C
001.642
UWA

OPTICAL CHARACTER RECOGNITION (OCR) SYSTEM FOR PRINTED TAMIL TEXT USING THE UNICODE STANDARD

A PROJECT REPORT PRESENTED BY

MOHAMED MOHIDEEN MOHAMED UWAIS

to the Board of Study in Statistics and Computer Science of the

POSTGRADUATE INSTITUTE OF SCIENCE

in partial fulfilment of the requirement

for the award of the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

of the

UNIVERSITY OF PERADENIYA

SRI LANKA

2009

635222

University of Peradeniya
06 NOV 2010
Library

OPTICAL CHARACTER RECOGNITION (OCR) SYSTEM FOR PRINTED TAMIL TEXT USING THE UNICODE STANDARD

Mohamed Mohideen Mohamed Uwais,
Postgraduate Institute of Science,
University of Peradeniya,
Peradeniya, Sri Lanka

ABSTRACT

Tamil language usage in computers started with Tamil font development and this has been around for a few decades. Developers have created fonts for screen displays and print environments. Almost all these efforts can be classified into two categories.

The first one attempts to replace the 7 bit ASCII characters with Tamil glyphs, so that a user will able to compose Tamil text with Standard English keyboard which is most of the times with the QWERTY layout. As there is an obvious limitation in the number of keys available in the keyboard, developers had creatively dissected the glyphs so that a single key can be assigned to commonly used shapes and combined keystrokes for Tamil characters.

The second one attempts to place the glyphs in the 8 bit space retaining the 7 bit ASCII slots. This required a piece of software to map the characters to keys and such software was freely available. While in the initial shapes, there were so many different schemes of placing glyphs, sometimes known as encoding schemes, three were most commonly used namely TSCII, TAB and TAM. None of these schemes endorsed by a Standards organization but they provided the opportunity for the exchange of text in Tamil even in legacy Operating Systems.

With more and more platforms providing system level support for Unicode and Complex Script rendering it is now possible to build very high quality fonts for Tamil and all other text / font related technologies like the Optical Character Recognition (OCR) systems using the Unicode standard. This assures unique glyphs for each Tamil letter towards best appearance.

Optical Character Recognition (OCR) based on Unicode standard refers to the process of converting printed text documents into software translated Unicode text.

This study attempted to create an OCR system for printed Tamil text using Unicode standard. In the process, printed documents are scanned using standard scanners which produce an image of the scanned document. As part of the preprocessing phase the image file is checked for skewing. If the image is skewed, it is corrected by a simple rotation technique in the appropriate direction. Then the image is passed through a noise elimination phase and is binarized. The preprocessed image is segmented using an algorithm which decomposes the scanned text into paragraphs using special space detection technique and then the paragraphs into lines using vertical histograms, and lines into words using horizontal histograms, and words into character image glyphs using horizontal histograms. Thus a database of character image glyphs is created out of the segmentation phase. Existing image processing algorithms were used for this project.

Then all the image glyphs are considered for recognition using Unicode mapping. Each image glyph is passed through various routines which extract the features of the glyph. The various features that are considered for classification are the character height, character width, the number of horizontal lines (long and short), the number of vertical lines (long and short), the horizontally oriented curves, the vertically oriented curves, the number of circles, number of slope lines, image centroid and special dots. The glyphs are now set ready for classification based on these features. These classes are mapped onto Unicode for recognition. Then the text is reconstructed using Unicode fonts.

Resultant recognised text is editable in any text editor or word processing software supports Tamil Unicode.

Keywords: Tamil, fonts, encoding, OCR, Unicode