APPLYING SOME TECHNIQUES OF IMAGE PROCESSING FOR AUTOMATIC MAP DATA ENTRY

Bach Hung Khang, Vu Duy Man, Ngo Quoc Tao, Luong Chi Mai, Do Nang Toan
Institute of Information Technology, National Center for Science and Technology

Email: nqtao@ioit.ncst.ac.vn

Tel: 84-47 560 536, Fax: 84-47 564 537

ABSTRACT

This paper presents some image processing techniques that are used for map data entry software such as: enhancement of image, optical character recognition, vectorization and geometrical correction.

– *Enhancement of image*: deleting noises and holes, connecting dot lines
– *Optical character recognition*: recognizing characters in the map
– *Vectorization*: convert an image map to vector map
– *Postprocessing*: Geometrical correcting, projection transforming a paper map into the coordinating map.

We implemented these techniques in the package MapScan for Windows – an automatic data entry. This software was supported by United Nation Statistical Division and is effectively used for many kinds of features maps such as:

– *Topographical , hydrograph, settlements, roads and transformation maps etc.*
– *Design, electronic circus, fingerprint etc.*

Key words: *Enhancing image, Optical character recognition, Vectorization, Postprocessing.*

1. INTRODUCTION

A Geographic Information System (GIS) supports spatial decision-making and links descriptions of location with the characteristics of the phenomena found there. A complete GIS consists of the supporting methodology and the required technology: spatial data, hardware, software and organizational structure. GIS technology is powerful for spatial information management and analysis.

Geographic information systems are widely used in facilities management, planning, environmental monitoring, population census analysis, health service provision, hazard mapping and many other applications. GIS technology opens a new way of presenting and analyzing information from different perspectives and in a meaningful way. The tremendous benefits to a very wide and diverse group of users have contributed to its dynamic growth and application.

The proliferation and widespread use of GIS has increased efforts to the development of systems and effective techniques for computerizing paper maps.

A map represents various features of the earth's surface, shows where these features are in the real world and their relations to each other. A GIS captures and stores spatial data from external sources or from paper maps. The process of converting printed or hand-drawn maps into digital format is costly and time-intensive. More than 70% of the cost of an average GIS project is spent on data capture, and this is why the main asset of a GIS is the database.

Map data are captured from a paper map through manual, semi-automatic or automatic means, and the outputs are in either raster or vector formats.

Raster format represents points, lines or areas using a matrix of values. The accuracy of a representation depends on the size or resolution of the individual grid cells (pixels). A point is a single cell; a line is several adjacent cells; and an area is an aggregation of cells. Each feature consists of sets of similarly numbered cells. In vector representation the points, lines and areas are produced from *x, y* coordinate pairs. A point is represented by a single coordinate pair, a line by a string of coordinate pairs, and an area by a string of coordinates that start and end at the same point.

Deciding on whether to use a raster or vector format depends on what you will do with the maps. Raster implementation is quick and easy and requires minimal training. It is reliable and supports high volume processing. However, raster maps have certain disadvantages. They are static; quality is lost when you zoom in; large storage space is required; and editing is limited to basic erase and redraw options. Raster maps are good for archiving and printing, like document imaging systems.

Many mapping applications use vector maps, but currently the tools for converting printed maps into vector format are inadequate, labor-intensive and very costly.

MapScan is a software package that accepts various formats of scanned maps or drawing, reads and converts scanned images into vector maps with text references of different formats that can be used by popular mapping systems. With MapScan users are able to move printed maps or drawing into a mapping system much more quickly and easily compared to using traditional digitizing techniques. MapScan has been developed in the framework of the UNFPA-assisted project INT/92/P23 "Computer Software and Support for Population Activities". MapScan is fully operational and can be used easily and successfully for small and large scan map vectorization.

2. MAIN FEATURES OF MAPSCAN

The main features of MapScan (see Figure 1) are:

a) **Scanning**: Scan the paper map and save the scanned map as a raster image.

Pre-processing or Raster Image Editing: edit the raster image to improve its quality, eliminate unnecessary items, connect broken lines, rotate an image, merge multiple pages to form the entire map image...

OCR (Optical Character Recognition): locate texts which are identifiers for regions, areas, cities, towns, ... ; perform OCR processing to identify the reference texts and determine their spatial coordinates; produce

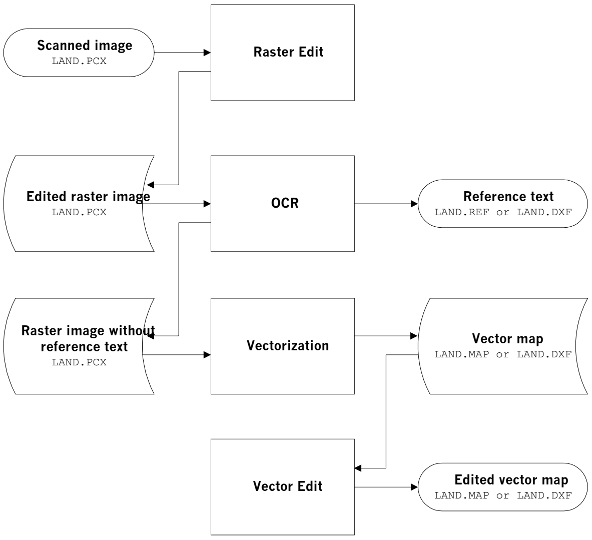


Figure 1. Flowchart showing MapScan's tasks

text reference file of appropriate format for specific mapping and GIS software; erase reference text from the raster image to avoid unnecessary vectorization of these items.
b) **Cleaning**: make additional editing of the raster image if necessary.

c) **Vectorization:** performs raster-to-vector conversion on the raster image that now contains only points, lines and polygons, and no reference text; generate map coordinate file in appropriate format for specific mapping and GIS software.
d) **Post-processing** (or Vector Map Editing): edit the generated vector map, close polygons, remove dirt, join line segments and assign specific layer attributes.

The reference text file and vector map file are then ready for use with the mapping system.

e) **Georeferencing.** In order to transformat a relative-coordinate map to real-world coordinate map.

2.1 Scanning

Scanner hardware can produce three kinds of output raster: binary, grayscale and color. Many scanners, whether color or black and white, allow the user to select a binary thresholding level as part of the scanning process. Binary scanning works best on black and white line drawings such as administrative maps.

Grayscale scanning should be used for all non-color materials that have intensity gradations for which the scanner's binary controls cannot handle easily. Grayscale is also better than binary for source material that has text labels that will be able to be read easily in the resulting raster. Color scanning should be used for materials that present the point and line information in a variety of colors. Color raster may be helpful in the interactive line following process, or in making color separation to isolate the particular lines.

2.2 Raster Editing

This feature of MapScan performs general raster editing and clean-up to remove unnecessary features such as annotations and noise before the automatic raster line vectorizing process. Raster editing could work with the black/white, grayscale and color images.

The raster editing environment (see Figure 2) allows to clean and repair a large raster object while scrolling over it. The process is fast and integrated, and provides many useful tools.

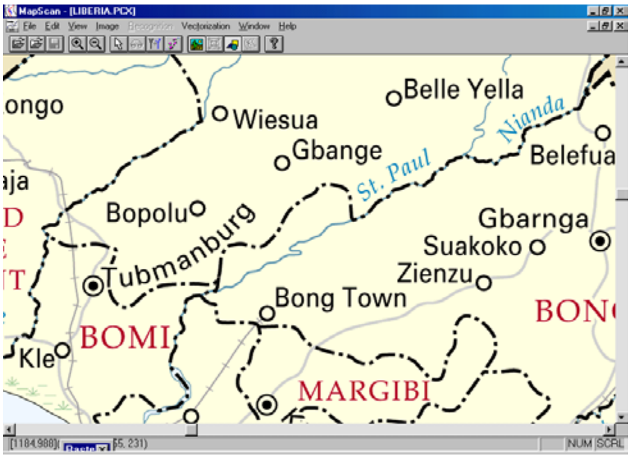


Figure 2. Raster Edit with black/white and grayscale images

- Cleaning-up to remove unnecessary features: it is able to cut and/or erase unnecessary items, draw additional items or correct broken lines, copy part of an image, zoom-in and out and rotate an image, etc. Two raster images can be merged to form a new raster image. This features is useful when scanning multiple pages of a large map.

- Improving the quality of the images: A histogram is a graphical representation of the distribution of gray and color values in an image. A gamma curve adjust midtone values in an image. It is necessary to change the gamma curve to correct different image bugs:

  – To correct for the distortion that occurs from the initial image source (for example, a camera) through other devices (scanners, monitors) to the final output source.

  – To make the image look better on the printed page. Usually this is done by increasing the gamma (that is, boosting the midtone) of the image.

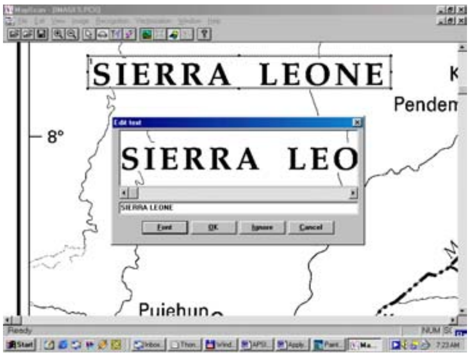
  There is two ways to adjust the gamma curve of a MapScan image:

  – Brightness & Contrast command on the Raster Editing submenu of the Edit menu make quick adjustments to the gamma value, but the user cannot see the actual gamma curve.

  – With the Gray/Color Map command on the Tune submenu of the Image menu, the user can create precise gamma curves. The user can also save the curves to a file for later use.

2.3 OCR

This module reads text labels on the image, recognizes and stores the ASCII text with coordinates into reference file. In the Figure 3, to perform an OCR, the user selects the text labels using the text box. The selection is interactive. The user places the mouse cursor at the upper left corner of the text area, then drags the mouse cursor to the lower right corner of the text area in order to cover the entire text.



SIERRA LEONE

Figure 3. OCR feature

Upon selecting text, user can perform OCR and recognize the marked text one by one. MapScan displays the raster image of the text box and the recognized ASCII text at the bottom of the screen. The recognized text and reference coordinates will be kept in the text reference file, and the text box will be erased from the raster image.

Once the text labels is processed, the corresponding area is erased. However, if this text label overlaps with line segments, MapScan will recognize the situation and keeps the overlapping parts intact. The Raster Edit functions help to correct the map.

2.4 Vectorization

MapScan accepts more than 30 input raster file formats, e.g. TIFF, PCX, BMP, GIF ... and produces vector map files in either MAP (PopMap) or DXF (AutoCAD) formats. There is an option to run a batch job to vectorize a list of raster images. MapScan allows to vectorize the lines in the center line and boundary line modes with both of automatic and interactive vectorizations.

Center Line: Extract lines by tracing the center of all line features.

Boundary Line: Extract lines by tracing the outer boundary of all features.
The options are suitable for black/white, grayscale and color images and for processing large images unattended.

2.4.1 *Automatic vectorization*

This feature performs the automatic vector extraction algorithm to detect all the lines from the image and display the vector data for verification and editing.

The task of automatic raster-to-vector conversion in the center line mode consists of three basic operations: skeletonization or line thinning, line extraction or vectorization and topology reconstruction.

- Line thinning is the process which automatically thins the lines in a raster object until they are uniformly just one cell wide. The line thinning process works from the edges of the line inward to the center, successively peeling off outside layers of cells. Thus, the thinned raster line represents the center-line of the wider lines of the original drawing and scanned image.
- Line extraction is the process of identifying a particular series of data entities or coordinates that constitute an individual line segment as portrayed on the input document. The process runs automatically. For input, it requires a raster object that contains lines that are continuous and uniform on pixel width, such as the output from the raster line thinning process. It is necessary to be sure that the boundaries of the mapping units form closed polygons, and that all unnecessary notation and text have been removed with the appropriate raster editing tools.
- The line extraction process creates unwanted vector elements corresponding to any such features that remain. If such vector elements should be accidentally produced, user can remove them later in the vector editing process. An automatic vector tolerance can be set as part of the vectorization step. The lower the tolerance, the more precise the vectorization, but more nodes will be generated for the vector output file.
- Topology reconstruction is the process of determining the adjacency relationships among the line segments. The individual line segments are joined into whole lines features and maps are build as continuous area representation.

2.4.2 *Automatic vectorization for grayscale and color images*

For a grayscale image, use threshold value to define the border of the regions to be vectorized (see Figure 4). The threshold can be set by Set Image Threshold option. The requirement for grayscale thinning are as follows:

If the object is connected, the result should also be connected; more generally, thinning guarantees that the topology structure of the object is not damaged.
The core line to be obtained should pass through the pixels with highest value, that is, ridge points, if grayscale values are viewed as altitudes.

- For a color image, supervised and unsupervised classification algorithms are used to generate a region boundary map.

For supervised classification: before starting the vectorizing, it is necessary to select an Area of Interest (AoI) to indicate the color feature to be used as a classifier to extract all the similar features from the image.

For unsupervised classification: the modification of K-means algorithm is used to classify the image into classes and then apply the line extraction and topology reconstruction. User specify number of classes based on the image and the number of iterations to run the algorithm. In the current version, the maximum number of classes to be used is limited to 32 and 500 is the limit for number of iterations. More iterations will increase the computing time but may improve the classification quality. To avoid excessive fragmentation in the final region map, the number of classes should not be specified too big. If the user want to generate a region boundary map, apply classification first and then use "Vectorize" with the option "boundary mode" to get vectors from the classified image. To get the boundary from one class, simply draw a selection rectangle within the class and then use "Vectorize" function.

With vectorization of the boundary mode, it is necessary to use only the line extraction and topology reconstruction functions.

2.4.3 *Interactive vectorization*

The vectorization is interactive. In this mode, user locates a section of line work on the scanned image, places the screen cursor on the feature to be digitized, then the software will take over and automatically follow the line feature by drawing along the raster line. The software generally stops and waits for operator input when encounter problem, such as reaching the end of the line or intersection.

User can save tracing results and continue the work in a later time. MapScan will load both the raster image and already vectorized line segments to facilitate tracing work.

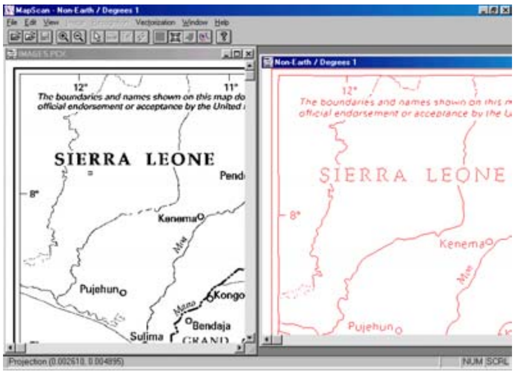

MapScan - Non-Earth / Degrees 1
File Edit View Vectorization Window Help
IMAGES.PCX
Non-Earth / Degrees 1
The boundaries and names shown on this map do not imply official endorsement or acceptance by the United
SIERRA LEONE
Kenema
Pujehun
Kongo
Bendaja
Sulima
Projection (0.002618, 0.004895)
NUM SCRL

Figure 4. Vectorizing feature

2.5 Vector Editing

After the vectorization process, the generated map contains line segments. The generated vector map might need to be edited for broken line segments, open polygons, and dirt. The user might also want to select line segments and assign specific layer attributes to them. The Vector Edit module shown in Figure 5 is for these post-processing purposes

<u>**Correct Cross**</u>: correct broken line intersections. The opened line segment is automatically connected to the closest line node producing a line junction.

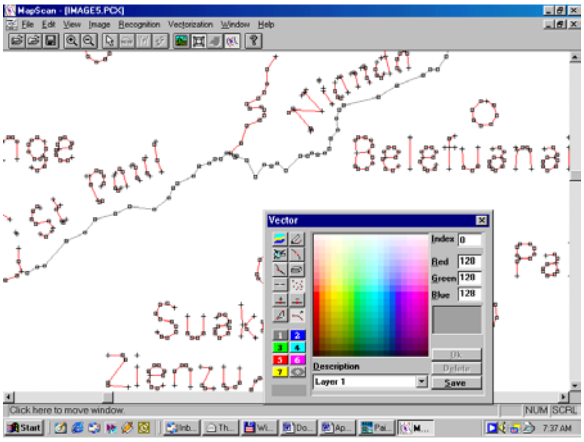


Figure 5. Vector Edit feature

Connect: connect gaps between two lines and nodes.
Join segments: joint to line segments.
Delete segment: delete unnecessary line segments.
Remove Dust: delete all the segments with the size less than tolerance inside selected rectangle.
Add new segments: add new segment if necessary.
Rotation: rotate vector map +90°, -90° or upside down.
Layer Assignment: Set the layer attribute for the line segments. There are 8 layers differentiated by 8 colors. The active layer is identified by its color in the Active box.
Merge*:* two vector maps can be merged to form a new vector map with the rubbersheeting algorithm.
Fill*:* to detect unclosed regions.

2.6 Georeferencing

After the vectorization process the map is in a relative coordinate system (plain *x,y*). The map is geo-referenced at this stage in order to convert the generated vector data into the real-world coordinate system (Latitude/Longitude) using known ground control points and information about the map projection.

3. APPLICATIONS

- GIS and Mapping: Topographic Contours.
- Automatic Input of mechanical drawing into CAD software.
- Logo, Art design and Font making.

4. REFERENCES

Bach Hung Khang, Ngo Quoc Tao, Luong Chi Mai, Do Nang Toan et al., "An examination of techniques for Raster-to-vector process and implementation of software package for automatic map data entry - MapScan ", *Journal of Computer Science and Cybernetics*, Vol. 12, No.2, 1996.

Ngo Quoc Tao et al., "An examination of techniques for raster-to-vector process and its implementation - MapScan software package", *Proceeding of AMPST'96*, Bradford, UK, 3/1996.

Donna J Peuquet, "An examination of techniques for reformatting digital cartographic data", *Cartographica,* Vol. 18, No.1, 1981, 34-48.

Pavlidis, T., "A thinning algorithm for discrete binary images", *Computer Graphics and Image Processing 13*, 1980, 142-157.

Wang P.S.P, Zhang Y.Y. A fast and flexible thinning algorithm, *IEEE Transaction on Computers*, Vol. 38, No. 5, 1989, 741-745

Bach Hung Khang, Ngo Quoc Tao, Luong Chi Mai, Do Nang Toan, Nguyen Duc Dung, Vu Van Thinh, "An examination of techniques for Raster-to-vector process and implementation of software package for automatic map data entry - MapScan ", *Journal of Computer Science and Cybernetics*, Vol.12, No.2, 1996.

US. Department of the Interior, "*GCTP General Cartographic Transformation Package*", September, 1990.