International Journal of Research in Advent Technology (IJRAT) Special Issue
E-ISSN: 2321-9637
Available online at www.ijrat.org
National Conference on "Role of Information Technology in Social Innovations"
26th & 27th February 2019

Language Classification from Text Documents

Santosh Chinchali[1], S.K.Honawad[2]
Department of Information Science and Engineering ,Dr.P.G.H College of Engineering and Technology vijayapura ,Affiliated V.T.U Belagavi,Karnataka. India[1,2]
Email: santoshchinchali@gmail.com[1] ,shivakunar.honawad@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.classification 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 impossible to design single recognizer which can find a large number of languages. So it is necessary to identify the language region of the document before feeding the document to the corresponding Optical Character Recognition (OCR) system . classifiation aims to extract information presented in digital documents like articles, newspapers, magazines and e-books 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, the usage of physical documents is still prevalent in most of the communications. For instance, the fax machine remains a very important means of communication worldwide.work carried out deals with physical document So, there is a great demand for software, which automatically extracts, analyzes and stores information from physical documents for later retrieval. in Fig. 1.1 Example for such pages contain different languages in a single 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 హైద్రాబాద్

Figure 1.1 different languages in single document.

One script could be used to write more than one languages. For example, Devanagari script is used by Hindi, Marathi, Rajasthan, Sanskrit and Nepali languages. One important task of document image analysis is automatic reading of text information from the document image.OCR can extract the data of an languages which are predefined.So It is difficult to feed OCR with different languages as a predefined. This can be solved by developing script classification systems. This addresses the need of developing tools that can recognize and analyze varied documents.

It can be seen that, most of the Telugu/Kannada characters have tick shaped structures at the top portion of their characters as shown in Fig.1.2. Also, it could be observed that majority of Telugu characters have upward curves present at their bottom portion. These distinct properties of Telugu characters are helpful in separating them from Hindi and English languages.

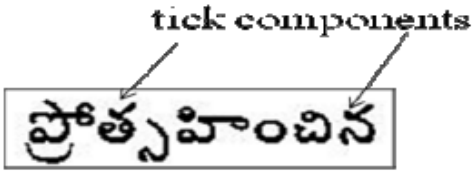

Figure 1.2 ExampleTelugu word.

It could be noted that many characters of Devanagari script have a horizontal line at the upper part called headline which is named as *sirorekha* in Devanagari as shown in Fig. 1.3. It joins two or more basic or compound characters to form a word. These head lines are present at the top portion of the characters and they are used as supporting features in identifying Devanagari script. Another strong feature in a Devanagari text line is that most of the pixels of the headline happen to be the pixels of bottom profile. This results in both top and bottom profiles of a Hindi text line to lie at the top portion of the characters. However this distinct feature is absent in both Kannada and English text lines where the density top and bottom profiles occur at different positions. Using these features Hindi text line could be strongly separated from Kannada and English languages.

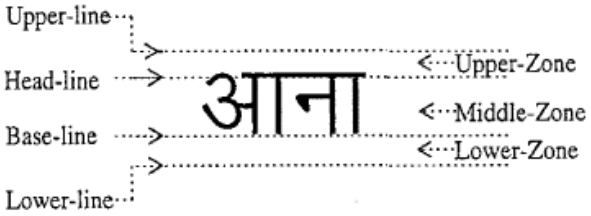

Figure 1.3 Hindi word with different portion

It is observed that the pixel distribution in most of the English characters is found to be symmetric and regular. This uniform distribution of the pixels of English characters results in the density of the top profile to be almost same as the density of the bottom profile. However, such uniformity found in pixel distribution of the top and bottom profiles of an English text line is not found in the other two anticipated languages Kannada and Hindi. Thus, this characteristic attribute is used as supporting visual features in the proposed model.

Although differences between different scripts are distinct in semantic level, its hard for computers to comprehend them directly. Taking this into account, research on special distribution and visual attribute is necessary for document image analysis. And abstraction of structure feature becomes the key technology of script identification. Although The skew of a camera-based image is often more severe and unpredictable than that of a scanned image. Therefore, it is difficult for a component-based approach to train an appropriate representative character set from images of all possible skew angles.

2. EXISTING WORK

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. and hence they are well suited to the documents where the script type differs at line or word level. most of the existing techniques are either line, sentence or block level in this paper define two word-level script identification methods namely PNN approach and KNN based approach. These methods use various visual features to recognise the script type from different language script documents.

3. PROPOSED WORK

Two word-level script identification methods are PNN approach and KNN based approach. These methods use various visual features to recognize the script type from different language script documents. 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. These features are stored in the feature library as a database. The database consists of distinct features of the scripts such as horizontal lines(head line),vertical lines and circle like structures in English and combination of vertical and horizontal, half rounded symbols and special symbols at the bottom. The features which are stored in database are also used for future references.

In the classification phase, based on the features extracted we classify the languages. The texture features are extracted from the test sample using the feature extraction algorithm and then compared with the corresponding feature values that are stored in the feature library.

4. IMPLEMENTATION

Work implementation involve preprocessing & segmentation, then extraction of features from the segmented image and storing of those features in order to train for classification of the Script image.

a) Image pre-processing is the name for operations on images at the lowest level of abstraction whose aim is an improvement of the image data that suppress undesired distortions or enhances some image features important for further processing and analysis task. It does not increase image information content.

b)Image segmentation is process is used to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. As the premise of feature extraction and pattern recognition, image segmentation is one of the fundamental approaches of digital image processing. Image Segmentation is the process that is used to distinguish object of interest from background.

c)feature extraction

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

Algorithm

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

Step 2: Preprocessing of an image to performed to suppress undesired distortions or enhances some image features important for further processing and analysis task.

Step3:Segmentation.

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

Step 5: Build the probabilistic neural network(PNN) or KNN for training & classification of images.

Step 6: Once the image is classified, the system shows the type script.

5. RESULTS

TABLE 6.1 RESULTS OF HEURISTIC BASED METHOD

Language	C1	C2	C3
Kannada	150	135	90
English	300	280	93.33

Hindi	400	365	91.25

C1:Number of test. C2:Recognised correctly
C3:Accuracy

TABLE 6.2 CLASSIFICATION RESULTS WITH KNN BASED METHOD

Language	C1	C2	C3
English	500	470	94
Hindi	450	410	91.6
Kannada	400	360	90

C1:Number of test. C2:Recognised correctly
C3:Accuracy

6. CONCLUSION

Approach could successfully classify different language based on feature extraction. The KNN based classifier could successfully classify script words with an average accuracy is less then the PNN. It is based on the average 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, (2005).

[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, (2003).

[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, (2002).

[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, (1997).

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