Advances in Computational Research
ISSN: 0975-3273 & E-ISSN: 0975-9085, Volume 4, Issue 1, 2012, pp.-54-56
Available online at http://www.bioinfo.in/contents.php?id=33

EFFICIENT ALGORITHMS FOR KANNADA AND ENGLISH SCRIPT IDENTIFICATION

SANGAME S.K.[1], RAMTEKE R.J.[1] AND GUNDGE Y.V.[2]

[1]Department of Computer Science, NMU, Jalgaon, MS, India.
[2]Department of MCA, Rural Engineering College, Bhalki, Dist: Bidar, KA, India.
*Corresponding Author: Email- sunsang2003@gmail.com[1], rakeshramteke@yahoo.co.in[1], yogeshv_in1@rediffmail.com[2].

Received: February 21, 2012; Accepted: March 06, 2012

Abstract- Script identification is required for a multilingual OCR system. In this paper, we present a novel and efficient technique for Kannada/English script identification. The proposed approach is based upon the analysis of Kannada pages and English pages. Experimental results demonstrate that the proposed techniques are capable of identifying Kannada/English scripts from the real handwritten Kannada and English pages .
Keywords- Kannada script, English script, filled images, cut images, features

Citation: Sangame S.K., Ramteke R.J. and Gundge Y.V. (2012) Efficient Algorithms for Kannada and English Script Identification. Advances in Computational Research, ISSN: 0975-3273 & E-ISSN: 0975-9085, Volume 4, Issue 1, pp.-54-56.

Introduction

Script Identification is part of Character recognition and the character recognition is the important area in image processing and pattern recognition fields. Automatic script identification plays an important role in processing large volumes of documents of unknown origin. In addition, the ability to reliably identify script using the least amount of textual data is essential when dealing with document pages that contain multiple languages. Since script analysis takes place prior to the actual text recognition, it cannot rely on character identification and ideally should consume only a small fraction of the total processing time. The aim of character recognition is to translate human readable characters to machine readable characters. Handwritten character recognition (HCR) has received extensive attention in academic and production fields. Handwritten documents present two challenges for script identification. First handwriting styles are more diverse than printed fonts. Cultural differences, individual differences, and even differences in the way that people write at different times, enlarge the inventory of possible character and word shapes seen in handwritten documents. Third, problems typically addressed in preprocessing, such as ruling lines and character fragmentation due to low contrast, are common in handwritten documents due to the variety of papers and writing instruments used. Lot of work is done in the recognition system. The recognition system can be either on-line or off-line. In on-line handwriting recognition words are generally written on a pressure sensitive surface (digital tablet PCs) from which real time information, such as the order of the stroke made by the writer is obtained and preserved. This is significantly different to off-line handwriting recognition where no dynamic information is available [1]. Off-line handwriting recognition is the process of identifying script from the various document images. It is the subfield of optical character recognition (OCR). Several methods of recognition of English, Latin, Arabic, Chinese scripts are excellently reviewed in [1, 2, 3, 4].

This paper is organized as follows. The section 2 describes the overview of Kannada Language The section 3 describes Data collection and discriminating features of Kannada and English Scripts. The section 4 describes proposed techniques for identifying the scripts, respective algorithms and the results obtained. Conclusion is given in section 5.

Overview of Kannada Scripts

In this section, we will explain the properties of popular South Indian scripts. Most of the Indian scripts are originated from

Brahmi script through various transformations. Writing style of Indian scripts considered in this paper is from left to right, and the concept of upper/lower case is not applicable to these scripts.

Kannada is one of the major Dravidian languages of Southern India and one of the earliest languages evidenced epigraphically in India and spoken by about 50 million people in the Indian state of Karnataka, Tamil Nadu, Andhra Pradesh and Maharashtra. The characters are classified into three categories: swaras(vowels), vyanjans (consonants) and yogavaahas (part vowel, part consonants). The script also includes 10 different Kannada numerals of the decimal number system.

Data Collection, Preprocessing and Discriminating features

A. Data Collection

Data collection for the experiment has been done from the different individuals. We have collected 2000 Kannada Character samples from different professionals. The database is totally unconstrained and has been created for validating the Identification system. The collected documents are scanned using HP-scan jet 5400c at 300dpi which is usually a low noise and good quality image. The digitized images are stored as binary images in JPG format.

B. Preprocessing

The standard database for Kannada handwritten character is not available; therefore we have created our own database. Data has been collected from different professionals belonging to schools, colleges, and commercial sectors. We have collected 2000 images from 150 writers for the experimentation purpose. A flat bed scanner was used for digitization. Digitized images are in gray tone with 300 dpi and stored as JPG format. We have used global threshold binarizing algorithm to convert them to two-tone (0 and 1) images (Here '1'represents object point and '0'represents background point). Scanned images often contain noise that arises due to printer, scanner, print quality, etc. therefore, it is necessary to filter this noise before we process the Identification of Kannada characters and English characters. The noise has been removed by using median filter and scanning artefacts are removed by using morphological opening operation

C. Some Discriminating Features of Kannada and English Script

Modern Kannada character set has 47 basic characters, out of which the first 13 are vowels and the remaining 34 characters are consonants. Some books report 14 vowels and 36 consonants. By adding vowels to each consonant, Kannada characters are obtained or a Kannada character is combined by another consonant to form a compound character. Hence, Kannada text words consists of combination of vowels, consonants, modified consonant and/or compound characters. The compound characters may have descendants called 'vattaksharas' found at their bottom portions. Some examples of Kannada compound characters with descendants are given in Figure 1. The presence of these descendants is one of the discriminating features of Kannada script, which is not present in the English, hence it could be used as a feature named bottom-component to identify the text word as a Kannada script.

It could be observed that most of the Kannada characters have either horizontal lines or hole-like structures present. Also, it could be observed that majority of Kannada characters have upward curves present at their bottom portion. Some characters have double-upward curves found at their bottom portions. In addition, left curve and right curve are also present at the left and right portion of some characters. Thus, the presence of the structures such as – horizontal lines, hole-like structures, bottom-up-curves, descendants, left-curves and right-curves could be used as the supporting features to identify Kannada scripts. Figure 1 shows the sample of Kannada script. The density of the occurrence of these features is thoroughly studied and the features with maximum density are considered in the proposed model.

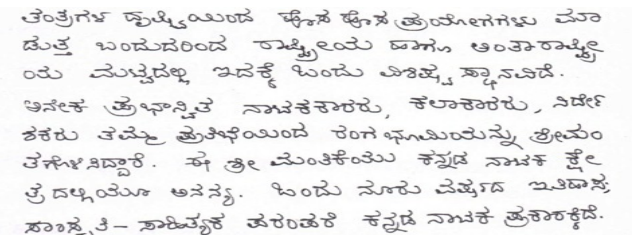

Fig. 1- Sample of Kannada Script

English character set has 26 alphabets in both upper and lower cases. One of the most distinct and inherent characteristics of the most of the English characters is the existence of vertical line – like structures. It could be observed that the upward – curve and downward – curve shaped structures are present at the bottom and top portion of majority of English characters respectively. So, it was inspired to use these distinct characteristics as supporting features in the proposed script identification model. Figure 2 shows the sample English script.

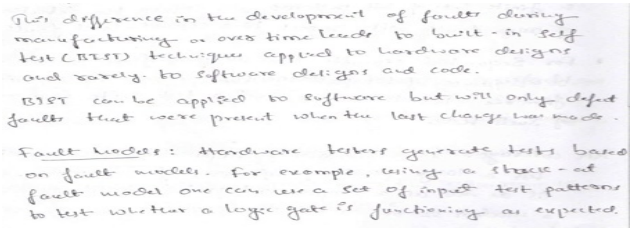

Fig. 2- Sample English Script

Feature Extraction

Features extraction is the identification of appropriate and unique characteristics of the component images. There are many popular methods to extract features. In this paper two major features are proposed to identify the Kannada and English scripts.

A. Feature 1

By thoroughly observing the structural outline of the characters of the two scripts, it is observed that the distinct features are present at some specific portion of the characters. So, in this paper, the discriminating feature is well projected by filling all the holes of Kannada and English scripts.

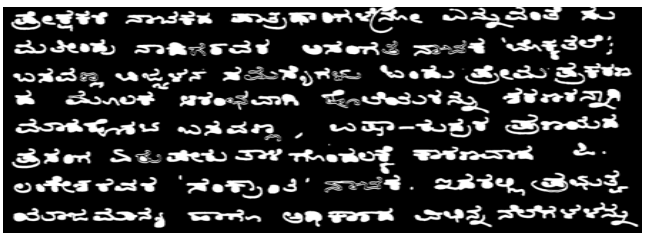

Fig. 3- Kannada Script Filled images

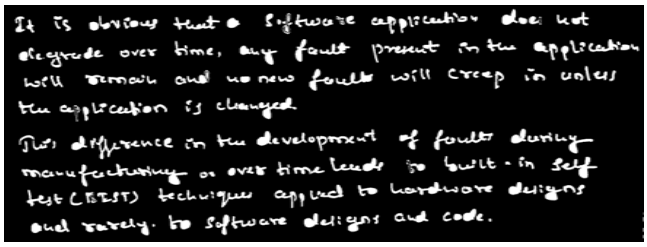

Fig. 4- English script filled images

Let us closely observe Figure 3 and Figure 4.Kannada filled page contains the maximum filled circle type elements whereas English filled page contains almost less number of filled circle type elements. This is the key for identifying Kannada page and English Page.

Input: Kannada and English Pages
Output: Identified Kannada and English pages.

1. Preprocess the input document image.
2. Fill the holes
3. Find the total number of white pixels in the page.
4. If number of white pixels is greater than 900000 then classify it as Kannada page otherwise English page.
5. Repeat step 1 to step 4 for all the document images.

B. Feature 2

The limitation of feature 1 is, if the half portion of the page contains written data and the remaining half portion is empty then the feature 1 does not give the correct result. Therefore, we have proposed second technique. This technique is summarized in the below algorithm. Figure 5 illustrates the algorithm.

Algorithm
Input: Kannada and English document images
Output: Identified Kannada and English Words

1. Preprocess the input document image
2. Segment the Page images to words images
3. Draw a horizontal line in the middle of each word image
4. Cut the top part up to horizontal line of each word image
5. Fill all holes of the word image.
6. Find the number of white pixels of filled images.
7. If the total number of white pixel is greater than 3000 then identify as Kannada word else identify as English word.

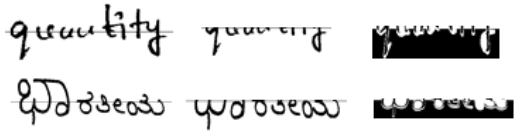

Fig. 5- Illustration of Feature 2

Results

The proposed algorithm has been tested on a test data set of 200 document images containing about 25 text lines from each script and found accuracy of 95%.

Conclusion

In this paper we presented the techniques for identifying offline handwritten Kannada and English characters. In future we work on recognition of characters.

Our future work aims to improve classifier to achieve better recognition rate and also to develop new feature extraction algorithms, which provides efficient results.

References

[1] Plamondon R. and Srihari S.N. (2000) *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 22(1), 63-84.
[2] Nafiz Arica and Fatos Yarman-Vural T. (2001) *IEEE Transactions on System. Man. Cybernetics-Part C: Applications and Reviews*, 31(2), 216-233.
[3] Liana M. Lorigo and Venu Govindaraju (2006) *IEEE Transactions on Pattern Ana. and Machine Inte.*, 22(5), 712-724.
[4] Nagy G. (1988) *ICPR*, 109-114.
[5] Pal U., Chaudhuri B.B. (2004) *Pattern Reco.,* 37, 1887-1899.
[6] Anil K. Jain and Torfinn Taxt (1996) *Pattern Recognition*, 29 (4), 641-662.
[7] Majumdar A. and Chaudhuri B.B. (2006) *first IEEE ICSIP06,* 1, 190-195.
[8] Abdur Rahim, Shuvabranta Saha, Mahfuzur and Abdus sattar, (2007) *International Conference on Information and Communication Technology,* 96-99.
[9] Pal U., Sharma N., Wakabayashi T. and Kimura F. (2007) *Ninth International Conference on Document Analysis as Recognition*, 496-500.
[10] Patil P.M., Sontakke T.R. (2007) *Pattern Recognition,* 40, 2110-2117.
[11] Pal U., Wakabayashi T. and Kimura F. (2007) *10th International Conference on Information Technology*, *IEEE*, 227-229.
[12] Hanmandlu M., Grover J., Madasu V.K. and Vasikarla S. (2007) *International Conference on Informational Technology*, 2, 208-213.
[13] Arun K. Pujari, Dhanunjaya C., Naidu M., Sreenivasa Rao and Jinaga B.C. *Image vision Computing,* 1221-1227.
[14] Suresh R.M., Arumugam S. (2007) *Image Vision Computing,* 25, 230-239.
[15] Lajish V.L. (2008) *IEEE International Conference on signal processing, Communication and Networking,* 374-379.
[16] Pal U., Wakabayashi T. and Kimura F. (2007) *Ninth International conference on Document Analysis and Recognition,* 2, 749-753.
[17] Rajaput G.G. and Mallikarjun Hangarge (2007) *PReMI*, *LNCS*.4815, 153-160.
[18] Rajashekararadhya S.V. and Vanaja Ranjan P. (2008) *International Conference on Cognition and Recognition,* 134-140.
[19] Rajashekararadhya S.V., Vanaja Ranjan P. and Manjunath Aradhya V.N. (2008) *First International Conference on Emerging Trends in Engineering and Technology*, 1192-1195.
[20] Rajashekararadhya S.V. and Vanaja Ranjan P. (2008) *second International Conference on information processing*, 162-167.
[21] Rajashekararadhya S.V. and Vanaja Ranjan P. (2004) *Digital Image Processing using MATLAB, Pearson Education*.
[22] Rajashekararadhya S.V., Vanaja Ranjan P. *Journal of Theoretical and Applied Information Technology.*