RECOGNITION AND DESCRIPTION OF TOPONYM IN KANNADA CARTOGRAPHIC MAPS

[1]Jagadishappa B Akkimaradi , [2]Jagadish S Kallimani

[1,2] Department of Computer Science and Engineering

M S Ramaiah Institute of Technology, Bangalore, (India)

ABSTRACT

Graphical documents such as cartographic maps contain great variety of textual elements appearing in different spatial positions, in different sizes and colors and overlapping graphical symbols. This greatly complicates automatic optical recognition of textual elements in the process of raster-to-vector conversion of graphical documents. At present a lot of methods and programs for automatic text recognition exist. However there are no effective text recognition systems for graphic documents. Graphic documents usually contain a great variety of textual information. There are no sufficient number of works on Indian language character recognition especially Kannada script among twelve major scripts in India. The goal is to present a review of existing work on printed Kannada script and their results. The characteristics of Kannada script and Kannada Character Recognition System KCR are discussed in detail. Finally fusion at the classifier level is proposed to increase the recognition accuracy. This dissertation work is an effort to toponym recognition in cartographic maps of Kannada language.

Keywords: Cartography, Toponyms, Segmentation, Template Matching

I INTRODUCTION

A huge amount of geographic information collected in the last centuries is available in the form of maps printed or drawn on paper. Currently, Google maps serves as an excellent tool in retrieving almost every details from different types of maps. To store, search, distribute, and view these maps in the electronic form they are to be converted in one of the digital format developed for this purpose. The simplest way of such a conversion is scanning the paper map to obtain an image (a picture) stored in any of the raster graphical formats such as TIFF, GIF, etc. After that, a raster-to-vector conversion may be applied to include obtained vector maps into Geographic Information Systems (GIS).

Cartography is the study and practice of making maps. Combining science and technique, cartography builds on the premise that reality can be modelled in ways that communicate spatial information effectively.

The fundamental problems of traditional cartography are to

- Set the map's agenda and select traits of the object to be mapped. This is the concern of map editing. Traits may be physical, such as roads or land masses, or may be abstract, such as toponyms or political boundaries.
- Represent the terrain of the mapped object on flat media. This is the concern of map projections.

- Eliminate characteristics of the mapped object that are not relevant to the map's purpose. This is the concern of generalization.
- Reduce the complexity of the characteristics that will be mapped. This is also the concern of generalization.
- Orchestrate the elements of the map to best convey its message to its audience. This is the concern of map design.

Modern cartography is largely integrated with geographic information science (GIScience) and constitutes many theoretical and practical foundations of geographic information systems.

The earliest known map is a matter of some debate, both because the definition of "map" is not sharp and because some artifacts speculated to be maps might actually be something else. In cartography, technology has continually changed in order to meet the demands of new generations of mapmakers and map users. The first maps were manually constructed with brushes and parchment; therefore, varied in quality and were limited in distribution. The advent of magnetic devices, such as the compass and much later, magnetic storage devices allowed for the creation of far more accurate maps and the ability to store and manipulate them digitally.

The separation of overlapping text and graphics is a challenging problem in document image analysis. This problem is found in many applications, including forms processing, maps interpretation and engineering drawings interpretation, where text and graphics are processed in fundamentally different ways. [1]

The problem of text extraction from an imaged document still remains an important issue in the field of image processing. Applications such as map interpretation, referencing system for digitized manuscripts, and news article search from microfilms require some form of text extraction and categorizing of text images into logical groups.

Document Processing in the Indian environment has special significance, since eighteen official languages are in use in the country. Throughout the country, every government office uses at least two languages, English and the official language of the corresponding state. The state of Karnataka has an official language as Kannada, however many national organizations such as Banks, use English and Kannada. Even all the documents in the government offices of Karnataka state usually appear in two languages, Kannada and English. The aim of the automation of document processing is to convert the scanned paper document to the machine readable codes such as ASCII.

Computer vision research today is still facing greater difficulties in understanding graphical objects than character strings which can be handled using OCR with reasonable success. Any words that can be found in a document image by the computer will certainly assist in the understanding of the image. One major technique reported in the literature for text/graphics separation relies on connected component analysis through examination of the relationships among neighbouring components.

Advances in mechanical devices such as the printing press, quadrant and vernier, allowed for the mass production of maps and the ability to make accurate reproductions from more accurate data. Optical technology, such as the telescope, sextant and other devices that use telescopes, allowed for accurate surveying of land and the ability of mapmakers and navigators to find their latitude by measuring angles to the North Star at night or the sun at noon.

Advances in photochemical technology, such as the lithographic and photochemical processes, have allowed for the creation of maps that have fine details, do not distort in shape and resist moisture and wear. This also eliminated the need for engraving, which further shortened the time it takes to make and reproduce maps.

Advancements in electronic technology in the 20th century ushered in another revolution in cartography. Ready availability of computers and peripherals such as monitors, plotters, printers, scanners (remote and document) and analytic stereo plotters, along with computer programs for visualization, image processing, spatial analysis, and database management, have democratized and greatly expanded the making of maps. The ability to superimpose spatially located variables onto existing maps created new uses for maps and new industries to explore and exploit these potentials.

These days most commercial-quality maps are made using software that falls into one of three main types: CAD, GIS and specialized illustration software. Spatial information can be stored in a database, from which it can be extracted on demand. These tools lead to increasingly dynamic, interactive maps that can be manipulated digitally.

With the field rugged computers, GPS and laser rangefinders, it is possible to perform mapping directly in the terrain. Construction of a map in real time, for example by using Field-Map technology, improves productivity and quality of the result.

The source maps are from a local Street Directory. The maps are printed in color. The focus here is on the black layer, which includes text, small icons, road lines, outlines of buildings and parks, and so on. Some icons in the maps are solid components, such as solid arrowheads, solid overhead pedestrian bridge. The road lines are either thick or thin lines, mostly curves. Some tracks and outlines of parks are dashed lines. Most outlines of buildings and parks are polygons.

The text strings in the maps are mainly road names, names of buildings and parks, and so on. The road names are labeled along the roads. That means the labels may be along curved lines. The touching of road names with road lines is somewhat similar to the case of strings touching the underlines or upper lines. On the other hand, the connection of the names and the outlines of the buildings are much more complicated. The reason is that the outlines of the buildings can be of any shapes.

II RELATED WORKS

Text detection and recognition are very difficult problems which have generated a lot of research work. The main difficulty lies in the graphics that surround or intersect the text. The main feature of a good text detection algorithm is to be independent in respect to text font, size and orientation. One can distinguish two main families of methods: (i) methods devoted to detect text in paper-based documents, (ii) methods developed for text detection in videos or textured images. The later methods are mostly based on texture analysis. [3]

An early approach in [4] introduced an algorithm which permits to extract text on binary images such as electronic circuits. Using connected components, a selection of possible characters is performed using the components' characteristics. The characters are merged into words using Hough Transform. This method, which inspired our paper, gives good results and has the advantage to be usable for various orientations and sizes of texts.

In [5], authors develop a model to extract black characters in urban maps. They use street lines data to recover the characters composing street names. Due to the possible overlapping between street names and street lines describing the street, they develop a specific OCR algorithm to produce an efficient text file of street names.

The move from paper-based documentation towards computerized storage and retrieval systems has been prompted by the many advantages to be gained from the ‘‘electronic document” environment. Document update and revision is efficiently achieved in the computerized form. For efficient processing and storage of documents, however, it is necessary to generate a description of graphical elements in the document rather than a bit-map in order to decrease the storage and processing time. Thus, increasing emphasis is being placed on the need for the realization of computer-based systems which are capable of providing automated analysis and interpretation of paper-based documents**.** Much of the attention paid to automated document analysis systems in the literature has been in relation to engineering drawings and diagrams**.** Such systems provide a means for originating technical information (text and graphics) in a digital form suitable for interactive graphics editing, reproduction, and distribution. [4]

The initial processing stage in an automated document analysis system requires conversion of paper-based graphics/text to a digital bit-map representation. Wherever the primary goal of the automated document analysis system is interpretation of graphic data, text strings present within the digitized document must first be separated from the graphics in order that subsequent processing stages may operate exclusively on the graphic information. The extracted text may be stored separately for input to a character recognition system for later retrieval or revision. Since document types vary widely in style and content of both graphic and text data, an algorithm to perform text string removal must be able to accommodate documents containing text of various font styles and sizes.Further, the documents may contain text strings which are intermingled with graphics, and text characters which are similar in size or shape to graphics. In general, text strings may be of any orientation in the image; not simply horizontal or vertical but possibly diagonally aligned.

Several algorithms for text string separation have been reported in the literature [6][7]. However, many of these algorithms are very restrictive in the type of documents they can process and are therefore not useful in a general automated document analysis system. The Bley algorithm [7] is also sensitive to variations in text font style and size; the algorithm breaks connected components into subcomponents, which makes it difficult to process the components for graphics recognition. Thus there is no single algorithm which is robust enough to segment images containing mixed graphics and text, with multiple font styles and sizes and strings of arbitrary orientation.

In [8], proposed a fast algorithm for speeding up the process of template matching that uses M-estimators for dealing with outliers. In this a particular image hierarchy called the -pyramid that can be exploited to generate a list of ascending lower bounds of the minimal matching errors when a non decreasing robust error measure is adopted. Then, the set of lower bounds can be used to prune the search of the -pyramid, and a fast algorithm is thereby developed in this paper. This fast algorithm ensures finding the global minimum of the robust template matching problem in which a non decreasing M-estimator serves as an error measure. Experimental results demonstrate the effectiveness of our method. The basic idea of the M-estimator technique is to limit the

influence of outliers in the matching error. In principle, the effects of the outlier can be suppressed with the M-estimator technique and therefore better estimations are obtained.

Text recognition from documents that contain non homogeneous text, such as from raster maps [6], is a difficult task, and hence much of the previous research only works on specific cases. Fletcher and Kasturi [4] utilize the Hough transformation to group characters and identify text strings. Since the Hough transformation detects straight lines, their method cannot be applied on curved strings. Moreover, their work does not handle multi-sized characters.

Vel´azquez and Levachkine [2] and Pal et al. [9] present text recognition techniques to handle characters in various font sizes, font types, and orientations. Their techniques are based on detecting straight string baselines for identifying individual text strings. These techniques cannot work on curved strings.

In our previous work [10], a text recognition approach that locates individual multi-oriented text labels in raster maps and detects the label orientations to then leverage the horizontal text recognition capability of commercial OCR software. The previous work requires manually specified character spacing for identifying individual text labels and does not consider multi-sized characters. In this paper, we present a text recognition technique to dynamically group characters from non-homogeneous text into text strings based on the character sizes and maximum desired string curvature. The hypothesis is that characters in a text string are similar in size and are spatially closer than the characters in two separated strings. Our text recognition technique does not require training for specific fonts and can be easily integrated with a commercial OCR product for processing documents that contain non-homogeneous text. [10]

III METHODOLOGY

The main objective of our method is to generate a set of images containing potential strings by detecting them on the map and to match the templates in order to obtain a consistent vectorized toponyms data file. We discuss how such additional information can be used to work around the problems arising in recognition of the inscriptions in the maps, associating them with specific cartographic objects, and importing information on these objects from available databases.

The following are the list of responsibilities involved in this dissertation work:

- A detailed survey and study on various types of cartographic maps in Kannada language.
- Develop a dictionary of all places, rivers, mountains, etc, for a limited geographical area maps.
- Develop an algorithm to recognize and describe the various features of the identified regions.
- Conduct subjective and objective test result analysis.

3.1 Segmentation

Segmentation is the process of extracting objects of interest from an image. The first step in segmentation is detecting lines.The first step of our technique consists in a binary classification of each pixel in order to eliminate pixel values that can definitively not be a part of a character.

A raster map has to be segmented first. All its elements should be retrieved with their coordinates and features, and then sent to the corresponding thematic layers. The layers can be symbols, landmarks, isolines, natural and artificial surroundings, words and numbers, lakes and other "punctual", "linear" and polygonal bodies. Cartographic maps are the most complex graphic documents due to the high density of information that they contain. [2]

3.2 Template Matching Approach

Template matching is one of the Character Recognition techniques. It is the process of finding the location of a sub image called a template inside an image. Once a number of corresponding templates is found, their centers are used as corresponding points to determine the registration parameters. Template matching involves determining similarities between a given template and windows of the same size in an image and identifying the window that produces the highest similarity measure. It works by comparing derived image features of the image and the template for each possible displacement of the template. This process involves the use of a database of characters or templates. There exists a template for all possible input characters.

In this method the dictionary is created for storing the Kannada toponyms. And this is used to compare with recognized toponym. The dictionary contains toponyms with their description that contains the information like name of river, city, etc

Template of all the characters to be recognized by the system is formed. A type of similarity measure is performed between the test character and the templates. At present, template matching is popularly used algorithm in character recognition. This method is efficient and has high speed when dealing with character identification. In the proposed method, recognition is done using template matching.

IV RESULTS

The accuracy of segmentation module of the proposed method has been checked and the results are obtained as per the below table and figure. It is evident that majority of the names are identified correctly. This shows that the method adopted is reliable and can be enhanced in future.

Table 1: Detection percentage of samples

SI No	Districts Names	Segmentation Accuracy
1	ಬೀದರ (Bidar)	100%
2	ಬೆಂಗಳೂರು (Bangalore)	100%
3	ಬೆಂಗಳೂರು ಗ್ರಾಮಾಂತರ (Bangalore Rural)	100%
4	ಕೋಲಾರ (Kolar)	100%
5	ಚಿಕ್ಕಬಳ್ಳಾಪುರ (Chikkaballapur)	85%
6	ರಾಮನಗರ (Ramnagar)	100%
7	ಗದಗ (Gadag)	100%
8	ಬೆಳಗಾವಿ (Belgaum)	100%

9	ಬಿಜಾಪುರ (Bijapur)	100%
10	ಹಾವೇರಿ (Haveri)	100%
11	ಧಾರವಾಡ (Dharwad)	100%
12	ಬಾಗಲಕೋಟ (Bagalkot)	100%s
13	ಉತ್ತರ ಕನ್ನಡ (Uttar Kannada)	98%
14	ಶಿವಮೊಗ್ಗ (Shimoga)	100%
15	ಬಳ್ಳಾರಿ (Bellary)	100%
16	ಗುಲ್ಬರ್ಗ (Gulbarga)	100%
17	ಕೊಪ್ಪಳ (Koppal)	100%
18	ರಾಯಚೂರು (Raichur)	100%
19	ಯಾದಗಿರಿ (Yadgir)	100%
20	ಚಾಮರಾಜನಗರ (Chamarajnagar)	85%
21	ಚಿಕ್ಕಮಗಳೂರು (Chikmagalur)	100%
22	ದಕ್ಷಿಣ ಕನ್ನಡ (Dakshin Kannada)	100%
23	ಹಾಸನ (Hassan)	100%
24	ಕೊಡಗು (Kodagu)	100%
25	ಮಂಡ್ಯ (Mandya)	100%
26	ಮೈಸೂರು (Mysore)	100%
27	ಉಡುಪಿ (Udupi)	100%
28	ಚಿತ್ರದುರ್ಗ (Chitradurga)	100%
29	ದಾವಣಗೆರೆ (Davangere)	98%
30	ತುಮಕೂರು (Tumkur)	100%

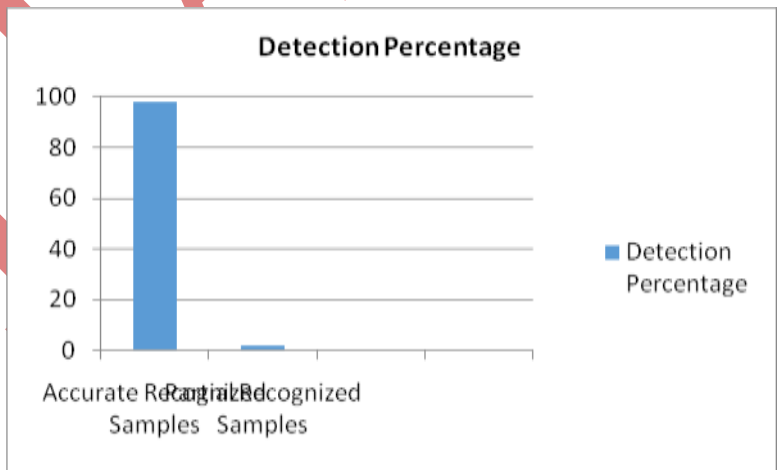

Figure 1: Graph of Detection Results.

V CONCLUSION

An automatic method to extract and recognize the textual layer of scanned color topographic maps has been described. The proposed method is mainly based on image segmentation and connected component processing. At each step, pruning or corrections are done using empirical filtering. A new algorithm for text string separation from mixed text/graphics images has been presented. The algorithm is robust to changes in text font style and size within an image. The algorithm presented has several advantages over previously applied techniques. With improvements made to the algorithm in terms of processing speed and efficiency of data representation, the application of the algorithm will become highly appropriate for use in document analysis systems.

REFERENCES

[1]. Detection and Extraction of Text Connected to Graphics in Maps

[2]. Vel´azquez, A. and Levachkine, S. (2004). Text/graphics separation and recognition in raster-scanned color cartographic maps. In GREC, vol 3088 of LNCS, pages 63–74.

[3]. Pouderoux, J., Gonzato, J. C., Pereira, A., and Guitton, P. (2007). Toponym recognition in scanned color topographic maps. In Proceedings of the 9th ICDAR, volume 1, pages 531 -535.

[4]. L. Fletcher and R. Kasturi. A robust algorithm for text string separation from mixed text/graphics images. IEEE Trans Pattern Anal. Mach. Intell., 10(6):910–918, 1988.

[5]. L. Li, G. Nagy, A. Samal, S. Seth, and Y. Xu. Cooperative text and line-art extraction from a topographic map. In Proceedings of ICDAR ’99, pages 467–470, 1999.

[6]. L. T. Watson, K. Arvind, A. W. Ehrich, and R. M. Haralick, “Extraction of lines and regions from grey tone line drawing images,” Pattern Recognition, vol. 17, pp. 493-506, 1984.

[7]. H. Bley, “Segmentation and pre-processing of electrical schematics using picture graphs,” Cornput. Vision, Graphics, Image Processing, vol. 28, pp. 271-288, 1984.

[8]. Fast Algorithm for Robust Template Matching With M-Estimators Jiun-Hung Chen, Chu-Song Chen, and Yong-Sheng Chen IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 51, NO. 1, JANUARY 2003

[9]. Pal, U., Sinha, S., and Chaudhuri, B. B. (2003). Multioriented english text line identification. In Proceedings of the 13th Scandinavian conference on Image analysis, pages 1146–1153.

[10]. Chiang, Y.-Y. and Knoblock, C. A. (2010). An approach for recognizing text labels in raster maps. In Proceedings of the 20th ICPR, pages 3199–3202.