Bilingual (Tamil – Roman) Text Recognition on Windows

Aparna K G, Dhanya D 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

Introduction

Optical Character Recognition (OCR) is the machine recognition of characters in a document image, obtained by scanning printed text on paper. In the Indian scenario, documents are often bilingual in nature. English, being the unofficial link language in India, is used along with the regional language in most of the important documents, reports, magazines and technical papers. Monolingual OCRs fail in such context and thus, there is a need to extend the operation of current monolingual systems to bilingual ones. We have developed an OCR system that recognizes multi-font, multi-size Tamil and English scripts. The basic techniques were presented in Tamilnet 2001 [2]. The entire scheme, from scanning to output, has been implemented on Visual C++ platform and is designed to run on Windows 95 and 98 platforms. The bilingual OCR system "MOZHI VALLAAN", which presently gives a recognition accuracy of around 94%, shall be demonstrated during the conference. Analysis of the input document to detect features, such as alignment, justification, font size and line spacing, has been satisfactorily accomplished.

Description of our Bilingual OCR system

The design of an OCR system capable of recognising bilingual text is a very challenging task. There are two ways of recognising a bilingual document. The first one is to initially identify the script and then pass on sections of the test image to the individual OCR's; the other one is a combined database approach. But script identification can be very tedious. In the bilingual OCR system that we have designed, combined database approach is employed. Both English and Tamil characters are handled similarly irrespective of the script they belong to. To the author's knowledge, this is the first system that recognises both Tamil and Roman text.

The input documents obtained from various magazines are scanned at 300 dots per inch (dpi). The basic steps like skew detection [3], correction [4] and segmentation remain similar to monolingual OCR explained in [1]. The features used are geometric moments and discrete cosine transform coefficients. A hierarchical classification scheme is designed wherein all the gross discriminations are made first postponing the subtle ones to later stages. The use of such a classification scheme has a lot of advantages like

- Reduced number of classes at each level.
- Ease of choice of features and hence the possibility of using simpler features.
- Higher processing speed and lower measurement cost since the number of features required to classify a test sample is less.
- Reduction in computational complexity.

The proposed system uses a three level hierarchical scheme as shown in Fig. 1

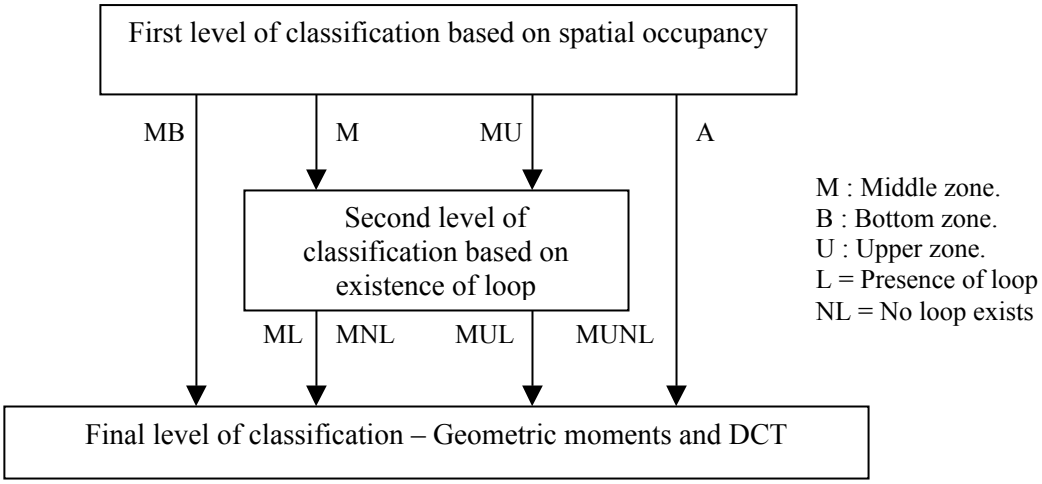

Figure 1: Hierarchical classification scheme

In the first level, characters are classified into one of the 4 groups - middle (M), middle-bottom (MB), middle-upper (MU) and all three (A), depending on their spatial spread in the vertical direction [2].

In the second level, the characters are normalised to different predefined sizes depending on the group they belong to. The characters are then thinned and presence or absence of loop is detected. This is performed for groups M and MU only. A contour-tracing algorithm is used for this purpose [4].

In the final level, feature [6] based classification is performed. Features such as geometric moments and discrete cosine transform coefficients are employed. Classification is performed with nearest neighbour classifier, based on Euclidean distance. TAB (Tamil Bilingual) code is assigned to the recognised characters.

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 Tamil and English books. The total database exceeds more than 6000 samples. This database also includes bold characters, italicized characters along with numerals and few keyboard characters. We have handled font sizes from 14 to 20 in testing the system.

Document Analysis

Obtaining visual similarity between the input document image and the recognised document in terms of its font size, line spacing and justification is a very important step in any character recognition system. The user will prefer an output document that is an exact replica of the input image. In order to accomplish this, the document is graphically analysed and the recognised format is stored in the Rich Text Format (RTF). RTF files are actually ASCII files with special commands to indicate formatting information, such as fonts and margins. The RTF specification is also a method of encoding formatted text and graphics for easy transfer between applications. Simple algorithms are used to find the font size, text line gap and text alignment.

Implementation Details

The system is designed to work on Windows 95 and 98. It is designed using C++ and graphic user interface (GUI) is provided using Visual C++. For a scanned A4 page containing around 1200 characters, it approximately takes two minutes on a 500 MHz Pentium III machine with 128 MB RAM. The GUI of our software is shown in Fig 3.

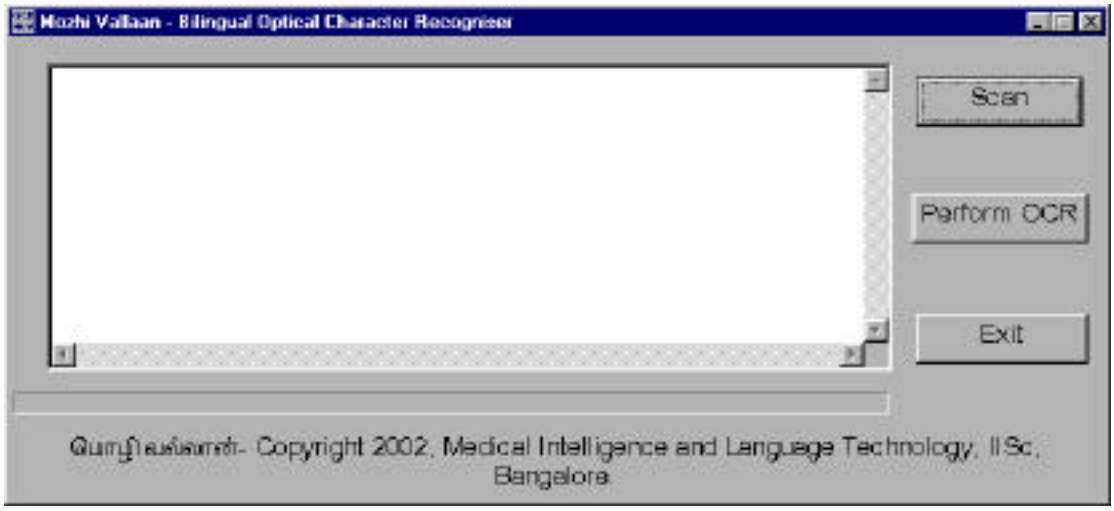

Figure 3. GUI of our software

There are only 3 buttons in the software developed to make it easier for the user. The description about each button is given below.

Options	Description
Scan	The scanning operation is independent of the scanner. The software detects the scanner. Once the scanning is done, recognition process starts automatically.
Perform OCR	This option is to process any document image already available, without the need for scanning. The user can select the document in the displayed explorer window. The file extensions for the images should be BMP (Bitmap), RAW or PGM (Portable Gray Map) formats only.
Exit	This quits the application.

Results and Discussion

The system is being tested on files obtained from various magazines, newspapers and books containing variable font styles and sizes scanned at 300 dpi for digitizing the documents. The training set consists of more than 60 samples per pattern. Results on a set of 30 test documents (disjoint from the Training Set) are discussed below. The overall recognition accuracy on the above test set is around 95%. A sample input document and the corresponding recognised output are shown in Fig 2.

இரண்டு பக்கமும் ஜெயம்தான்'
என்றேன். :
வெற்றிகரமான வறக்கறிஞராக
இருந்துள்ள நீங்கள், அரசியல் சாசன
பரிசிலனைக் கமிட்டிமில் இடம்
பெற்றுள்ரிகள்t அதன் அதிகார
வரம்பு என்ன?
அடிப்படையில் இந்தக் கமிட்டி
பரிந்துரைக்கும் தன்மை கொண்டது.
இதன் பரிந்துரைகளை நாடாளுமன்றம்
பரிசிலித்து எந்த முடிவையும் எடுக்கும்.
இந்தக் கமிட்டி முடிவு எடுக்கும்
அதிகாரம் கொண்ட அமைப்பல்ல.
என்னென்ன விஷயங்களை
கமிட்டி பரிசிலனைக்கு எடுத்துக்
கொள்ளும் ?
இன்னமும் Terms of reference
கைக்கு வரவில்லை. எனவே அது பற்றி
ஏதும் கூறக்கூடாது.
இந்தக் கமி குறித்து ஒவ்வொரு
பத்திரிகையும் ஒவ்வொரு விதமாக
எழுதுகிறது. ஒரு நாளிதற்
ConsLtution review panel என்கிறது.
மற்றொரு இதர், Constitution review
committee என்கிறற. இன்னொரு
நாளிதழ், Commission to review the
constitution என்கிறது. எது சரி?
இப்போது சொன்னேன் அல்லவா'.
Terms of reference வரலில்வை
என்று. இந்நிலையில் நான் எது சரி

Figure 2. Input document and recognised document

Output Details

Total Tamil characters = 502
Total English characters = 115
Total Special characters = 20
Mis-recognised (Tamil) = 10
Mis-recognised (English) = 3
Mis-recognised (Special) = 1
Recognition accuracy (Tamil) = 98%
Recognition accuracy (English) = 97%
Recognition accuracy (Special) = 95%
Overall Recognition = 97.8%
Rejected ("~" symbol) = 1

· Skew detection [3] is performed in 2 steps. The first step (coarse skew) detects the angle to an accuracy of ± 0.25° and the second step (fine skew) to the accuracy of 0.06°. Such a high skew accuracy is necessary because the first level of grouping in our hierarchical classification scheme is based on the spatial occupancy of the characters. If the skew is not detected to such a high accuracy then characters gets misgrouped in the first level itself and hence will be recognised wrongly. By

including fine skew detection, the recognition accuracy increased by 2% - 10% on various documents

- Skew correction [5] is performed on gray scale image rather than on binary image to reduce quantization effects. When tested on various input documents, this rotation of the gray level image results in an increase of recognition accuracy by a factor of 2% - 15% over the value when rotation was performed on a binary image. Binary rotation introduces cuts in the character, especially if the font size is small, which affects the recognition accuracy.
- The recognition accuracy without post processing is around 87%. By analysing the confused characters, we are able to obtain an appreciable improvement to around 94%. Some of the post processing techniques used are shape analysis and certain grammatical rules.

Speciality of the software

- The software works efficiently for multicolumn images also i.e. presence of more than one column in the same page.
- Ability to access any scanner with the same software.
- Document analysis done to make the output visually similar to the input.
- The output or the recognised document saved as Rich Text Format (RTF).

Conclusion

The package Mozhi Vallaan has been tested on different fonts. Attempts have been made to make the output (recognised document) very similar to input document visually. The overall recognition rate is around 94% with the presence of some special characters and numerals. A hierarchical classification scheme has been followed.

References

[1] K G Aparna and A G Ramakrishnan, "Tamil Gnaani - A complete Tamil OCR on windows", *Proc. Tamil Internet 2001*, Kuala Lumpur, Malaysia, Aug. 26-28, 2001, pp. 60-63.

[2] D Dhanya and A G Ramakrishnan, "Simultaneous Recognition of Tamil and Roman Scripts", *Proc. Tamil Internet 2001*, Kuala Lumpur, Aug 26-28, 2001, pp. 64-68.

[3] Kaushik Mahata and A G Ramakrishnan, "Precision Skew Detection through Principal Axis", *Proc. Intern. Conf. Multimedia Processing Systems*, Chennai, Aug. 13-15, 2000, pp. 186-188.

[4] R C Gonzalez and R E Woods, Digital Image Processing. Addison – Wesley, Massachusetts, 1993.

[5] D Dhanya, A G Ramakrishnan and Peeta Basa Pati "Script Recognition in Bilingual Documents", Sadhana, Vol. 27 Part I, pp. 73-82, Feb. 2002.

[6] O Trier, A K Jain and T Taxt, "Feature extraction methods for character recognition – a survey" Vol. 29. Pattern Recognition, pp. 641-662, 1996.