Volume 5, Issue 5, May 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Classification of Indian Language Scripts in Multilingual Documents

Komal
Assistant Professor, Department of CSE, Amity University
Haryana, India

Abstract— Being a multilingual country, Indian ancient and current literature is having quite a significant share for multilingual documents. The advancement of the computer technology has increased the online availability of such documents and hence the demand for identification, recognizing and retrieval of such documents has increased enormously. Current Optical Character Recognition (OCR) systems are not able to read the documents written in more than two scripts. In this paper a contour based classification scheme is proposed to identify the scripts of Devanagari, Gurumukhi, Urdu, Bengali and Kannada. The proposed methodology uses a hierarchical classification of scripts based on features and projections of the scripts.

Keywords— OCR, Scripts, Hierarchical Classification, Contour, Projections.

I. INTRODUCTION

There has been an enormous increase in online availability of all types of literature, study material, news, religious books etc. India has 22 officially recognized languages and more than 1000 dialects, which have been identified. At many places in Indian literature one can find multiple language scripts. Current OCR systems are unable to recognize and convert such document images into machine readable format due to inadequate discrimination capability of classifier. Other reasons which may attribute to non-processing of multilingual documents may be poor image quality, low resolution scanning, incorrect image pre-processing etc.

Due to the difficulty encountered by current OCR systems in processing multi-script documents, image analysis and classifiers of documents has gained insight of researchers in past few years. Automated tools will provide a great help to OCR systems for image preprocessing, classification, indexing, archiving and searching of online documents. Script identification also plays a vital role for online translators. Identification of scripts can be done on the basis of various features, contours and connected components of the scripts like single character, word formation, line, dot, breaks, horizontal or vertical projection, straight or curvy characters.

Many researchers have worked and contributed a lot in this direction. Judith Hochberg, Michael Cannon, Kevin Bowers and Patrick Kelly[1] utilized skew of connected components to classify and analyze documents using linear discrimination. Rajneesh Rani, Renu Dhir and Gurpreet Singh Lehal[2] proposed a Gabor feature based approach for identifying scripts in multilingual documents at line level. Sanghamitra Mohanty, Himadri Nandini Das Bebartta[3] have done a comparative assessment of K-nearest neighbour, Support Vector Machines and convolutional neural networks classifires for bi-lingual English Oriya printed documents.M Swamy Das, D Sandhya Rani, C R K Reddy, A Govardhan [4] present a script identification fapproach for multilingual English, Telugu and Hindi Text documents bu utilizing features like top max row, bottom max row, top horizontal lines, vertical lines, bottom components, tick components, top and bottom holes in hierarchical and K-Narest neighbor approaches.

U. Pal, T. Wakabayashi , F. Kimura[5] presented a comparative assessment of handwritten Devnagari character Recognition using various classifiers and features. Mohamed Farouk Abdel Hady, Mina Asham[6] discussed Canonical Correlation Analysis for multilingual topic modelling and for extracting language-independent topics from specific script labeled documents to address the problem of multilingual text classification. Sk Md Obaidullah , Anamika Mondal , Kaushik Roy[7] described a structural feature based methodology for finding out scripts of Indian printed documents. They implemented the suggested approach for 10 scripts using 5 times cross validation an average. Rumaan Bashir, S. M. K. Quadri[8] proposed a novel approach for script identification based in entropy analysis of multilingual image documents.

II. OVERVIEW OF INDIAN LITERATURE SCRIPTS

A. Devanagiri Script

Devanagari is a script which has evolved from Brahmi script and has become most prominent scripts of India and Nepal. Various languages such as Hindi, Sanskrit, Nepali, Marathi use Devanagiri script for writing purposes and their characters are derived from it. It is written from left to right, does not encompass distinctive letter cases and has a horizontal line that runs along the top of all letters of a word, known as “shero-rekha” and a line is terminated by a straight vertical line called “purna-viram”.

हालांकि सूर के जीवन के बारे में कई जनश्रुतियां प्रचलित हैं, पर इन में कितनी सच्चाई है यह कहना कठिन है। कहा जाता है उनका जन्म सन् १४७८ में दिल्ली के पास एक ग़रीब ब्राह्मीण परिवार में हुआ। जनश्रुति के अनुसार सूरदास जन्म से ही अंधे थे। आजकल थी अंधे आदमी अक्सर 'सूरदास' कहलाते है। कई लोगों ने उन्हें गुरु के रूप में अपनाया और उनकी पूजा करना शुरु कर दिया ।

Fig. 1 A sample of Hindi Script Excerpt

B. Gurumukhi Script

Punjabi is one of the famous speaking language of Northern India and Gurumukhi Script is used for writing Punjabi language. It is also written from left to right. Consonants in Gurumukhi Script have an innate vowel. Diacritics may appear above, below, after or below the consonant they belong to and are used to modify the innate vowel. Particular conjunct symbols are used to combine various essential parts of word for certain consonants.

ਕਈ ਕੰਪਿਊਟਰਾਂ ਵਿੱਚ ਫ਼ਾਇਲਾਂ ਨੂੰ ਇਸ ਸਾਈਟ ਤੋਂ ਸਿੱਧਿਆਂ ਪ੍ਰਿੰਟ ਕਰਨ ਵਿੱਚ ਮੁਸ਼ਕਲ ਹੋ ਸਕਦੀ ਹੈ। ਜੇ ਤੁਹਾਨੂੰ ਮੁਸ਼ਕਲਾ ਹਨ, ਤਾ, ਪਹਿਲਾ ਤੁਸੀਂ ਕੰਪਿਊਟਰ ਵਿੱਚ ਫ਼ਾਇਲ ਸਾਂਭਣ ਦੀ ਕੋਸ਼ਿਸ਼ ਕਰੋ। ਅਜਿਹਾ ਕਰਨ ਲਈ; ਲਿੰਕ ਤੇ ਰਾਈਟ ਕਲਿੱਕ ਕਰੋ ਜਾਂ ਲਿੰਕ ਨਾਲ ਟੈਬ ਕਰਕੇ ਰੀਟਰਨ ਕੀ ਦਬਾਉ ਅਤੇ ਚੁਣੋ, "Save Target As..." ਇਸਤੋਂ ਬਾਅਦ ਆਪਣੇ ਕੰਪਿਊਟਰ ਤੇ ਇਸਨੂੰ ਸਾਂਭਣ ਲਈ ਥਾਂ ਚੁਣੋ। ਫਿਰ ਇਸ ਨੂੰ ਸਾਂਭਣ ਵਾਲੀ ਥਾਂ ਤੋਂ ਖੋਲ੍ਹਣ ਲਈ Acrobat Reader ਦਾ ਇਸਤੇਮਾਲ ਕਰੋ ਅਤੇ ਪ੍ਰਿੰਟ ਕੱਢੋ।

Fig. 2 A sample of Gurumukhi Script Excerpt

C. Urdu Script

Urdu Script has evolved from its ascendants, that is, Persian and Arabic Scripts with an accommodation of Indian Phonology. Urdu Script is written from right to left, does not encompass distinctive letter cases. An important feature of this script is the use of subscript and superscript marks in place of short vowels. Another characteristic feature is the duplicate and triplicate existence of consonants which stand for the same sound.

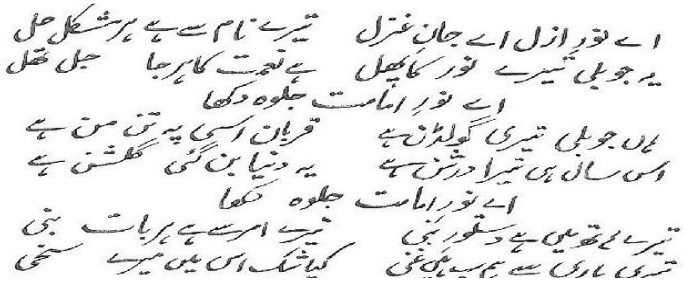

Fig. 3 A sample of Urdu Script Excerpt

D. Bengali Script

Bengali Script is the writing system of Bengali language and is written from left to right. Letters of Bengali script are grouped together and are not cased. A horizontal line runs on top of all letters of the word, called the head stroke. The vowels take dependent shape in case they have an onset consonant; else they take the independent vowel shape. Complex conjuncts may be formed by joining two or more consonants.

''সর্বভুতে ঈশ্বর দর্শন* ই মানুষের আদর্শ - উদ্দেশ্য। যদি সকল বস্তুতে তাঁহাকে দেখিতে না পার, অন্ততঃ যাহাকে সর্বাপেক্ষা ভালোবাসো, এমন এক ব্যক্তির মধ্যে তাঁহাকে দর্শন করিবার চেষ্টা কর - তারপর আর এক ব্যক্তির মধ্যে; এইরূপে অগ্রসর হইতে পার। আত্মার সম্মুখে তো অনন্ত জীবন পড়িয়া রহিয়াছে, অধ্যাবসায়ের সহিত চেষ্টা করিলে তোমার শুভ বাসনা পূর্ণ হইবে।''

দ্রষ্টব্য: * অংশটি জ্ঞানযোগ উদ্বোধন কার্যালয়, কলিকাতা, সেপ্টেম্বর, ১৯৯৯ সংস্করণ, পৃষ্ঠা ১৫৯ এ পাবেন।

* আসলে কথাটি ছিল ব্রহ্মদর্শন, কিন্তু ম এ ভ টা ঠিক মত লিখতে পারছি না।

Fig. 4 A sample of Bengali Script Excerpt

E. Kannad Script

Kannad script is an abugida of the Brahmic family. It is used primarily to write the Kannada language, one of the Dravidian languages of southern India. Kannad is a alphasyllabary writing system in which all consonants have an inherent vowel. Other vowels are indicated with diacritics, which can appear above, below, before or after the

consonants.Kannad script is written from left to right. The characters are classified into three categories: swaras (vowels), vyanjanas (consonants) and Yogavaahakas (part vowel, part consonants).

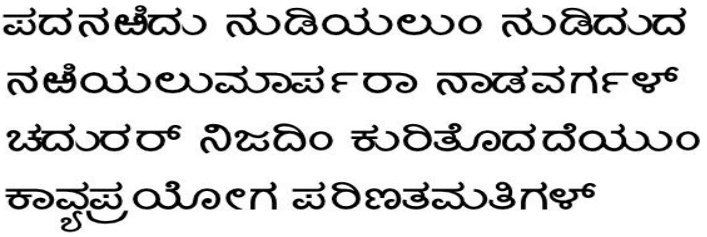

Fig. 5 A sample of Kannad Script Excerpt

III. PROPOSED METHODOLOGY FOR SCRIPT IDENTIFICATION

A. Image Pre-processing

- The digitalization of the text is done using an HP scanner. The digitalized images are the grey scale images so the intensity values vary from (0- 255).
- Binarize the image where the black pixels are represented by zero (0) and white pixel is represented by (1).
- Scanned documents always have few degree of skew may be due to human or mechanical error. De-noising of the binarized images is done to remove the skew and isolated pixels and small segments.

B. Line Segmentation

- Line segmentation is applied by partitioning every connected component of the script excerpt into equally- sized blocks. The width of each block is defined by the average width of character.
- After the block division, we use horizontal projection profile (means total no. of black pixels in the row) of the lines of the document. Horizontal projection profile becomes zero if there is no black pixel present in a row. Such points are used to segment the line into groups.
- Gurmukhi, Bengali and Devanagari scripts have the longest run of the black pixel in the row due to headline feature. This longest run is missing in Urdu and kannad scripts. Thus, after horizontal projection, the vailable scripts are divided into two groups- one with Gurmukhi, Bengali and Devanagari and second group having Urdu and Kannad.

C. Vertical Projection and Profile factor

- After line segmentation, Vertical Projection profile is applied to second group(that is Urdu and Kannad Script) to calculate the number of white pixels in divided line components/words. Urdu script has more number of white spaces especially in the beginning of the selected line segment.
- Along with the vertical projection profile, top profile and bottom profile of the two scripts are computed by scanning each column of the text line from top or bottom until it reaches a first black pixel. Thereafter, we calculate maximum number of black pixel in the top profile and bottom profile of the script.
- The profile factor is obtained by dividing the value of maximum pixel in top profile by the value of maximum pixel in top profile. The value of profile factor of the Kannada script is always greater than the Urdu script.

D. Feature Extraction

- Middle profile of the line segment component is taken into consideration to extract features like vertical line, angular skew, slant line connected with bottom of vertical line, curve shape connected with vertical line, horizontal line connecting to the middle of vertical line etc.

E. Template Matching

- After the feature extraction done from excerpts of various scripts of first- Devanagiri, Gurumukhi and Bengali, we match the features of each script with the corresponding template of Gurumukhi and Bangla. Whichever template is found to have maximum match, that script will be identified for the excerpt. However, if least matching with the two templates, it is devanagiri (Hindi) script.

Fig. 7 Bengali Script Template

Fig. 8 Alphabets matching with the Bengali Script Template

Fig. 9 Gurumukhi Script Templates

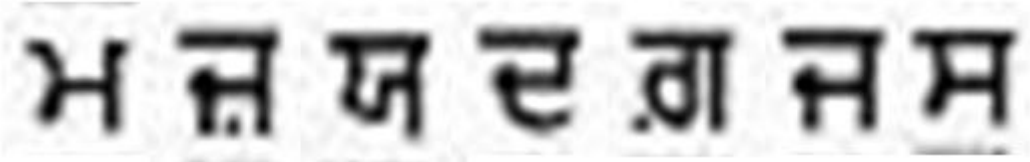

Fig. 10 Alphabets matching with the Gurumukhi Script Templates

F. Classification

- Based on all the previous steps, we are able to categorize the scripts and further able to conclude and identify the actual script, which can then be used as input to corresponding OCR or translator.

Fig. 6 Algorithm for classification of Devanagiri, Bangla, Gurumukhi, Kannad and Urdu in Multilingual Documents

IV. CONCLUSION

Script Recognition has been a more focused area of research in past 15 years. Online archival and retrieval of multilingual documents is necessitated due to growth of Internet savvy generation. A hybrid approach has been suggested in this paper for identification and classification of five most prominent Indian language scripts, which are Devanagiri (Hindi), Urdu, Gurumukhi, Bengali and Kannad. The approach is good enough to classify these scripts in trilingual documents as well.

REFERENCES

[1] Judith Hochberg, Kevin Bowers, Michael Cannon,Patrick Kelly, "Script and Language Identification for Handwritten Document Images",Proceedings 1999 Symposium on Document Image Understanding Technology, pp. 161-165.

[2] Rajneesh Rani, Renu Dhir and Gurpreet Singh Lehal, "Gabor Features Based Script Identification of Lines within a Bilingual/Trilingual Document", International Journal of Advanced Science and Technology Vol.66 (2014), pp.1-12.

[3] Sanghamitra Mohanty, Himadri Nandini Das Bebartta, "A Comparative Analysis of Classifiers Accuracies for Bilingual Printed Documents (Oriya-English)", International Journal of Computer Science and Information Technologies, Vol. 2 (2), 2011, pp. 916-923.

[4] M Swamy Das, D Sandhya Rani, C R K Reddy, A Govardhan, "Script identification from Multilingual Telugu,Hindi and English Text", International Journal of Wisdom Based Computing, Vol. 1 (3), December 2011,pp. 79-85.

[5] U. Pal, T. Wakabayashi , F. Kimura, "Comparative Study of Devnagari Handwritten Character Recognition using Different Feature and Classifiers", 2009 10th International Conference on Document Analysis and Recognition, pp. 1111-1115.

[6] Mohamed Farouk Abdel Hady, Mina Asham, "Class-Dependent Canonical Correlation Analysis for Scalable Cross-Lingual Document Categorization", 2013 IEEE Symposium on Computational Intelligence and Data Mining, pp. 308-315.

[7] Sk Md Obaidullah , Anamika Mondal , Kaushik Roy , "Structural Feature Based Approach for Script Identification from Printed Indian Document", 2014 International Conference on Signal Processing and Integrated Networks (SPIN),pp. 120-124.

[8] Rumaan Bashir, S. M. K. Quadri ,"Entropy Based Script Identification of a Multilingual Document Image", 2014 International Conference on Computing for Sustainable Global Development (INDIACom), pp. 19-23.

