Identification of Printed Bilingual Documents (Odia + English)

Prangya Paramita Pradhan[1], Priti Priyadarsani Pradhan[2]

[1]*Lecturer, Department of Electronics and Instrumentation Engineering, College of Engineering And Technology, Bhubaneswar, India*

[2]*Lecturer, Department of Computer Science and Engieering, College of Engineering And Technology, Bhubaneswar, India*

***Abstract*: This work deals with, identification of Oriya and English scrips from bilingual documents using an Optical Character Recognition (OCR) system. In the proposed system, first the document image is captured by flatbed scanner and then passed through preprocessing module for skew detection and correction. Thereafter, line segmentation is done. Then the word segmentation is performed. The main characteristic of this scheme is that, different scripts are identified, as Oriya script, Roman script and bilingual script, during the word segmentations, phase itself. In this project the words are differentiated with the help of vertical stroke based feature.**

***Keywords*: Word segmentation, Bilingual Script Identification, Vertical Stroke feature.**

1. Introduction

Script is the orthography of writing system. Script is the collection of set of symbols having some, sort of rules and is represented in a graphic form on any media. That media may hit papers; leafs. stones, metal plates of electronic media. Script is one of the media for representing the languages. India is a multi-lingual multi-script country. Here a single document page may contain two or more language’s scripts. For that many of the documents in India are multi script in nature. Indian documents are written in three different languages. For example, money order form of Orissa is written in English, Hindi and the local language Oriya. It helps in many applications like office automation, cheque verification, business and data entry applications etc.

In this system the document image is first taken using a scanner. Then the image is processed through different preprocessing modules like skew correction, line segmentation, and word recognition [4]. This work is concerned with the character recognition of printed Bilingual script that is for both Oriya Roman scripts along with Roman numbers. Thus OCR development for bilingual script is difficult.

2. Properties of Scripts

By comparing the structures of the letters some properties of both the scripts are described below:

- There are 12 vowels and 38 consonants are present in Oriya alphabets. These are called basic characters basic characters [1,4]. The basic characters of Oriya scripts as shown in Figure 1. When the basic characters follow any vowel is known as ‘matra’. Matras attached to the first consonant ‘ka’ (K). Besides that, Oriya scrips are also having some compound characters i.e. consonant follow other consonants in the characters. In Oriya scripts consonants follow another consonant that type character is known as compound character. For that reason, in Oriya nearly 200 characters have to be recognized. But where as in Roman 52 characters have to be recognized.
- The word of both the scripts are present in upper zone, lower zone and middle zone. It is shown in Figure 4. But in Roman scrip lower part of ‘j’, ‘g’, ‘p’, ‘q’ are present in lower zone. Where as in Oriya scripts lots of characters are present in lower zone. Because in Oriya scripts matra are used in upper zone and lower zone.
- The structures of Roman alphabets contain more straight and slant shape whereas Oriya scripts are round and curvy in nature.
- If matras of upper zone and lower zone are extracted, then Oriya scripts are in the same level-but whereas Roman scripts the characters are not in the same level as shown in Figure 1.

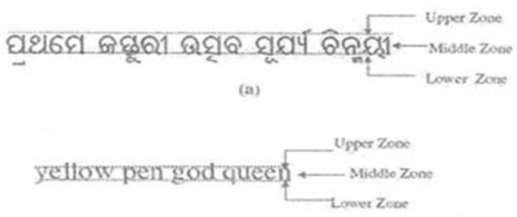

Fig. 1. Characters present in upper and lower zone.

3. Systems Description

Skew correction and line segmentation is same as like monolingual OCR systems. Here the word identification is different from the other system [1], [3], [4].

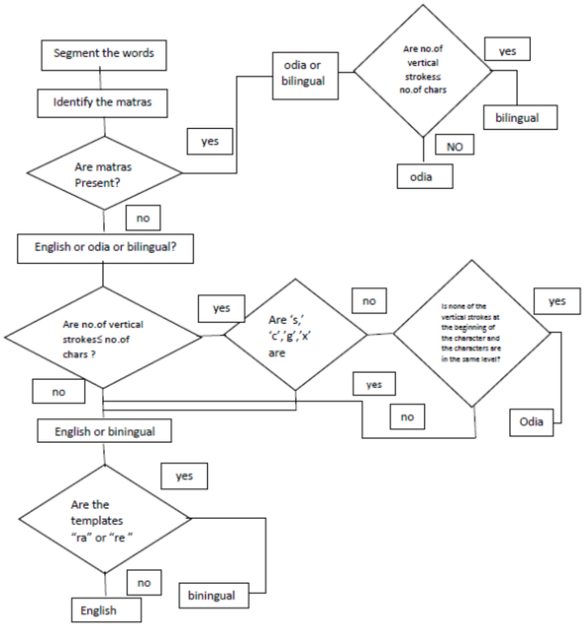
Segment the words
Identify the matras
Are matras Present?
yes
odia or bilingual
Are no.of vertical strokes≤ no.of chars
yes
bilingual
NO
odia
no
English or odia or bilingual?
Are no.of vertical strokes≤ no.of chars ?
yes
Are 's,' 'c','g','x' are
no
Is none of the vertical strokes at the beginning of the character and the characters are in the same level?
yes
Odia
yes
no
no
English or biningual
yes
Are the templates "ra" or "re "
no
English
biningual

Fig. 2. Flow chart for word identification in segmentation level

4. Word identification by vertical stroke feature

Feature is a vector which differentiates the specific image signal from the other image signals. In properties of scripts, it is observed that English letters are straight and slanted. The straight part of the character is considered as Vertical stroke feature as given in flow chart Figure 2.

Steps for Word identification.

- *Step 1.* Characters of roman scripts are not at the same level. However, it matras item Oriya characters are extracted. Then all the characters would be at the same level. Some of the English words can be classified using this concept.
- *Stcp 2:* A word is classified as Oriya or English depending upon the presence or absence of matras.
- *Step 3:* The vertical stroke feature in a word is identified. In roman script number of vertical strokes is greater than number of chanters. If number of vertical stroke is greater than number of character, it is Roman else Oriya.
- *Step 4:* However, in some cases Oriya scripts are also contained greater number of vertical strokes as compared to the number of characters. In this case, the position of the vertical stroke is to be considered, as it is always present. at the end of letters in Oriya script.
- *Step 5:* Some letters like "S", "Q" "g", "C" do not have tiny vertical stroke. Template matching using correlation method is applied for classification of these letters.

5. Results and Discussion

Considering the above steps, it is tested separately for Oriya and. English characters for large amount of data. A paragraph of Oriya script is tested and its output is shown in Figured. Similarly, it is also tested for English script and the output graph is shown in Figure 3.

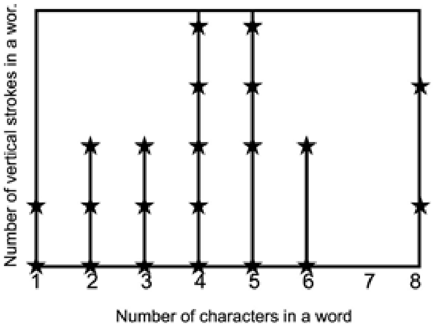

Fig. 3. No. of vertical strokes w.r.t individual for odia script

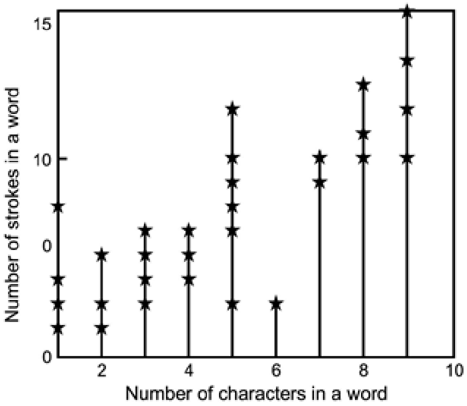

Fig. 4. No. of vertical strokes w.r.t individual for odia script

6. Conclusion

The method is quite simple, if the line segmentation is perfect and the matras are identified in line segmentation. Template matching by correlation is necessary some times. So further modification is needed.

References

[1] D Dhanya and A.G. Ramakrishnan "Script Identification in Printed Bilingual Documents" Springer-Verlag Berlin Heidelberg 2002 pp. 13-24, 2002.

[2] D Dhanya, A Ramakrishnan and Peeta basa pati 'Script identification in printed bilingual documents" Sadhana Vol. 27, Part 1, February 2002, pp. 73-82.

[3] Peeta Basa Pati, S Sabari Raju Nishikanta Pati, A G Ramakrishnan, 'Gabor filters for Document analysis in Indian Bilingual Documents' 2004 IEEE.

[4] B. B. Choudhury, U. Pal and M. Mitra, 'Automatic recognition of printed Oriya script', Vol. 27, Part 1. February 2002. pp. 23-34.

[5] S Mohanty, and H K Behera." A complete OCR Development System for Oriya Script". Proceedings of SIMPLE' 04, IIT Kharagpur, 2004.

[6] U. Pal, and B. B Chaudhuri, "Script Line Separation from Indian Multi-Script Documents". IETE Journal of Research, 49, 3-11, 2003.

[7] Srihari, S.N. and Govindaraju. Analysis of textual images using the Hough transform. Machine Vision, pp. 141- 153, 1989.