IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 7, September 2014.
www.ijiset.com

ISSN 2348 – 7968

Identification of Multilingual Words Using Profile Based Features

Kapil Kumar Kaswan [*], Manisha bansal**
[*] Department of Computer Science & Applications,Chaudhary Devi Lal Univeristy, India
**Department of Computer Science & Applications,Chaudhary Devi Lal Univeristy, India

Abstract

In a multi Language environment, majority of the documents may contain text information printed in more than one script/language. For automatic processing of such documents through Optical Character Recognition, it is necessary to identify different Language regions of the document. In this dissertation, it is proposed to develop a model to identify the Language type of trilingual words printed in Punjabi, Hindi and English Languages. The distinct characteristic features of Punjabi, Hindi and English Languages are thoroughly studied from the nature of the top bottom and middle profiles. The proposed model is applied on the words present in all the three languages. Experimentation conducted involved 200 text words for testing. The results are encouraging and prove the efficiency of the proposed model. The average success rate is found to be 94.50% for data set constructed from scanned document images and created data set.

Keywords: Multilingual words processing, Language Identification, Top Profile, Bottom Profile, Middle profile, Feature extraction

Introduction

Automatic language identification plays an important role in processing large volumes of document images, particularly for a multilingual OCR system. In addition, the ability to reliably identify the language type using the least amount of textual data is essential when dealing with document pages that contain multiple languages. An automatic language identification scheme is useful to (i) sort document images, (ii) to select specific OCRs and (iii) search online archives of document image for those containing a particular language In a multi-lingual country like India (India has 18 regional languages derived from 12 different scripts), a document page like bus reservation forms, question researches, bank challen, language translation books and money order forms may contain text lines in more than one language/language forms. Under the three language formulae, adopted by most of the Indian states, the document in a State may be printed in its respective official regional language, the national language Hindi and also in English. Accordingly, in Punjab, a state in India, generally any document including official ones would contain the text in three languages-English-the language of general importance, Hindi-the language of National importance and Punjabi –the language of State/Regional importance. Further there is a growing demand for automatically processing the documents in every state in India including Punjab. With this context, this report focuses on identifying the language type of a document containing only these three languages Punjabi, Hindi and English.

For automatic processing of such tri-lingual documents through the respective OCRs, a pre-processor is necessary which could identify the language type of the text words. In this thesis, it is proposed to develop a model to identify the text words of Punjabi, Hindi and English language from a trilingual document.

Data Collection

Standard dataset of Indian languages is currently not available. Data set construction with respect to the language identification problem seems to be complex since the factors like the font type and font size of each language needs to be considered. In this research, it is assumed that the input data set contains text words of the three languages - Punjabi, Hindi, and English. Also, it is

ISSN 2348 – 7968

assumed that the language type, font and size of the text words are same.

For the experimentation of the proposed model, three separate datasets are constructed, out of which one dataset is used to train the proposed system and the other two datasets are constructed to test the system. Thus separate data sets are constructed for training and testing. The words of English language were created using the Microsoft word software and these text words were imported to the Micro Soft Paint program. In the Microsoft Paint, a portion of the text word was saved as black and white gif image. The words of Hindi and Punjabi languages were created with newspaper cuttings of both Hindi and Punjabi languages present in the web sites of the news papers. These words were imported to the Micro Soft Paint program. In the Microsoft Paint, a portion of the text word was saved as black and white gif image.

To test the proposed model, two different data sets were constructed out of which one dataset was constructed manually similar to the dataset constructed for training and the other data set was constructed from the scanned document images. The printed documents like application forms, language translation books, manuals and magazines were scanned through an optical scanner to obtain the document image. The HP Scan Jet 5200c series scanner was used to obtain the digitized images. The scanning was performed in normal 100% view size at 300 dpi resolution. Manually constructed dataset were comprised of 300 text words and the data set constructed from the scanned document images were comprised of 200 text words from each of the three languages.

Pre-processing

Pre-processing is a method of enhancing the image for better feature extraction. The choice of pre-processing method to be adopted on a document image depends on the type of application for which the image is used. There are many techniques that are generally available to accomplish pre-processing on images; however, several experiments on language identification suggest that pre-processing methods have got to be customized to suit the requirements of language identification. Any language identification method requires conditioned image input of the document, which implies that the document should be noise free and skew free. Apart from these, some recognition techniques require that the document image should be segmented and threshold. All these methods, help in obtaining appropriate features for language identification processes.

In this paper, the pre-processing techniques such as noise removal and skew correction are not necessary for the datasets that are manually constructed by downloading the documents from the Internet. However, for the datasets that is constructed from the scanned document images, pre-processing steps such as removal of non-text regions, skew-correction, noise removal and binarization is necessary. In this research, text portion of the document image was separated from the non-text region manually. A global thresholding approach was used to binarize the scanned gray scale images where black pixels having the value 0's correspond to object and white pixels having value 1's correspond to background. The text area is segmented from the document image by removing the upper, lower, left and right blank regions. It should be noted that the text block might contain lines with different font sizes and variable spaces between lines. It is not necessary to homogenize these parameters, as the input to the proposed model is the individual text words.

The Proposed Model

The new model is inspired by a simple observation that every script/language defines a finite set of text patterns, each having a distinct visual appearance, which serves as useful visual clues to recognize the language. The

character shape delanguageors take into consideration any feature that appears to be distinct for the language and hence every language could be identified based on its discriminating features.

The proposed approach has adopted the concept of the top and bottom profiles of the input text word. The two languages – Hindi and English are identified using only one feature, which is obtained by computing differences of the black pixels in top max row of the top profiles. The method is not applicable for the trilingual documents as only one feature is used. With this backdrop, in this research, a new model has been proposed that uses the concept of top, bottom and middle profiles of a connected component. However, the new proposed method uses features extracted from the top, bottom and middle profiles of an input text word to identify the three anticipated languages - Punjabi, Hindi and English. The terms top profile, bottom profile and middle profile of a text word are defined below:

Top profile, Bottom profile and Middle profile

The top profile (bottom profile) of a text line represents a set of black pixels obtained by scanning each column of the text word from top (bottom) until it reaches a first black pixel. Thus, a component of width N gets N such pixels. Middle profile is computed by removing the header line and bottom part of the given word. The top profile, bottom profile and middle profile of a text words are obtained through the algorithms 1, 2 and 3 respectively.

Algorithm 1: Top profile ()

Input: Pre-processed input text word - Matrix a.

Output: Top profile - Matrix b.

1. Initialize matrix b=ones (size (a)) // the elements of the matrix b are initialized to 1's.

2. Do for j =1 to n columns

{Do for i= 1 to m rows

{If (a (i, j) == black)

{b (i, j) = a (i, j) exit }

Else continue

}

3. Return Matrix b.

Algorithm 2: Bottom profile ()

Input: Pre-processed input text word - Matrix a.

Output: Bottom profile - Matrix c.

1. Initialize matrix c=ones (size (a)) // the elements of the matrix c are initialized to 1's.

2. Do for j =1 to n columns

{Do for i = m down to 1 rows

{If (a (i, j) == black)

{c (i, j) = a(i , j) exit }

Else continue

}

3. Return Matrix c.

Algorithm 3: Middle profile ()

Input: Pre-processed input text word - Matrix a.

Upr/btm // upper/bottom first/last row containing black pixels.

Output: middle profile - Matrix d.

1. Initialize matrix d=ones (size (a)) // the elements of the matrix c are initialized to 1's.

2. Do for j =1 to n columns

{Do for i = 1 to upr+2 rows

{If (a (i, j) == black)

{d (i, j) = white continue }

Else continue

}

3. Do for i = btm to m rows

{Do for j =1 to n columns

{If (a (i, j) == black)

{d (i, j) = white continue }

Else continue

ISSN 2348 – 7968

}

4. Return Matrix d.

Algorithm 4: Straight line view ()

Input: Pre-processed input text word - Matrix a.

Upr/btm // upper/bottom first/last row containing black pixels.

Output: middle profile - Matrix e.

1. Initialize matrix e=ones (size (a)) // the elements of the matrix c are initialized to 1's.

2. Do for j =1 to n columns

{If (a (upr, j) == black)

Do for i = upr to btm rows

e (i, j) == d (i, j) continue }

Else continue

3. Return Matrix e.

Discriminating Features of Punjabi, Hindi and English text words

It has been observed that the three languages - Punjabi, Hindi and English considered in this research possess their own distinct features. These distinct features could be used as supporting features in the process of language identification system. It could be observed that most of the Punjabi characters have horizontal line like structures present at top portion of the characters. The pixels of these horizontal lines Happen to be the pixels of the top profile. Also, it could be observed that majority of Punjabi characters have not touch its matra present at their bottom portion and also matra present at the middle or we can say the matra is present at just right of the character is contain half height as comparison to Hindi.

Many characters of Hindi language have a horizontal line at the upper part called headline or sirorekha. It could be seen that, when two or more characters are combined to form a word, the character headline segments mostly join one another and generates one long headline for each text word. These long horizontal lines are present at the top portion of the characters. The pixels of these horizontal lines happen to be the pixels of the top profile.

It is observed that the most of the English characters are symmetric and regular in the pixel distribution. 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 uniform distribution of the pixels in top and bottom profiles of an English text line is not found in the other two anticipated languages - Punjabi and Hindi. Thus, this characteristic attribute is used as a supporting feature to separate an English text word.

Feature Extraction from Top, Bottom and Middle Profiles

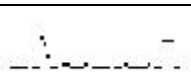
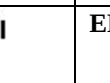
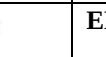
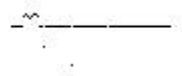
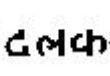
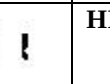
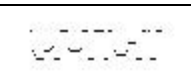
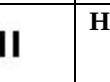
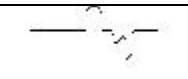
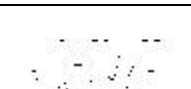
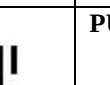
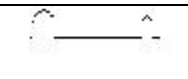
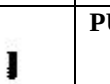
Choosing suitable features useful for discriminating the different text words from a set of trilingual words is an important step. By thoroughly studying the nature of the top, bottom and middle profiles of the three languages, a set of distinct features that yield discriminating values are extracted

- **Use of Top profile:** top profile is used to identify English language because English does not contain headline (siroh rekha) which is obtain in the top profile. Both Hindi and Punjabi contain headline.
- **Use of Bottom profile:** bottom profile is used to separate Hindi and Punjabi because Punjabi is different from Hindi in bottom when we analyze both languages as an image. The bottom matra of Hindi touches the character but this is not happen in Punjabi.
- **Use of Middle profile:** middle profile is used to separate Hindi and Punjabi because Punjabi is different from Hindi in middle when we analyze both languages as an image. The middle matra of Punjabi is half in height with respect to Hindi.All the experimental work is done with the help of a software tool named MATLAB. Matlab commands are used to process the document

ISSN 2348 – 7968

image. All the experimental work is describe in the following steps.

- Firstly convert the image in binary matrices.
- Create top, bottom and middle profile.
- In top profile find the limit of the word means first and last black element from left right
- Check whether the headline present or not with the assumption that top profile must contain a line with 75% black pixel of the total pixel in the line limit from left to right.
- If there is no such line then the word is related with English.
- After that bottom profile is come into picture.

WORD	TOP PROFILE	BOTTOM PROFILE	MIDDLE PROFILE	STRAIGT LINE VEIW	LANG.	RESULT
ENGLISH					ENG	PASS
VILAY					ENG	PASS
तेंदुलकर					HIN	PASS
लगाना					HIN	PASS
ਦੁਨੀਆ					PUN	PASS
ਮਿਠਾਂ					PUN	PASS

Matra present in bottom are different in Hindi and Punjabi. In Hindi matra touch the character but this is not happen in Punjabi.

ISSN 2348 – 7968

- At the same time also check the middle profile because matra present in the middle of Punjabi is usually less in height as comparison to Hindi.
    - Also straight line view is created from the middle profile is use to separate Hindi and Punjabi words according to their middle matra.
    - The algorithm for top, middle, bottom and straight line view is describe in above
    - After this step Hindi and Punjabi is separated

Results and discussion

The proposed system is also tested with scanned document images obtained from the text portions of application forms, manuals, language-translation books and such other trilingual documents. Data set of 200 text words from each of the three languages was considered from scanned images. Scanned images having text words in various font type and font size are considered. Experiments conducted indicate that scanned images are recognized efficiently with an average recognition rate of 99.5%. The average success Rate printed scanned data set has found to be 94.25%. This indicates the effectiveness of the proposed algorithm.

Conclusion

In this thesis, a new method to identify text words of the Punjabi, Hindi and English Languages. Experimental results show performance of the proposed model. The proposed model is developed based on the distinct features extracted from the top, bottom and middle profiles of the individual text words. The method looks simple, as it does not require any character segmentation. Experimental results demonstrate that relatively simple technique can reach recognition rate of 94.25% for data set constructed from scanned document images and created data set. The performance of the proposed algorithm is encouraging when the proposed algorithm is tested using manually created data set. However, the performance slightly comes down when the algorithm is tested on scanned document images due to noise and skew error.

References :

1 Spitz, A.L. ; Daimler Benz Res. & Technol. Center, Palo Alto, CA Pattern Analysis and Machine Intelligence, Browse Journals & Magazines IEEE Transactions on Volume:19 , Issue: 3(1997)

2 U.Pal, B.B.Choudhuri, “Script Line Separation From Indian Multi-Script Documents”, 5th Int.Conference on Document Analysis and Recognition (IEEE Comput. Soc. Press), 406-409,13(1999)

3 D Dhanya, AG Ramakrishnan and Peeta Basa Pati” Script identification in printed bilingual documents” Sadhan ¯ a¯ Vol. 27, Part 1, February 2002, pp. 73–82. © Printed in India.

4 Patil, B., SubbaReddy, N.V.: Neural network based system for script identification in Indian documents. In: Sadhana, India, vol. 27(part 1), pp. 83–97 (2002)

5 M. C. Padma and P.Nagabhushan, “Identification and separation of text words of Telugu, Hindi and English languages through discriminating features”, in proc. of Second National Conference on Document Analysis and Recognition, Karnataka, India, pp. 252-260, (2003)

6 S. Lu and C.L. Tan, “Script and Language Identification in Degraded and Distorted Document Images,” *Proc. 21st Nat'l Conf. Artificial Intelligence,* pp. 769-774, 2006

7 Dhandra.B.V, Nagabhushan. P, Mallikarjun Hangarge, Ravindra Hegadi, Malemath. V.S, “Script Identification Based On Morphological Reconstruction In Document Images “The 18th International Conference On Pattern Recognition (ICPR’06), pp 950-953(2006)

8 B.V.Dhandr and Mallikarjun hangrange “On Separation of English Numerals from Multilingual Document Images”; Journal of multimedia 2 (6), pp26-33(2007)

ISSN 2348 – 7968

9. V.N. Manjunath Aradhya, , G. Hemantha Kumar, S. Noushath "Multilingual OCRsystem for South Indian scripts and English documents: An approach based on Fourier transform and principal component analysis, Engineering Applications of Artificial Intelligence 21 (4), pp 658-668 (2008)

10. Prakash, Aithal, K., Rajesh, G., Dinesh Acharya, U., Krishnamoorthi, M., Subbareddy, N.V.: Text Line Script Identification for a Tri-lingual Document. In: International Conference on Computing Communication and Networking Technologies, Karur, India, pp. 1–3 (2010)

11. M.C padma "Word Level Script Identification in triilingual Documents", IEEE – ICSCN , MIT Campus, Anna University, Chennai, India. pp.630-635. (2010)

12. Prakash K. Aithal, Rajesh Gopakumar, and Dinesh U. Acharya, "Script line separation from Indian multi-Script documents," Journal of Computing, Volume 2, Issue 11, Nov 2010, ISSN 2151-9617 pp 107-111

13. Prakash, Aithal, K., Rajesh, G., Dinesh Acharya, U., Krishnamoorthi, M., Subbareddy, N.V.: Text Line Script Identification for a Tri-lingual Document. In International Conference on Computing Communication and Networking Technologies, Karur, India, pp. 1–3 (2010)

14.Priyanka P. Yeotikar,and P. R. Deshmukh:" Script Identification of Text Words from Multilingual Indian Document" International Journal of Computer Applications (0975 – 8887) National Level Technical Conference "X-PLORE pp.22-29 (2013)

15.P. Ramanathan "Automatic Identification of Handwritten Scripts" Middle-East Journal of Scientific Research 19 (7): 933-936, 2014 ISSN 1990-9233 © IDOSI Publications, 2014