

First of all, the map has to be scanned. Scan resolution relies on the type and quality of map. With low resolution, details might get lost. With high resolution, the amount of data and noise within the raster image will increase. In most applications a resolution of 50 μm (500 dpi) proved sufficient. The maps dealt with in Hannover used to be black and white, so raster data is organized in binary format. In case of colour maps, binary images are obtained either by scanner firmware (colour separation) or image processing (multi-spectral classification).

3.2 Raster-to-vector conversion

Raster-to-vector conversion is performed by the software tool RAVEL */Lichtner,1987/*. Using distance transformation and topologic skeletonization algorithms, the vectorization program extracts lines from the raster image and arranges them in an arc-and-node-structure. At further steps, connected lines are linked to line networks. Iconic polygons are computed from the line network. So far, only geometry and topology of the map are known. The map graphics are broken down into primitive elements.

3.3 Segmentation

To enable character recognition, letters and symbols have to be composed from arcs and separated from the rest of line graphics. Since such features are in most cases isolated networks, segmentation can be carried out quite easily by checking the size of a circumscribing rectangle. If texts and symbols intersect with the line network, additional information like line width or straightness have to be considered. Characteristic of segmentation procedures is that they take advantage of simple classification operations. Figure 2 demonstrates the effect of some threshold operations on the vector data. Segmentation procedures structure the data in a rough way and therefore help to reduce the expenditure on detailed classification.

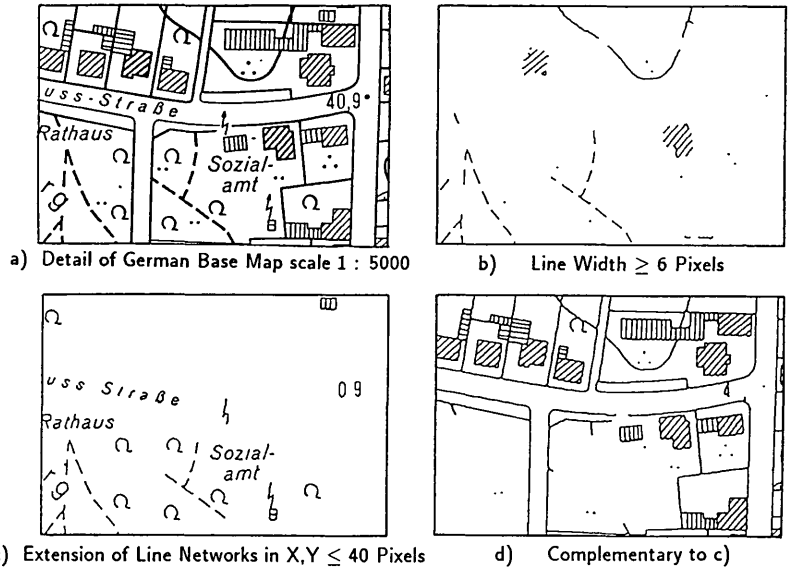

Figure 2 Segmentation of unstructured vector data

3.4 Recognition of texts and symbols

Next isolated texts, symbols and numbers have to be classified. This task is similar to the objectives of Optical Character Recognition, and so Cartography may make use of existing methods. Template matching counts as the easiest approach. A template is defined as an ideal pattern of the class. This template is matched against the raster image. The procedure reveals high success rates, although computations are very simple. Unfortunately, template matching works only on features with uniform size and rotation. Thus the method does not satisfy the demands of many applications.

So procedures have to be considered which extract characteristics independent of size and rotation. A method adopted in the CAROL system is expansion of the contour in a Fourier series */Zahn,Roskies 1972/*, */Illert,1988/*. While tracing the contour of a feature the angular change is summed up and perceived as a function of arc length from the starting point. This angle versus length function is expanded in a Fourier series. The Fourier descriptors (i.e. amplitudes and phase angles) are taken as chracteristics. Figure 3 demonstrates the method by reconstructing the original contour polygon from Fourier Descriptors of increasing degree. An expansion up to degree ten already yields sufficient information for shape recognition. In addition to Fourier Descriptors further characteristics like number of nodes or arcs may be included to improve the results of classification.

The classification itself is based on statistical analysis. The n characteristics of a feature define its location in the so-called n-dimensional feature space. Similar fea-

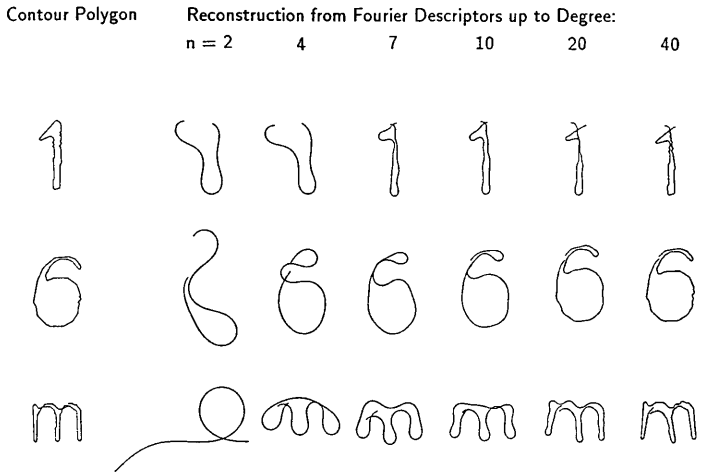

Figure 3 Expansion of contour polygons in a Fourier series

tures of a common class produce clusters in feature space. Therefore a class may be described by parameters of normal distribution. Limitations to the statistical model lead to different classification methods, such as maximum-likelihood-classification or minimum-distance-classification. However, experience reveals that success of classification depends much more on the choice of suitable characteristics than on the statistical model.

3.5 Analysis of complex features

The preceding steps structured the data into arcs, letters and symbols. Now these primitive features have to be combined into objects of higher level. The structure of a map feature in regard to its primitive components is reflected in the map legend. Some graphic structures are common to a lot of map types (e.g. dashed lines, hatching etc.), but a good part is unique for a special type of map. In the CAROL system, procedural analysis of complex features is performed by algorithms like:

- recognition of dashed lines : The system examines the data for repeated sequences of dashes, dots and gaps.
- recognition of hatched areas : The system extracts groups of parallel lines and computes an outline polygon.
- combine symbols in strings (digits to numbers, letters to texts) : The system checks relative position, rotation and size. Feature codes may be changed due to context (e.g. Number '0' and uppercase 'O' or number '1', uppercase 'I' and lowercase 'l').
- Decoding of attributes: Texts, numbers and symbols are assigned to spatial features (lines, polygons, points)

The set up of the procedure, the choice of parameters and the sequence of algorithms should be supervised by an expert. Extensive installation work is necessary whenever the procedural system faces a new type of map.

To enable knowledge-based interpretation, basic rules have to be derived from the algorithms and put into a knowledge base. When applied to a set of arcs, nodes and texts, the inference engine performs the interpretation task. With a global knowledge base no specific set-up would be required, but on the other hand technology does not yet support such ideas.

4. Applications

4.1 Hannover town plan scale 1 : 20.000

The city of Hannover has published a town plan at the scale of 1 : 20.000. Size of the map is 120 x 90 cm^2. The map is printed in 12 colours. Recently local authorities have introduced vector-based GIS software. One of its applications will be thematic mapping. In this context the town map acts as topographic base map.

The City of Hannover, Department of Cartography maintains about 20 printing separates, each of them showing a certain category of features like public buildings, industrial plants, forests or hydrography. Ten of these black-and-white separates were scanned with a resolution of 50 μm, resulting in ten raster images of 25.000 x 18.000 pixels. Raster-to-vector-conversion yields ten sets of vector data, either centre lines (in case of linear features like isolines or small rivers) or contour polygons (in case of areal features like buildings, forests etc.). The centre lines are organized in an arc-and-node structure, whereas polygons are structured hierarchically in regard to feature outlines and enclosed blank areas. Finally, lines are smoothened, and the data sets are merged by affine transformation. The whole process took about one week on microcomputer equipment, resulting in a data base of about 300.000 lines and 50.000 polygons.

4.2 Isoline maps scale 1 : 5000

German isoline maps at the scale 1 : 5000 show topography complementary to the ground situation in base map 1 : 5000. The graphic of isoline maps comprises solid lines with height numbers (height interval 10 m), dashed intermediate lines (intervals 5 m, 2,5 m or 1 m depending on gradient), height spots with numbers and slope symbols.

Computation of a DTM requires height data in digital form. For that the map sheets were scanned and vectorized. Next dashed lines were recognized. Then segmentation operations by parameters *number of nodes, extension in X* and *extension in Y* help to subdivide the data in slope symbols, isolines and numbers. Recognition of digits zero to nine is performed using Fourier Descriptors. After classification the digits were linked to height numbers and assigned to the 10 m isolines or to height spots respectively. Finally, height values have to be assigned to the intermediate isolines through interpolation within the 10 m intervals.

Interactive work is reduced to a maximum of one hour for each map sheet of 40 x 40 cm^2. The procedure is detailed in */Yang,1990a/*.

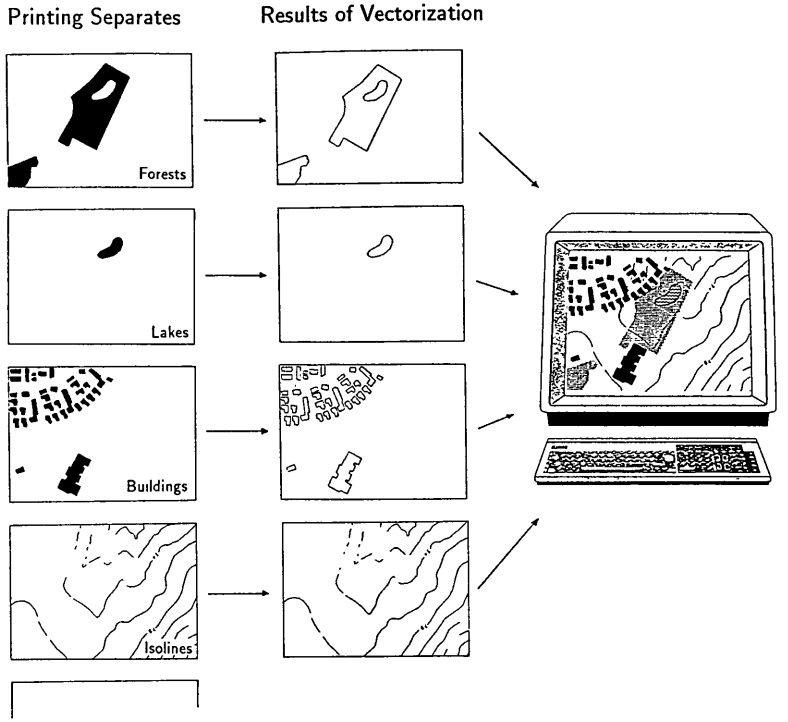

Figure 4 Digitization of Hannover town plan scale 1 : 20.000

4.3 German base map scale 1 : 5000

The base map covers the whole area of West Germany with few exceptions in the south. Ground situation is displayed in detail. The features like buildings, roads or forests are hardly affected by generalization due to the large scale. By this the map is an ideal source for GIS data bases.

The recognition procedure is set up as explained in section 3. Scanning and vectorization of a 40 x 40 cm^2 map sheet produce a set of unstructured vector data, ranging from 20.000 arcs in rural areas to 100.000 arcs in densely populated areas. Recognition of texts and symbols has to classify about 80 different features. Algorithms have been established to analyse some of the most common map features, such as

buildings	(hatched polygons)
forests	(polygons + texture of tree symbols)
meadows	(polygons + symbols: two neighbored dots in level)
gardens	(polygons + symbols: three dots arranged in a triangle)
roads	(long and narrow polygons, inside blank or street name)

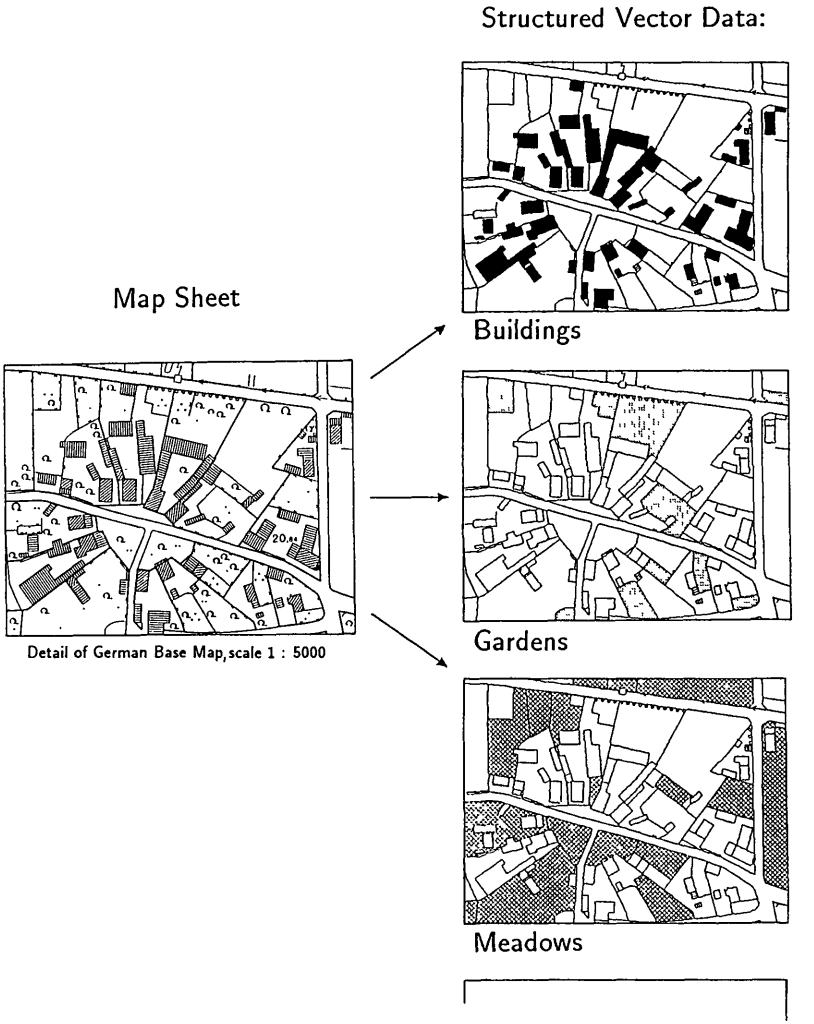

Figure 5 Digitization of German base map scale 1:5000

Examples are displayed in Figure 2 and 5.

First tests have been carried out with data aquisition for the ATKIS system of German Surveying Agencies which requires information of some 10.000 map sheets. Automatic digitization reveals success rates of 80 to 95 % (see Figure 5) */Illert,1990/*. However some problems may still arise:

1. Geometry of vector data obtained by scanning has to be enhanced to meet the high quality standards of German cadastre.
2. The classification software should be embedded in a CAD system to support interactive editing of errors during the recognition process.
3. Success rates rise with complexity of algorithms, but on the other hand the system becomes less flexible in regard to application on different map types. Because of that the knowledge-based approach should be kept in mind.

5. References

Grünreich,D. (1990) Das Projekt ATKIS: Konzeption und erste Erfahrungen aus der Aufbauphase des digitalen Landschaftsmodells 1:25000 (DLM25). Proceedings XIX FIG Congress, Helsinki/Finland June 1990, Commission 5, pp. 152-163

Illert,A. (1988) Automatic Recognition of Texts and Symbols in Scanned Maps. Proceedings EUROCARTO SEVEN, Enschede, The Netherlands Sept.88, ITC Publication No.8, pp.32-41

Illert,A. (1990) Automatische Erfassung von Kartenschrift, Symbolen und Grundrißobjekten aus der Deutschen Grundkarte 1:5000 Wissenschaftliche Arbeiten der Fachrichtung Vermessungswesen der Universität Hannover, Nr.166, 1990

Lichtner,W. (1987) RAVEL — Complex software for Raster-to-vector-conversion. Proceedings EUROCARTO VI, Brno/ Czechoslovakia 1987

Lichtner, Illert (1989) Entwicklungen zur kartographischen Mustererkennung. in: Geo-Informationsystems, Applications — New Trends. Edited by Schilcher/Fritsch, Wichmann-Verlag, Karlsruhe 1989, pp. 283-291

Meng,L. (1990) Potentialities of Quintus PROLOG in Cartographic Pattern Recognition. Proceedings EUROCARTO VIII, Palma de Mallorca / Spain, April 1990

Yang,J. (1990a) Automatische Erfassung von Höhenlinien mit Verfahren der Mustererkennung. Nachrichten aus dem Karten- und Vermessungswesen, Series I, No. 105, Frankfurt am Main 1990

Yang,J. (1990b) Automatic data capture for polygon maps from scanned data. Proceedings XIX FIG Congress, Helsinki/Finland June 1990, Commission 5, pp. 522-531

Zahn,Roskies (1972) Fourier Descriptors for Plane Closed Curves. IEEE Transactions on Computers, Vol C-21, No.3, March 1972 ,pp. 269-281