ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

Word-wise Script Identification of South Indian Document Images

SmitaBiradar[1], V.S. Malemath[2], Suneel C Shinde[3]

Department of CSE, KLE Dr.MSS College of Engg., & Tech. Belgaum[1]

Head of Computer Science & Engg, KLE DR M S Sheshgiri College of Engg. & Tech., Belgaum[2]

Faculty, Dept. of Master of Computer Applications, KLE DR M S Sheshgiri College of Engg. & Tech., Belgaum[3]

Abstract: A document page may consist of text words, numerical in different regional script along with the English and/or National language.Especially the documents in multilingual country or in the border area may have this scenario to convey information at mass. The monolingual OCR fails to identify such other script words and hence script identification becomes essential in such cases. Script identification is one of the challenging steps in the Optical Character Recognition system for multi-script documents. In this work we propose a word-wise script identifier considering all the south Indian languages. The proposed method uses morphological features such as dilation and erosion and reconstruction as base and a nearest neighbor classifier is used to classify the script. The method showed robustness in the estimation of script when tested on 600 word document images.The overall accuracy is found to be 98.1%

Keywords: OCR, Script Identification, morphological reconstruction, multilingual documents, multi script documents, NN classifier.

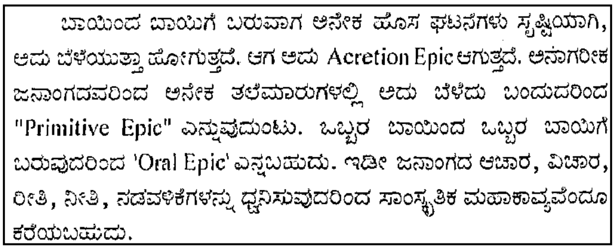
I. INTRODUCTION

Being updated with the recent trends of having paperless office digitalization of documents has become mandatory and of lot significance. Document image analysis involves optical character recognition where any printed or handwritten document is converted into the format known to a machine.

In a multilingual nation like India, an archive may contain words in more than one language as shown in fig1 (a) (b). In such scenario a multilingual situation, multi lingual Optical Character Recognition (OCR) framework is expected to peruse the multilingual records. In this way, it is essential to recognize distinctive language areas of the record before sustaining the report to the OCRs of individual language.

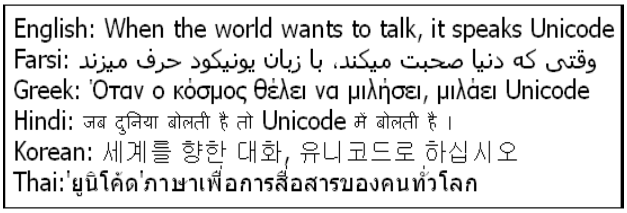
The handling of such complex multi-script documents is a testing issue for OCR programmers. The monolingual OCR frameworks fails to process such multi-script records without human contribution for outlining distinctive script zones of multi-lingual pages before actuating the script to a particular OCR system. The requirement for such manual contribution can bring about more noteworthy cost and essentially defers the general picture to-content transformation.

Hence, a programmed system is needed for the approaching documentimages to handover this to the specific OCR system upon the information of the characteristic scripts. In perspective of this, recognizable proof of script and/ or language is one of the rudimentary assignments for multi-script reporthandling. A script identification system, accordingly, improves the assignment of OCR by upgrading the precision of acknowledgment and decreasing the computational complexity.



ಬಾಯಿಂದ ಬಾಯಿಗೆ ಬರುವಾಗ ಅನೇಕ ಹೊಸ ಘಟನೆಗಳು ಸೃಷ್ಟಿಯಾಗಿ,
ಅದು ಬೆಳೆಯುತ್ತಾ ಹೋಗುತ್ತದೆ. ಆಗ ಅದು Acretion Epic ಆಗುತ್ತದೆ. ಅನಾಗರೀಕ
ಜನಾಂಗದವರಿಂದ ಅನೇಕ ತಲೆಮಾರುಗಳಲ್ಲಿ ಅದು ಬೆಳೆದು ಬಂದುದರಿಂದ
"Primitive Epic" ಎನ್ನುವುದುಂಟು. ಒಬ್ಬರ ಬಾಯಿಂದ ಒಬ್ಬರ ಬಾಯಿಗೆ
ಬರುವುದರಿಂದ 'Oral Epic' ಎನ್ನಬಹುದು. ಇಡೀ ಜನಾಂಗದ ಆಚಾರ, ವಿಚಾರ,
ರೀತಿ, ನೀತಿ, ನಡವಳಿಕೆಗಳನ್ನು ಧ್ವನಿಸುವುದರಿಂದ ಸಾಂಸ್ಕೃತಿಕ ಮಹಾಕಾವ್ಯವೆಂದೂ
ಕರೆಯಬಹುದು.

Fig. 1(a) Example of bi-script document with English words



English: When the world wants to talk, it speaks Unicode
Farsi: وقتی که دنیا صحبت میکند، با زبان یونیکود حرف میزند
Greek: Όταν ο κόσμος θέλει να μιλήσει, μιλάει Unicode
Hindi: जब दुनिया बोलती है तो Unicode में बोलती है ।
Korean: 세계를 향한 대화, 유니코드로 하십시오
Thai:'ยูนิโค้ด'ภาษาเพื่อการสื่อสารของคนทั่วโลก

Fig. 1(b) Example of multilingual document

II. LITERATURE REVIEW

The literature reveals that the work carried out in the past is based on threeways viz.

(1)Mono script identification
(2)Line wise script identification
(3)Word wise script identification

In mono script identification a document is solely and entirely in a single script and hence is produced for script identification.In line wise script identification each and every text line is extracted and the script is identified.The third type is considered to be the most difficult where each word is used for script identification.

DOI 10.17148/IJARCCE.2015.48103

BhatiaDifferent methods have been proposed to understand the focal point of character acknowledgment in an optical character acknowledgment framework. Despite the fact that, adequate studies and papers are portrays the systems for changing over textual content from a paper document into machine readable form. Optical character recognition is a procedure where the PC sees naturally the picture of written by hand script and move into group character. This material use as an aide and redesign for perusers working in the Character Recognition territory. Determination of a significant component extraction system is most likely the absolute most vital consider accomplishing high character acknowledgment with vastly improved precision in character acknowledgment frameworks with no variety.

M.C Padma.et.alhave proposed the script identification improved schemes based on visual discriminating features like directional strokes, equal and unequal sized blocks and variable sized characters etc. For classification, they have considered Tamil, Telugu, Devanagari, English and Kannada.U. Pal, S. Sinha et.al carried out word level script identification for different configurations. Topological and structural features like number of loops, headline feature, stokes, water reservoirs and projection profiles etc are focused by them in their algorithms. Manoj Kumar Shukla et. alproposed a method in which they show a system for script division of the individual content line for imprinted in Indian language report.Rajneesh Raniet.alproposed to exhibits a zone based Gabor highlight extraction system. The given word picture after standardization is partitioned into distinctive zones of diverse sizes and afterward highlights from each of these zones are removed in different headings utilizing gabor channels. Script is then dictated by utilizing SVM classifier. The trial tests completed in the field of Gurumukhi and English Script acknowledgment demonstrate that the proposed system prompts change over the conventional Gabor highlight extraction without zoning. In future, this can likewise be reached out for different scripts.

Dhandraet. alproposed a method in which they considered texture as a tool for deciding the script of written by hand record image,based on the perception that content has a particular visual texture.Further, K closest neighbor calculation is utilized to order English, Devnagari and Urdu scripts, in light of 13 spatialspread elements removed utilizing morphological channels.

From the above literature it reveals that there are works which are carried out considering three to four types of scripts but none of them have attempted identification for six scripts hence in this work an attempt is made to consider six south Indian scripts which consist of Kannada, Malayalam, Telugu, Tamil, English and Hindi.

III. PROPOSED METHOD

In the proposed work a set of nine features are extracted from the test image X and these feature valuesare compared with feature values stored in the knowledge base. For classification the minimum distance classifier is used.

Digitization and pre processing of Input Documents: As the standard dataset for script is not available hence we created our own dataset. During this process the documents from news papers articles books, magazines for each of the script were considered. The documents with varied font style and font sizes are considered to have variability in dataset and to check the robustness of the algorithm nad the feature extracted.

In all 6 scripts of south India are considered for the experimentation. These includethe data sets images of Kannada, Telugu, Devnagari, Tamil, Malayalam and English The collected documents are scanned using HP Scanner at 300 DPI, which usually yields a low noise and good quality document image. The digitized images are in gray tone and we have used Otsu’s global thresholding approach to convert them into two-tone images as shown in fig.3. The two-tone images are then converted into 0-1 labels where the label 1 represents the object and 0 represents the background. The small objects like, single or double quotation marks, hyphens and periods etc. are removed using morphological opening. The next step in pre-processing is skew detection and correction.

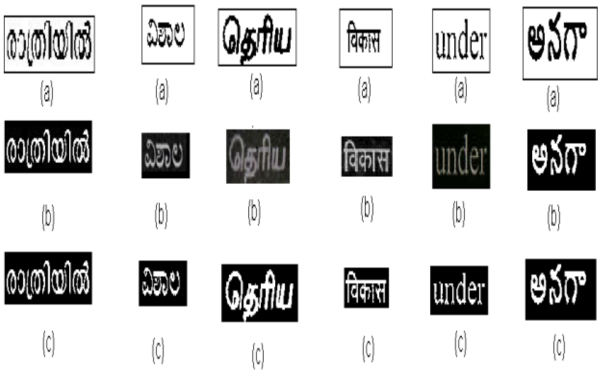


Fig. 3 (a): Document images of Malayalam, Kannada, Tamil, Hindi, English and Telugu
(A) Original images (b) grayscale images of (a)
(c) two toned images after thresholding.

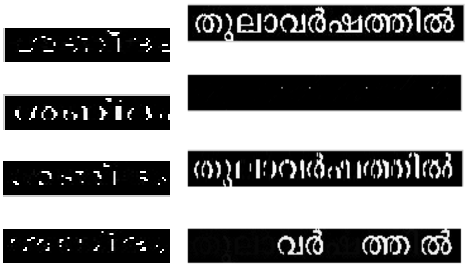
FEATURE EXTRACTION:

Feature extraction in any recognition system plays an important part. In this work a set of nine features is considered and evaluated for each set of document images belonging to that particular script. For the six south Indian languages Kannada, Hindi, Telugu, Tamil, Malayalam and the two documentation languages English and Hindi a set of nine features is extracted.

The set of nine features extracted to distinguish between the languages describe the properties. The feature set considered includes horizontal erosion, vertical erosion, right diagonal erosion,left diagonal erosion, fill hole, erosioncombined with morphological reconstruction forhorizontal, vertical, right, and left diagonal erosion.

The process is depicted in fig 4.

DOI 10.17148/IJARCCE.2015.48103



(a) (b)

Fig.4: sample example to show feature extraction process
(a): Extraction of features (horizontal, vertical, left diagonal. right diagonal deviation) for a Malayalam image.
(b): morphological opening and reconstruction for line as a structuring element for a Malayalam image.

IV. ALGORITHM

Input: Segmented Document image of a word
Output: Identification of Script Type

Steps:

Step1: Preprocess the document image for noise removal and to remove small artifacts such as colon, commas, quotation marks full stops etc.

Step2: Convert images into gray scale. Apply Otsu's method to binarize the inputted word.

Step 3: Perform horizontal, vertical, right diagonal and left diagonal erosionby using masks which are designed using 3 X 3 mask structural elements.

Step 4:Estimate the average value of density of pixels with original image. This constitutes four set of features.
Step 5: Repeat step 3 and perform morphological reconstruction and repeat step 4 which yields another four set of features.
Step 6: Perform fill hole operation on the input image and estimate average density of pixels with original image.

Step 7: For training extract nine set of features mentioned in above step for 25 images of each script type and store the average value of each feature for every script as knowledge base.

Step 8: Perform classification using nearest neighbor classifier and classify the test text words in tothe type of the class of its nearest neighbor.

Stop.

V. EXPERIMENTAL RESULTS

In order to carry out the experimentation 100 test document images for each of the script which is total of 600 document test images areconsidered and features were extracted as shown in fig 5 and fig 6.

In order to verify the robustness of the algorithm randomly 25 images are selected and the rest images are considered as test images.

The results are tabulated in table.

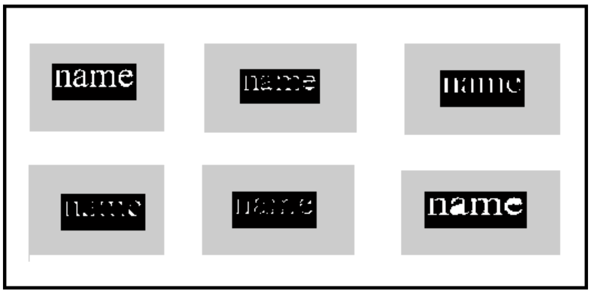


FIG 5: Intermediate stage images of English: (from left to right)(1)Input script word (2)Horizontal erosion of image, (3) Vertical erosion of image (4) Left diagonal erosion of image (5) Right diagonal erosion of image (6) Image after hole-filling

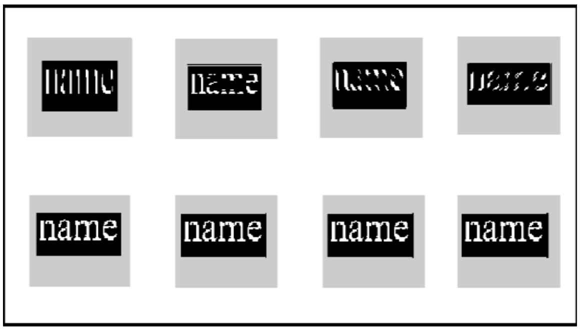


FIG 6: Intermediate stage images of English: (from left to right)(1)opening after vertical erosion (2)opening after Horizontal erosion, (3) opening after left diagonal (4)opening after right diagonal erosion (5)(6)(7)(8) Images after reconstruction

The tables below depict the values obtained after Bi-script classification of every script type with English script. This table is followed by confusion matrix.

Table 1: Bi-script Classification Results of six languages

Script	Kan	Hindi	Mal	Tam	Tel	% Accuracy
English	98.5	96.5	100	97.5	98.0	98.1%

Table 2: Confusion Table of English and Kannada Script Classification

Scripts	English	Kannada
English	98	2
Kannada	1	99

Table 3: Confusion Table of English and Hindi Script Classification

Scripts	English	Hindi
English	97	3
Hindi	4	96

Table 4: Confusion Table of English and Malayalam Script Classification

Scripts	English	Malayalam
English	100	0
Malayalam	0	100

Table 5: Confusion Table of English and Tamil Script Classification

Scripts	English	Tamil
English	3	97
Tamil	98	2

Table 6: Confusion Table of English and Telugu Script Classification

Scripts	English	Telugu
English	100	0
Tamil	99	1

VI. CONCLUSION

Script identification essentially is a basic task before optical character recognition. In thiswork, a system for script identification for all south Indian scripts is proposed. It is based on simple mathematical morphological operations. The experimental results are found to be encouraging i.e. recognition rate of 98.1% exhibited on 600 images. .The method also showed the robustness when it was exposed to variable font sizes and styles. The future scope of the work would be to test for larger data sets and also considering handwritten scripts.

REFERENCES

[1]M. C. Padma and p. A. Vijaya, "script identification from trilingual documents using profile based features," international journal of computer science and applications,vol. 7 no. 4, vol. 7, pp. 16-33, 2010.

[2]R. Rani, R. Dhir, and G. S. Lehal, "Gabor Features Based Script Identification of Lines within a Bilingual/Trilingual Document," International Journal of Advanced Science and Technology, vol. 66, pp. 1-12, 2014.

[3] V. N. ManjunathAradhya, G. Hemantha Kumar, and S. Noushath, "Multilingual OCR system for South Indian scripts and English documents: An approach based on Fourier transform and principal component analysis," Engineering Applications of Artificial Intelligence 21, pp. 658-668, 2008.

[4] G. D. Joshi, S. Garg, and J. Sivaswamy, " Script Identification from Indian Documents.," in Document Analysis Systems VII, 7th International Workshop DAS 2006, Nelson, New Zealand, February 13-15, 2006, Proceedings, nelson,newzealand, 2006.

[5] U. Pal, S. Sinha, and B. B. Chaudhuri, "Multi-Script Line identification from Indian Documents," in Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003), 2003.

[6] I. T. Young, J. Jan, Gerbrands, J. Lucas, and V. Van.(1995-2007) Fundamentals of image proccessing.Document.

[7] K. M. M. Rao. Overview of image proccessing : READINGS IN IMAGE PROCCESSING. document.

[8] N. Bhatia, "Optical Character Recognition Techniques: A Review," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4, no. 5, pp. 1219-1223, May 2014.

[9] M. K. Shukla and B. Haider, "Line-wise Script Segmentation for Indian Language Documents," International Journal of Computer Applications, vol. 108, pp. 34-37, Dec. 2014.

[10] R. Rani, R. Dhir, and G. S. Lehal, "Modified Gabor Feature Extraction Method for Word Level Script Identification- Experimentation with Gurumukhi and English Scripts," International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 6, no. 6, pp. 25-38, 2013.

[11] M. Hangarge and B. V. Dhandra, "Offline Handwritten Script Identification in DocumentImages," International Journal of Computer Applications, vol. 4, pp. 6-10, Jul. 2010.

[12] U. Pal and B. B. Chaudhuri, "Indian Script Character Recognition: a survey," PATTERN RECOGNITION, vol. 37, pp. 1887-1899, Feb. 2004.

[13] V. S. Malemath, A. H. Kulkarni, and H. Mallikarjun, "Word-wise Script Identification in Document," International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, no. 6, Jun. 2014.

[14] BV Dhandra, P Nagabhushan, MallikarjunHangarge, RavindraHegadi, VS Malemath "Script identification based on morphological reconstruction in document images" 18th International Conference on Pattern Recognition , 2006. ICPR 2006,Hongkong pp950-953

[15] P. B. Pati and A. G. Ramakrishna, "Word level multi-script identification," Pattern Recognition Letters, vol. 29, pp. 1218–1229, 2008.

[16] U. Pal and B. B. Chaudhuri, "Script line separation from Indian Multi-script documents," In Proceeding 5th International Conference on Document Analysis and Recognition, pp. 406-409, 1999.

[17] B V Dhandraet. al, Script Identification based on Morphological Reconstruction in Document Images", In Proc. of 18th International Conference on Pattern Recognition (ICPR2006) Hong Kong, Aug. 2006, Vol. II-3, pp 950-53.

[18] B V Dhandraet. al, Word wise script identification in bilingual documents based on Morphological reconstruction", In Proc. of 1st IEEE International Conference on Digital Image Management (ICDIM2006), Bangalore India, held during 6-8 Dec. 2006, pp 389-94.

[19] B V Dhandraet. al, Word- wise Script Identification based on Morphological Reconstruction in printed Bilingual Documents", In Proc. of IET International Conference on Visual Information Engineering (VIE2006) Bangalore, 28-29 Sept. 2006, pp 389-393.

BIOGRAPHIES

SmitaBiradar received B.E. degree in Computer Science Engineering from Visvesvaraya Technological University of Belgaum in 2013. Currently pursuing her M.Tech degree in Visvesvaraya Technological University of Belgaum. Her research interests include Image Processing and Wireless Sensor Networks.

Dr. Virendra. S. Malemathis currently a Head of Computer Science &Engg, KLE DR M S Sheshgiri College of Engg. & Tech., Belgaum. He did his Bachelors in Engg. in Electronics & Communication Engg. from Karnataka University, Dharwad in the year 1993, did his MS in Software Systems from BITS Pilani Rajasthan in 1998 and received his PhD in Computer Science from Gulbarga University, Gulbarga, India in 2009. His research interests are document image processing medical and pattern recognition. He has published more than 60 articles in peer reviewed international journals and conferences.

Prof. Suneel C Shinde is a Faculty in the Department of Master of Computer Applications, KLE DR M S Sheshgiri College of Engg. & Tech., Belgaum. He did his Bachelor from Karnatak University Dharwad and M Tech in Computer science & Engineering from University of Mysore. His research interest include Image processing and Pattern Recognition He has number of publications in to his credit in peer reviewed international journals and conferences.

DOI 10.17148/IJARCCE.2015.48103