STAR
Research Journal

Available online at www.starresearchjournal.com (Star International Journal)

COMPUTER SCIENCE

UGC Journal No: 63023

TEXT EXTRACTION AND LANGUAGE IDENTIFICATION FROM DIGITAL DOCUMENTS

VINAYAK JADHAV[1] NAGAVENI KADAKOL[2]

Department of Computer Science and Engineering, Smt. S. R. Nirani, Government Polytechnic Bilagi, Bagalkot, India[1,2]

Email: vinayakkjadhav.cs@gmail.com[1], nagavenikadakol@gmail.com[2]

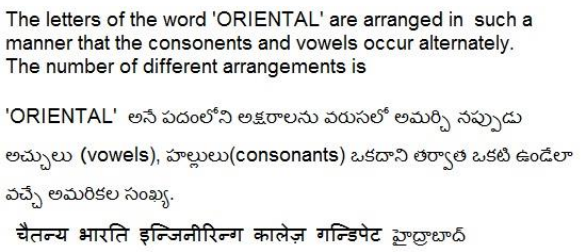
ABSTRACT

a Language classification is an essential feature in order to reach large community of people, a document contain more than one Language. Identification of these languages using OCR is practically a difficult task because language type should be predefined before applying to Optical Character Recognition (OCR) system. in turn it is much complicated to design one specific tool which can find a different types of languages. So , it is much needed task to find different part of the language portion of the script before sending the digital copy to the Optical Character Recognition tool. classification refers to extract information from scanned copy of physical documents such as letterhead, result copies, magazines and journals contain many languages for classification.

Index Terms-KNN, EDGE, PNN.

1. INTRODUCTION

Now a days the use of physical documents are converted into electronic document to facilate easy to store and retrieve easily for prolong duration. However, till today most of the places physical documents is major form of communication. as an example, the fax machine plays important means of communication worldwide. work carried out deals with physical document Hence, there is an requirement for an software, which performs multiple task such as retrieve, process or analyse and stores information from hard copyl of documents for later processing[5]. in Fig. 1.1 Example for such pages contain many languages in one document.


The letters of the word 'ORIENTAL' are arranged in such a manner that the consonents and vowels occur alternately. The number of different arrangements is

'ORIENTAL' అనే పదంలోని అక్షరాలను వరుసలో అమర్చి నప్పుడు అచ్చులు (vowels), హల్లులు(consonants) ఒకదాని తర్వాత ఒకటి ఉండేలా వచ్చే అమరికల సంఖ్య.

चैतन्य भारति इन्जिनीरिन्ग कालेज़ गन्डिपेट हैद्राबाद

Figure 1.1 many languages in one document.

A single script might contain two or more languages as shown in Fig. 1.1. main task is to read text data by its own for the given document using image processing technique.OCR can extract the data of an languages which are predefined. So It is difficult to feed OCR with different languages as a predefined. Above problems can be resolved by generating script classification systems. This addresses to design new tool which can perform multiple task such recognize, analyze and classify, from various documents .

From the Fig. 1..2. important observation noted that most of the characters in Telugu/Kannada languages contain tick shaped components at the top region of their characters.as well as, it has been be observed that measure part of Telugu characters contain upward blends present at their bottom region. These unique properties of Telugu characters are helpful in separating them from Hindi and English languages.

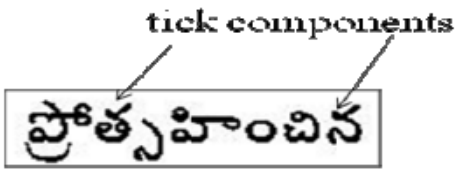


Figure 1.2 Example of Telugu word.

one key feature of Devanagari script is most of the character contain horizontal line at the upper portion which is called as head line and it is named as *sirorekha in* Devanagari script as shown in Fig. 1.3. these head lines are basic building block to form complete word by joining multiple basic charcter and they act as supporting feature in recognizing Devanagari script.

Hindi text differs from kannada and English text based on Another important feature of head line is that there is a close resemble between pixel found in headline and pixel found in bottom profile so,both bottom and top profile appears in top portion of the character text in hindi this key feature is not present in English and kannada text.

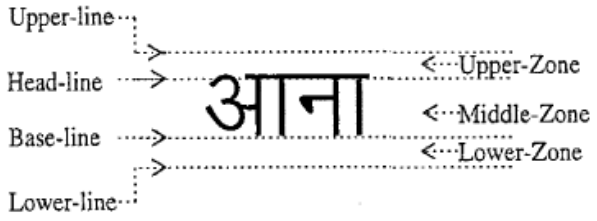


Figure 1.3 Hindi Text with different portion

as we found hindi text from multilingual text document we can recognise English text based on key feature such that the regular and symmetric pixel distribution found in most of the characters in English language.This feature makes uniformity between density of pixels in bottom and top profile same. However, this key feature not available in other two languages hence this attribute helps in specifying visual features in proposed system.

Recognizing the different languages from multilingual script based on different portion of the script become hard for computer to comprehend them directly because of differences value obtained from different languages of different scripts are in distinct semantic level. Taking this into account, research to be carried out to perform document image analysis based on pixel distribution and visual attributes. Although the image captured from camera and scanned physical document are varied in the skew angle this makes camera based images are unpredictable. Therefore, it is hard for the computer to train camera based images in all possible skew angles.

2. EXISTING SYSTEM

The Currently working system on automatic script identification are having two method that are Local and Global . Local method objective is to extract the key features from the image document like of list of character, line. In contrast, for global approach fine segmentation not required since this approach rely on analysis of region based on two lines and global approaches applicable only where document is processed based on complete document at a time or paragraph script only. work carried out define two word-level script identification methods refer pixel distribution and different visual features to recognise

the script type from scanned copy of physical documents.

3.PROPOSED WORK

Two text based script identification methods are PNN approach and KNN based approach. Different visual feature in these two tools identifies type of the script form script contain many different languages. Modes of Classification are, wavelet transform applied to perform feature extraction and linear analysis for classification during Training features are selected randomly from sample script. obtained features are stored in the database. These features in database are obtained from top bottom, middle portion of the multilingual script consists of different structure and symbols respectively such as vertical and horizontal line, half rounded symbols. The extracted features are stored in a future library as a references.

In the next phase Languages are identified based on extracted information which was placed in database for classification purpose. Then these test samples are compared with corresponding features stored in feature library.

3.1.ALGORITHM

Step 1:scan the document contain different language and store it in database.

Step 2: Preprocessing of an image produce improvement of image data that suppresses unwilling distortions this will helps for further analysis.

Step3:Segmentation.

Step4: Extract the Eigen features, Store it in database.

Step 5: develop PNN (probabilistic neural network) or KNN (K nearest neighborhoods) for operation such as training extracted features and classification based on feature stored in database of images.

Step 6:after images are classified based on portions(features extracted), the software identifies the types of the languages present in the script.

4. IMPLEMENTATION

Once segmentation operation is completes , then precede with segmented image to extract features, then these features are stored in to a database, in order to train for classification of the single image contain many languages.

a) Initial operation to be performed on images is image pre-processing which increases a quality of an image by reducing noise, increasing contrast of an image, increasing brightness, changing quantized number of level, in order to increase the detection rate of required values in the different portion of an multilingual document.

b) Image Segmentation is the process of differentiate between object of interest and background. image segmentation, feature extraction, classification are the basic building blocks of digital image processing.

c) feature extraction

feature extraction based on portion of text extracted from database that are top, bottom and middle portion to classify languages.

5. RESULT

TABLE 6.1 Results of Heuristic Based Method

Language	C1	C2	C3
Kannada	150	135	90
English	320	286	91.22

Hindi	450	366	92.25

C1:Number of test. C2:Recognised correctly

C3:Accuracy

TABLE 6.2 Classification Results with Knn Based Method

Language	C1	C2	C3
English	550	472	92
Hindi	450	410	91.6
Kannada	400	360	90

C1:Number of test. C2:Recognised correctly

C3:Accuracy

Approach could successfully classify different language based on feature extraction. The KNN Method produce successful classification of script words with an average accuracy is less than the PNN. it is based on the ratio of individual values to total values of given input text images.

REFERENCES

[1] S.Chanda, U.Pal, "English, Devanagari and Urdu Text Identification", *Proc. International Conference on Document Analysis and Recognition,* 538-545.

[2] M.C.Padma, P.Nagabhushan, "Identification and separation of text words of Kannada, Hindi and English languages through discriminating features", Proc. 2nd National Conference on Document Analysis and Recognition, Mandya, Karnataka, 252- 260.

[3] M.C.Padma, P.Nagabhushan, "Horizontal and Vertical linear edge features as useful clues in the discrimination of multilingual (Kannada, Hindi and English) machine printed documents", Proc. National Workshop on Computer Vision, Graphics and Image Processing (WVGIP), Madhurai, 204-209.

[4] U.Pal B.B.Choudhuri, "Automatic Separation of Words in Multi Lingual multi Script Indian Documents", Proc. 4th *International Conference on Document Analysis and Recognition,* 576-579.

[5]A.H.Kulakarni,P.S.Upparman "Script Identification from multilingual text documents",IJAR,(2015)