Tamil Gnani - an OCR on Windows

Aparna K G and A G Ramakrishnan

Biomedical Laboratory, Department of Electrical Engineering
Indian Institute of Science, Bangalore - 560 012
e-mails: {prjocr, ramkiag}@ee.iisc.ernet.in

ABSTRACT

A complete working model of Optical Character Recognizer for Tamil script is developed. The system works in a multi-font and multi-size scenario. The input to the system is a scanned or digitized document and the output is in TAM code. The basic techniques were presented in Tamilnet 2000 [1]. Now, we have added the recognition of other symbols, such as punctuations and numerals. Further, the entire scheme, from scanning to obtaining the TAM codes, has been implemented on Visual C++ platform. The product is designed to run on Windows 95 and 98 platforms. The current overall recognition rate is around 98%.

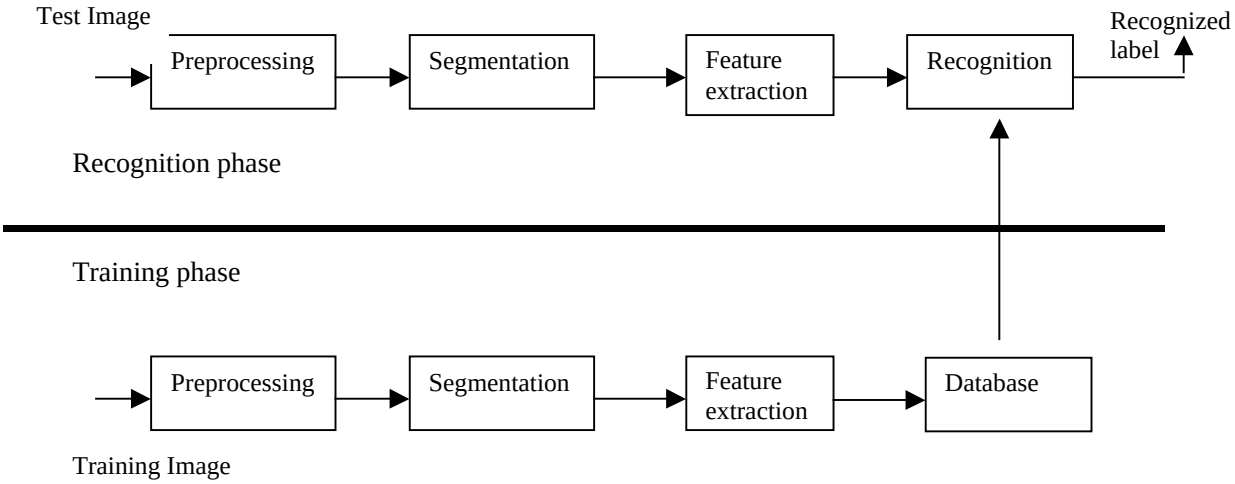

Fig. 1. Block Diagram of the OCR System.

The first step in the whole process is to scan the printed text page and convert it into a digital image. We scan the document at 300 dots per inch (dpi). An option has been provided in the product for scanning. Once the digital document is obtained, the recognition process proceeds as shown in the block diagram in Fig. 1. This has two phases - the training phase and the recognition phase.

PREPROCESSING

This involves binarisation, skew detection and skew correction. Binarisation is the process of converting the gray level image into a binary image with foreground as white and background as black.

The skew may be caused while placing the paper on the scanner, or may be inherently present in the paper. Even with lot of care, some amount of skew is inevitable. For skew detection, algorithm proposed by Kaushik et al [2] is employed. An estimate of the skew angle is found to an accuracy of ± 0.06^0.

After finding the skew angle, we need to correct the skew. While skew detection is performed on the binarized document, correction, which involves rotating the image in the appropriate direction, is performed on the gray scale image to reduce the quantization effects. The whole process of skew detection and correction on an A4 sized paper containing text of font size 16, takes about 24 seconds on a 500 MHz Pentium 3 processor with 128 MB RAM.

SEGMENTATION

Segmentation involves breaking the text in the page to lines, words and then characters. Horizontal projection profile is employed for line detection and vertical projection profile is employed for word detection. Connected component analysis is performed to extract the individual characters. The segmented characters are normalised before the recognition phase.

FEATURE EXTRACTION AND RECOGNITION

Depending on the spatial spread of the characters in the vertical direction, they are grouped into 4 classes. These classes are further divided into groups based on the type of ascenders and descenders in the characters. Second order moments are employed as features to perform this grouping. Block Discrete Cosine transform (DCT) based features are used for the final classification. Nearest neighborhood classifier is employed. The recognised characters are then stored in TAM codes.

DATABASE

In order to obtain good recognition accuracy, we have created a vast database. Each character has around 25 to 50 samples collected from various magazines, novels, technical papers and from various Tamil shloka books. The total database exceeds more than 4000 samples. This database also includes bold characters and italics. Some of the Tamil editors like Kamban, Murasu Anjal, iLEAP etc. provided some standard fonts like TM-TT Valluvar, TAB_Arulmathi, Inaimathi, TM-TT Bharathi and TAM-Aniezhai, which are included in the database. Some of the special characters like comma, semicolon, colon and numerals are also present in the database. We have handled font sizes from 14 to 20 in testing the system.

IMPLEMENTATION DETAILS

The whole product "Tamil Gnani" is implemented using Visual C++ and it is in dialog-based format. This is used to provide graphical user interface for Windows. The codes are written in C++ and the Graphic User Interface (GUI) is provided with the help of Visual C++. The brief outline of the product is as shown below. The recognized characters are stored using the TAM format.

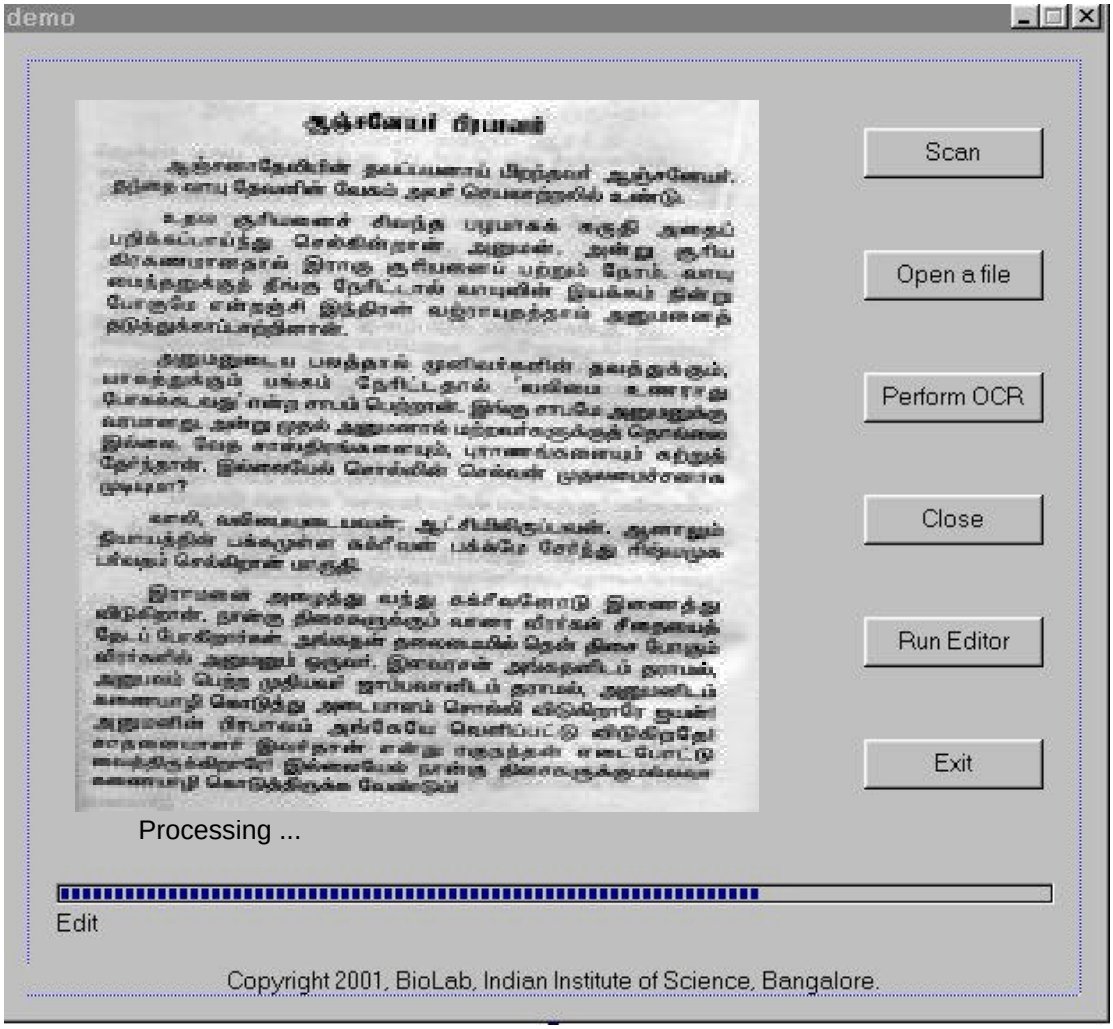

Fig. 2. User Interface Window of Tamil Gnani

CONCLUSION

The product has been tested on different fonts and the recognition rate is around 98% with the presence of some special characters and numerals. The time taken for recognition depends on the number of characters in the document. It may take around 2 minutes on a scanned A4 paper

containing around 1200 characters on a 500 MHz Pentium III machine with 128 MB RAM. Efforts are on to make this product run on Windows NT also.

OPTIONS	DESCRIPTION
Scan	The scanning operation is independent of any scanner. Once the scanning process is completed, the corresponding scanned image will be displayed. The scanned document will be in BMP format.
Open a file	This option is provided in order to avoid scanning when a scanned document is already present. The user can select the necessary document in the displayed explorer window. The file extensions of the images for which recognition is to be done should be either BMP or raw
Perform OCR	It initiates the process of Optical Character Recognition
Close	This is to terminate the current process without exiting the application
Run Editor	An in-built editor will be provided in the software, which displays the recognized text
Exit	This is to quit the application

ACKNOWLEDGEMENT

We thank the Department of Information Technology, Ministry of Information Technology, Government of India for funding part of this project under the Indian Language Technology Solutions Initiative (Ref. No. 21(3)/2000-CDD-TDIL0.

REFERENCES

1. Kaushik Mahata and A. G. Ramakrishnan, "A complete OCR for printed Tamil text ", Proc. Tamil Internet 2000, Singapore, July 22-24, 2000, pp. 165-170.
2. Kaushik Mahata and A. G. Ramakrishnan, "Precision Skew Detection through Principal Axis", Proc. Intern. Conf. on Multimedia Processing and Systems, Chennai, Aug. 13-15, 2000, pp. 186-188.