IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

Word Wise Script Identification in Indian Context

Sugaladevi Choukimath[1], V S Malemath[2], Suneel C Shinde[3]

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

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

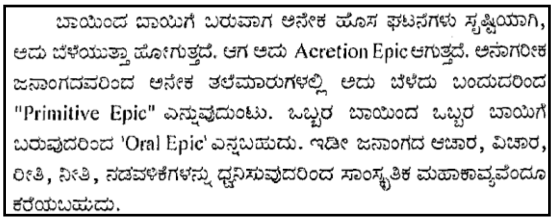
Faculty, Dept. of Master of Computer Applications, KLE DR M S Sheshgiri College of Engg & Tech., Belgaum [3]

Abstract: Script is a set of symbols and rules used to express or convey the information in a graphic form. Script Identification is one of the challenging steps in the Optical Character Recognition system for multi-script documents. In Indian and Non-Indian context some results have been reported, but research in this field is still emerging. This paper presents study on word wise script Identification which is based on scale Invariant Feature Transform, The system is developed and tested for 500 document images representing English, Hindi, Kannada, Bengali and Gurumukhi scripts. The system is developed includes a feature extractor which is based on scale invariant feature transform and for classification nearest neighbour classifier is used. The method is found to be robust and classification accuracy across five scripts is found to be 97.8%.

Keywords: Script Identification, Image Processing, SIFT, KNN.

I. INTRODUCTION

Script is a mean of communication which is in the form of set of rules and symbols. India is a country of multiple languages which are the member of different language families. The Indo Aryan languages are spoken by 75% of Indians.20% of Indians speaks Dravidian Languages. Other languages spoken in India belong to few other minor language families. The recent trends of having paperless office digitalization of documents has become mandatory and of lot significance. Presently, there is a considerable amount of work for computerization of document analysis system, due to the advancement of software and hardware technologies. Optical Character recognition (OCR) system contributes to this area by providing techniques to convert the large volumes of physical documents into an electronic form. Many papers and patents can be seen with recognition rates of 99% and above; this gave the impression that the digital document problems have been solved. The OCR technology for Indian scripts is in infant stage. Most of these OCR systems can read the documents written in single script and /or language only. In Asian countries most of the official and commercial documents are multi-script/multi-lingual in nature.We have multiple languages to communicate and any document may have content words expressed in multiple languages. In a multilingual, multiscript country like India, automatic script reorganization becomes quite a challenging task



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.Bi script document that contains English words

The monolingual OCR systems will not process such multilingual documents without human intervention for demarcating different script and / or language zones of multi-lingual pages before activating the script and language specific OCR engine. The need for such manual intervention can be labor intensive, which results in greater expense and significantly slows down the overall image-to-text conversion.

In this regard there is a need to develop pre OCR script and language Identification system to enable to select the appropriate OCR system for processing the document containing different scripts and languages which provide a new approach to script Identification environment As Script Identification supports vital applications such as image sorting, Appropriate Script selection, better text understanding system and online archiving of document image containing specific script.

II. RELATED WORK

Many works are carried out for script identification in document images. Script Identification works are always carried out in three different ways:

(1) Mono Script Identification
(2) Line Wise Script Identification
(3) Word wise Script Identification

The entire document is produced as single script for script Identification. The second way is Line wise script Identification where a document is divided into lines and those lines are produced for script Identification. Word wise script identification is one of the toughest ways where each word of a document is produced for script identification.

DOI 10.17148/IJARCCE.2016.55115

Malemath et.al [1] have carried out a study on word wise script Identification based on Steerable Gaussian filter for printed document Images .The system Developed includes a feature extractor which is based on Steerable Gaussian filter technique and K-nearest neighbor classifier is used. The System is tested for 3000 document images of English, Kannada, Hindi, Urdu and Telugu images of 600 words each.

The results are analyzed using two different classification techniques to study the robustness of the features extracted as well as the properties and structural shape differences of 5 Indian scripts.

The results are found to be encouraging with both classification methods. The Linear discriminate classifier gives an accuracy of 99.125%. The KNN classifier provided the accuracy of about 96%.The work can be enhanced for other native languages.

Hangarge et.al [2] carried out a work at word level and they have identified a new tool for morphological opening which is done by the reconstruction of the images .The proposed System is carried out for three Indian languages which are Kannada, Telugu and Hindi.

David et.al [3] have compared the performance of three classifiers for word level script Identification and the classifiers used are as follows.

SVM (Support Vector machines).
GMM (Gaussian mixture model).
KNN (k-Nearest Neighbor).

These classifiers are used for different bilinguals and their performance comparison is carried out. The three classifiers have provided the considerably good performance but the accuracy rate is effected by the noise, line spacing, word spacing and so on and this drawback need to be considered in proposed methodology.

Padma et.al [4] have carried out the script identification for the documents containing three languages. The proposed system has a model to identify the scripts of Kannada, Hindi and English using K-nearest neighbor. The authors have carried out the experiment using 1500 lines for learning and 1500 lines for testing and have achieved the average accuracy of 99.5%.In future the author suggest to carry out the work at word level for scanned images.

Rajesh et.al [7] have carried out a work which is helpful to identify only a single script. In case of Bilingual/Trilingual document, OCR fails to identify the script. Thus here authors work on this particular problem. In this paper, the script Identification is carried out Line Level for a document image using Gabor Filter. The work has achieved the recognition rate of 99.85% for trilingual document images and carries out a performance comparative study of SVM, KNN and PNN classifiers and the results are encouraging and The work can be enhanced for word level script identification and for Indian and Non-Indian Scripts.

From the above Literature it reveals that ,many works are carried out considering block and line level script Identification and few works are carried out at word level script identification considering three to four types of scripts Hence In this work an attempt is made to identify five native Indian scripts which consist of Kannada, Bengali, Gurumukhi, Hindi and English .

III.PROPOSED METHOD

Digitalization and pre-processing:

In this work, five data sets are used for experimentation. The data sets include digital images of English, Hindi, Kannada, Bengali and Gurumukhi. These are collected from different books, daily papers, magazines and from other different online sources which tend to have different textual styles and sizes.

Utilizing HP Scanner at 600DPI, the reports are in examined, which provide a low commotion and great quality. After digitalization, Binarization and noise removal is carried out and the symbols like hyphens, periods, single quotes and double quotes are vanished at the step of preprocessing.

Feature Extraction:

Scale Invariant Feature Transform (SIFT)

Scale Invariant Feature Transform algorithm is helpful in detecting and describing the local features in the images. Correct identification with lesser mismatch.

Scale Invariant Feature Transform key points of the object are first extracted from a set of training images and are stored in the image. A script is recognized in a new image by individually comparing each feature from the new image to this database .Scale Invariant Feature Detection is one of the key stage of the SIFT which is taken in our work. In Scale Invariant Feature Detection, an image is transformed into a large collection of feature vectors, which are invariant to image scaling, translation and rotation, partially constant to illumination changes.

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

IV. ALGORITHM

Input: Segmented Document image of a word
Output: Identification of Script type
STEPS:
START
Step1: Uploading the Document Image
Step 2: Preprocessing the image for noise removal to remove small artefacts commas, Colons, quotation marks, full stops etc.
Step 3: convert the images into grey scale by identifying the RGB combination in the uploaded image.
Step4: Perform the SIFT Feature Operations. SIFT obtains the special features in the image, in order to identify special characters in the images and train it to temporary database.
Step5: Train and store the above feature special characters to temporary database.
Step 6: Perform fill whole operation on the input image and estimate the average density of pixels with original image.

DOI 10.17148/IJARCCE.2016.55115

Step7: For Training, Feature values mentioned in above step for 50 images of each script type and the average value of features for every script is stored in the knowledge base.

Step 8: Perform classification of the scripts using nearest neighbour classifier and identify the text words into the type of the nearest class.

STOP

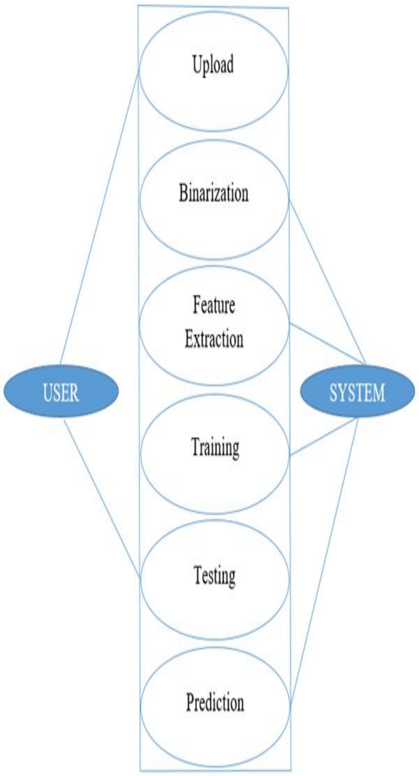


Fig 2. Use case diagram for script identification

The above Use case diagram specifies that, in the proposed work only Uploading of the image and taking the image for testing is done by the user here as the system frame work carries out the essential tasks like Binarization, Feature Extraction, Training and Prediction. Thus making the script Identification quite more robust.

V. EXPERIMENTAL RESULTS

As there is no standard database for the document images for the native Indian scripts, there was a great need to create own database of datasets in the following five languages English Hindi, Kannada, Bengali and Gurumukhi. Dataset of document images were searched and found through Internet, Magazine, web documents online newspapers and books etc.In order to carry out the experimentation of the work, 100 test document images for each of the script which is total of 500 document test images are considered and features were extracted using Scale Invariant Feature Transform and the nearest neighbor classifier was used in script Identification. In order to verify the robustness of the algorithm randomly 50 images were selected for training and the rest images are considered as test images.

The table below depicts the experiment values obtained after Bi script classification of every script type with English script. The tables below show the confusion matrix of the scripts by taking the English

Table 1: Bi-script Classification Results of five languages.

Script	Kannada	Hindi	Bengali	Gurumukhi	Accuracy in %
Eng	97.2%	97.4%	98.0%	98.3%	97.8%

Table 2: Confusion Table of English and Kannada Script Classification.

Script	English	Kannada
English	94	6
Kannada	3	97

Table 3: Confusion Table of English and Hindi Script Classification.

Script	**English**	Hindi
English	97	97
Hindi	3	3

Table 4: Confusion Table of English and Bengali Script Classification.

Script	English	Bengali
English	92	8
Bengali	4	96

Table 5: Confusion Table of English and Gurumukhi Script Classification.

Script	English	Gurumukhi
English	97	3
Gurumukhi	1	99

VI. CONCLUSION

In this work, a framework for script identification for native Indian scripts is proposed. The work is based on the scale Invariant feature transform operations. The experimental results are found to be encouraging with the recognition rate of 97.8 % exhibited on 500 images and the results obtained are found to be accurate. The work may be enhanced to other Indian and non-Indian scripts by considering the larger datasets.

REFERENCES

[1] V .S. Malemath, A. H. Kulkarni and H. Mallikarjun," Word-wise Script Identification in Document Images based on Steerable Gaussian Filtering Technique" ,International Journal of Advanced Research in computer and communication and Engineering" ,vol 3,no.6,Jun 2014.

[2] M. Hangarge and B.V .Dhandra, "Morphological Reconstruction for Word level script identification", International Journal of Computer Science and Security, vol.1, no.1, pp 41-51.

[3] David Doermann and Huanfeng Ma, "word level script identification for scanned images", Language and Media Processing

Laboratory Institute for Advanced Computer Studies University of Maryland, College Park, MD 20742, USA

[4] 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.

[5] R .Rani, R. Dhir and G. S. Lehal,"Gabor feature Based script identification of lines within a bilingual/trilingual document", International Journal of Advanced Science and Technology, Vol 66, pp.1-12, 2014.

[6] Rajneesh Rani1, Renu Dhir1 and Gurpreet Singh Lehal2 "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.5 (2013), pp.25-38

[7] Rajesh Gopakumar, N.V.SubbaReddy, Krishnamoorthi Makkithaya, U.Dinesh Acharya," Script Identification from Multilingual Indian Documents using Structural Features ",Journal Of Computing, Volume 2, Issue 7, July 2010, Issn 2151-9617.

[8] Smita Biradar, V.S. Malemath, Suneel C Shinde," Word-wise Script Identification of South Indian Document Images", International Journal of Advanced Research in Computer and Communication Engineering,,Vol. 4, Issue 8, August 2015

[9] 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

BIOGRAPHIES

Sugaladevi Choukimath received B E degree in Computer science and engineering from visveshwaraya Technological University of Belgaum in 2013.Currently pursuing her M.Tech degree in visveshwaraya Technological University of Belgaum.Her research intresets include Image Processing and Big Data Analytics

Dr.Virendra S Malemath is currently working as Professor in Computer Science & Engg, KLE Dr. M S Sheshgiri college of Engg & Tech, Belagaum.He did his Bachelors of Engg in Electronics & Communication Engg from Karnataka Universiyt Dharwad in the year 1993,did his MS in Software Systems from BITS Pilani Rajasthan in 1998 and received his Ph.D in Computer science from Gulbarga University,Gulbarga in 2009. His research interests are Document Image Processing , medical and pattern recognition.He has published more than 75 articles 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 Karnataka 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.2016.55115