International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 10
October 2017

Ground Truth Creation for Odia Historical Document Character Analysis and Recognition

Mamata Nayak , Ajit Kumar Nayak
S'O'A University, Bhubaneswar, India

ABSTRACT

In the Completely Automated recognition systems for any language (i.e. to eliminate Computer and Humans interaction) large set of characters or symbols serves as a connecting link between the preceding and the following. Odia language is the 10th popular language in India, 33rd language of world and is used by around 40 million peoples as native language in the state of Odisha. Existence of a large dataset of printed Odia documents is a vital step in the Optical Character Recognition system as its accuracy greatly influences the overall recognition performance. It also assists the researcher to implement new methods or techniques for analysis and to compare their results with least effort. Many researchers ended their efforts towards the recognition of the Odia script, however practically no such standard data sets are available in any public domain for Odia language. We aim at autonomously collecting of all text from a single letter-size page based only on the information the page contains. This paper aims to extract characters by considering 50 different fonts, again both noise-less and noisy version from each of the print.

KEYWORDS:

Printed Data Set, Document Segmentation, Odia documents, OCR

1. PROPERTIES OF ODIA CHARACTER SET

To design the data set it is necessary to understand the characteristics of the language clearly [3-5],

i. The Odia script is developed from the Kalinga alphabet, one of the many descendants of the Brahmi script of ancient India

ii. It consists of 11 vowels, 37 consonants and 10 numerals. The language is characterized by its circular form in most of the characters. Few of those characters, have been appended with a vertical line as shown in the Tab.1.

a	e	kha	ga	gha	ŋa	dha	ma
ଅ	ଏ	ଖ	ଗ	ଘ	ଣ	ଧ	ମ

Tab.1: character having vertical line

i. The reading and writing style is from left to right.

ii. All vowels can combine with consonants to modify them and a special symbol gets added at left, right, top or bottom of the consonant. The special symbol is known as *matra*. There are 10 matra's used by the script shown in Tab.2.

Matra	ା	ି	ୀ	ୁ	ୂ	ୃ	େ	ୈ	ୋ	ୌ
Ka(Consonant)	କ	କ	କ	କ	କ	କ	କ	କ	କ	କ
Consonant combine with matra	କା	କି	କୀ	କୁ	କୂ	କୃ	କେ	କୈ	କୋ	କୌ

Tab.2: Matra with character ka

Also there are four different diacritics symbols (ଂ, ଁ, ଃ, ଼) can combine with consonants.

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 10
October 2017

iii. Consonants also combine with other consonants, by following some rules, which form a new character known as *conjunct/compound* character. In the compound character, either the base character gets modified or a new symbol may be appended. At most three characters can combine, at a time, to form a compound character, some of them are shown in Tab.3.

2 or 3 Consonants	Conjuncts
ଲ + ଲ	ଲ୍ଲ(New symbol append)
ଙ + କ	ଙ୍କ(Base Character changed)
କ + ତ + ର	କ୍ତ୍ର (New symbol append)
କ + ଷ + ମ	କ୍ଷ୍ମ(Base Character changed)

Tab.3: Compound characters of Odia script

2. WORTH OF THE DATA

- Character set of Odia script are very complex in shape, and the research for recognition are carried out in generally two direction i.e. template matching and classification. Conversely both approaches need a large dataset.
- Integration of the script in the new field of information communicationtechnology that helps for visual impaired persons also causes the major necessity of research.
- Also challenging problem of this script, similarity in structure of different fonts as well as same character is exist in different form.
- The data set is very useful for the remaining area of research of this script.
- The data set is first free and online data set for the Odia language.

3. MATERIALS AND METHODS

The design specification used to prepare the data is shown in the Tab.4.

Subject Area	Computer Science
Application Area	Character Recognition
Data source location	Odisha
Acquiring of Data	Image Documents, Scanned Documents taken 300dpi
Experimental data obtained	2800 .jpg files and 9 Excel worksheets of image names with class labels

Tab.4: Specification table

At first we prepare a document of 352 characters of Kalinga font which is used by numerous historical ancient documents of the script and saved in three forms: Normal, Italic and Bold. Then create an images of .jpg (with no LZW compression) format of 300dpi and also obtained by using a scanner. The Fig. 1(a)(b) show part of the images obtained from both the approaches.

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 10
October 2017

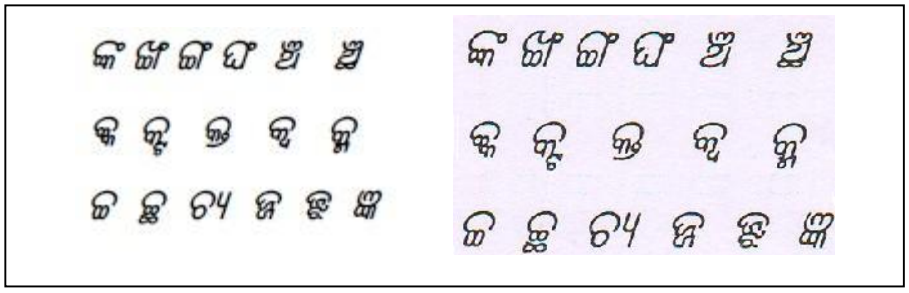

Fig.1 (a): Italic image document (b) Italic Scanned image

Then, the image is imported to a graphical user interface (GUI) which is designed using Matlab. For further processing of the image it needs to convert the gray scale image into binary images. So we use a threshold and represent the image in two values i.e. 0(white) and 1(black) and perform a morphological operation to the binary image until the image changes no longer. Since the image is consist of multiple lines, we use horizontal projection approach to extract each line and extend vertical projection to each of the obtained line to get each word. Connected component analysis is used to detect each individual separable symbol of the script. Each extracted sub images are named as shown below. As the property of the script few symbols are repeated, so we consider only single instant of a particular type as a result of which 2800 number of distinct symbols are collected few of them are shown in Fig2.

Fig.2. Elementary Odia symbols

Therefore 6 different types for a single character of the script helps the researcher to explore the techniques used in diversified field of research of this language.Fig. 3. show the character ‘ka/କ’ in its 6 different types.

13_111.jpg 13_112.jpg 13_113.jpg 13_112.jpg 13_122.jpg

Fig.3. Different form of an Odia symbols

To automatically explore the dataset and also use in different research direction of the whole dataset we propose analgorithm and its Matlab implementation code is shown below.

Algorithm:

Step1: Read an image file of .jpg or .tif

Step2: Convert RGB images to grayscale by eliminating the hue and saturation information while retaining the luminance

Step3: Find a global threshold (Th) that can be used to convert a grayscale image to an image of two intensity

Step4: If image_intensity>Th

Current_intnsity=1

else

Current_intnsity=0

```
    end
Step5: Perform a morphing operation such that the image changes no longer.
    For each pixel
        If 5 or more pixels of its 3×3 neighborhood are 1s
            Pixel_value=1;
        Else
            Pixel_value=0;
Step6: Perform horizontal projection of the image to extract lines
        For each line perform vertical projection to extract words
Step7: For each word connected component analysis has done to find each symbol
```

REFERENCES:

[1] P. Pujari, B. Majhi, “A Survey on Odia Character Recognition”, International Journal of Emerging Science and Engineering (IJESE), Vol-3 No-4, 2015
[2] Omar Bencharef, Younes Chihab, Nouredine Mousaid, Mustapha Oujaoura, “Data set for Tifinagh handwriting character recognition”, Data In Brief, 4, 11-13, 2015
[3] A.Alaei, P.Nagabhushan, U.Pal, “Handwritten Dataset, Ground Truth and Data Annotation”, International Journal of Pattern Recognition and Artificial Intelligence, Vol.26, No.4, 2012.
[4] S. Mohanty, H. K. Behera, “A Novel Approach for Bilingual (English - Oriya) Script Identification and Recognition in a Printed Document”, International Journal of Image Processing (IJIP), CSC Journals, Kuala Lumpur, Malaysia, 175 – 191, No. 2, Vol.4.
[5] M. Nayak, A. K. Nayak, "Odia Characters Recognition by Training Tesseract OCR Engine", International Journal of Computer Applications (0975 – 8887), pp.25-30, 2013