MALAYALAM OCR

ABSTRACT

The project aims to develop Malayalam optical Character Recognition software, mOCR. Using the software, a `scanned image of the printed/handwritten Malayalam document ` can be converted into a computer editable text file. This will avoid cumbersome task of typing it again.

The printed/handwritten document is first scanned into an image file of any format (preferably Bitmap file). Next the user will load the image file into the mOCR and mOCR will recognize the glyphs in the document and save it as a computer editable text file. Correcting the errors or editing the text content in the recognized text will be made possible in mOCR. There will be provision to concatenate the recognized text into a large file, so that it may be in the form of a book.

There will be facility to recognize text from documents having colored text. Facilities to correct errors using a dictionary, to recognize text from documents having embedded images, remove salt and pepper noise from the scanned image and correcting the skew of the scanned image can be incorporated later mOCR can be further integrated as a plug-in into a document processing software like Microsoft Word.

For the recognition lines and words are isolated based on the distance between the glyph groups. Then the glyphs in each word are isolated by identifying the pixel groups. The seven `Hu moments` of each glyph is then extracted and these moments are matched with the library of feature. Character having the highest degree of match will be then the recognized character.

Feature Library of each font can be created by a training process. A scanned image of alphabet or a font file can be given as the training data.

The potential of OCR systems is enormous in situations like digitalization of old books in library which got no electronic copies and can't be preserved. Handwriting recognition will allow very easy and fast input of data than typewriting.

mOCR will be platform independent and developed using Wx Dev-C++ in windows XP and will be done as an internal project.

INTRODUCTION

Optical Character Recognition refers to the technique of analysing images to find and recognise the text in them. The input to an OCR system is therefore an image file and the output is the extractor text, in any machine readable form.

Optical Character Recognition is extremely relevant in today's world in the backdrop of the digital revolution. While 92-96% of human knowledge is locked up in printed and handwritten form, the amazing computing power that we seem to have at our disposal is entirely in the digital realm. OCR systems play an important part in bridging the gap and bridging these together. This technology assumes great significant in the local context, where OCR systems in Malayalam do not exist.

Malayalam OCR is a software device which will convert a scanned image of Malayalam printed/handwritten document into a computer editable

Malayalam text. With the help of such a program we can digitalize old books in Malayalam so that it can be preserved and we can have new copies very easily. Handwriting recognition will have breath taking impacts in the world of data entry.

OCR will make the text machine readable and should therefore be considered if

- The text is to be reused, edited or reformatted.
- The text should be available for full text information retrieval.
- The text is to be coded in HTML or SGML.
- The text should be available to adaptive equipment for the visually impaired.
- File size is of concern.
- Resources are available to perform OCR and correct the output.

1. OBJECTIVES

- Develop a Malayalam OCR.
- Support maximum number of fonts.
- Recognize handwriting.
- Recognize colored text.
- Provide facilities to correct and edit recognized text.
- Make the OCR compatible with all OS.

2. PRODUCT PERSPECTIVE

- Malayalam OCR will be an independent program.
- It will provide menu option to select the image of a page from a stored file or directly from the scanner.
- Malayalam OCR will convert the input image to a computer editable format and keep it in another window.
- Malayalam OCR will work in various platforms like windows.
- Malayalam OCR will be semi automatic ie, the recognition will be done page by page and user involvement is required for corrections.
- Malayalam OCR contains training facility so that it can be trained for recognizing new fonts.

REFERENCES

1. BOOKS

- The History of OCR, Optical Character Recognition, Herbert F.Schantz

2. RESEARCH PAPERS

- Optical Character recognition for Model-based Object Recognition Applications, Quig Chen and Emil M. Perlu, HAVE 2003 Proceedings, pp 77-82.
- Skeltonization of 2D Binary Images, David Eberly, Geometric Tools, Inc.

3. TUTORIALS

- IPL98 Tutorial v2,20

4. WEBSITES

- www.ccs.neu.edu/home/feneric/charrec.html
- www.wikipedia.org/wiki/Optical_character_recognition
- http://gocr.sourceforge.net/index.html

Done by

DEEPTHI M K

REMYA K V

SEMEER N M

VIPIN KV

VIPIN MOHAN