Advances in Computer Science and Information Technology (ACSIT)
Print ISSN : 2393-9907; Online ISSN : 2393-9915; Volume 2, Number 7; April – June, 2015 pp 77 – 80
Krishi Sanskriti Publications
http://www.krishisanskriti.org/ACSIT.html

An Efficient Algorithm for Characters recognition of Printed Oriya Script

Sneha Choudhary[1], Tushar Patnaik[2], Shweta Singh[3]

[1,2,3]CDAC, Noida, India
B-30, Sec-62
Noida
[1]9540543767, sneha.choudhary0106@gmail.com
[2]9811063041 tusharpatnaik@cdac.in
[3]9650457503, singh.shweta2721@gmail.com

Abstract: *Abstract –The subject of character recognition has received considerable attention in recent years. Character recognition is a process of converting handwritten or printed text images into machine readable code or text. Optical character recognition is used for many applications such as 1) Handwriting recognition systems, 2) Number plate recognition systems, 3) text recognition systems, 4) Data entry for business documents.*

In this, we are concerned with the recognition of printed Oriya script a popular Indian script. The development of OCR for this script is challenging as number of identified classes are more than 380.
In the proposed approach, the digitized document image is first passed through preprocessing modules. The preprocessed data is segmented to the symbol level using horizontal and vertical projection profiling. The proposed approach use HOG feature extraction technique and SVM (Support Vector Machine) is used as a classifier. This approach is able to distinguishing between characters that have very similar shapes.
A prototype of the system has been tested on a variety of printed Oriya material, and currently achieves 97.2% character level accuracy on average.
Keywords-Oriya script; HOG (Histogram of oriented gradient); optical character recognition (OCR); SVM.

1. INTRODUCTION

Character recognition of Oriya script is a topic which is being discussed among the researchers for long time. Character recognition is a process of converting handwritten, typewritten or printed text images into machine encoded code or text. Character recognition is the research field of pattern recognition, artificial intelligence & computer vision. In optical character recognition (OCR) many errors are not caused by inadequate classifier power, but by segmentation errors. Broken characters and merged characters constitute the major remaining problem. These problems occur during high-speed printing process in the machine vision industry. These problems are needs to be identified and corrected for the efficient OCR system. As a result, It is difficult for OCR system to segment and recognize characters properly. The number of characters in Oriya is large, and two or more characters may combine to form compound characters. Thus, the development of OCR for this script is challenging as number of identified classes are more than 380.

Our work is concerned with the techniques of character recognition of Oriya script. In the proposed approach, the digitized document image is first passed through preprocessing modules like binarization, noise cleaning, skew correction. The preprocessed data is segmented to the word and character level using the projection profiling. The merged symbols are identified and segmented into the sub-symbol by using water reservoir algorithm. The proposed approach is based on the HOG feature which is called Feature points. As Feature points increases results will be more accurate but complexity and time require for testing will be more. SVM (Support Vector Machine) is used as classifier. This approach is able to distinguishing between characters that have very similar shapes.

This paper is divided into five sections. Section II describes properties of Oriya script; Section III describes proposed method; Section IV describes Test results and Section V describes conclusion of an approach.

2. PROPERTIES OF ORIYA SCRIPT

We describe here some properties of the Oriya script that are useful for building the recognition system.

The alphabet of the modern Oriya script consists of 52 basic characters first 11 are vowels and rest 41 are consonants. The basic characters of Oriya script are shown in Figure 1.

The first vowel can occur only at the beginning of a word and never printed after a consonant in a word. Writing style in the script is from left to right. The concept of upper/lower case is absent in Oriya script.

In Oriya script a vowel (other than the first one) following a consonant takes a modified shape, which, depending on the vowel, is placed to the left, right (or both), top or bottom of the consonant. These are called modified characters.

Fig1. Basic characters of Oriya script

A consonant or vowel preceding another consonant sometimes takes a compound shape, which is termed as compound character. In some cases, it is represented by modifier called consonant modifier. Compound characters can be combinations of consonant and consonant, as well as consonant and vowel.

Fig2. Compound Characters.

In Oriya script, a text line may be partitioned into three zones. The upper-zone denotes the portion above the meanline, the middle zone covers the portion of basic (and compound) characters below mean-line and the lower-zone is the portion below base-line. An imaginary line, where most of the uppermost (lowermost) points of characters of a text line lie, is referred as mean-line (base-line). Examples of zoning are shown in Figure 3. In this case, the mean-line along with base-line partition the text line into three zones.

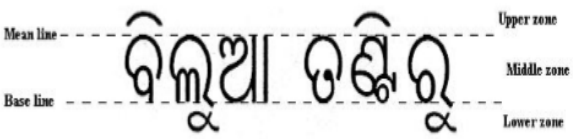

Fig3. Different Zones in Oriya Text line

3. PROPOSED METHOD

In this section we present our approach of Character Recognition of Oriya Script which includes various modules. Figure 4 shows an overview of the proposed method.

3.1 Preprocessing

Preprocessing includes binarization, noise cleaning and skew correction. Binarization can be described as the process of converting a grey scale image into one, which contains only two distinct tones, that is black and white. The intensity values of the grey scale images varies from 0 to 255 which we convert into 0 and 1 only. In this process, a global thresholding approach is used to binarize the scanned gray scale images where black pixels having the 0's correspond to object and white pixels having values 1's correspond to background.

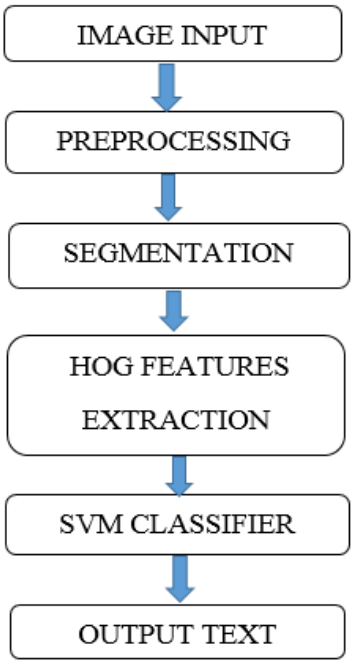

Fig4. Flow Diagram of Proposed Approach

Skew detection and correction are important preprocessing steps. Skew correction can be achieved in two steps, namely 1) Skew angle Estimation, and 2) Rotation of the image by the skew angle in the opposite direction. In our approach, we used a Hough transform based technique for skew angle estimation of Oriya script. The uppermost and lowermost points of most of the Oriya text line characters lie on the mean line and base line respectively. The lowermost and uppermost points of characters in a skewed Oriya text are shown in Figure 5.

Fig5. Uppermost and lowermost points of characters in a skewed Oriya text line.

3.2 Segmentation

The proposed OCR system automatically detects individual text lines, segment the words from the line, and then segment the characters in each word. Since Oriya text lines can be partitioned into three zones. It is convenient to distinguish these zones. Character recognition becomes easier if the zones are distinguished because the lower zone contains only modifiers, while the upper zone contains modifiers and portions of some basic characters.

3.2.1 Line Segmentation

The lines of a text block are segmented by finding the valleys of the horizontal projection profile computed by counting the number of black pixels in each row. The horizontal projection profile will have peaks at text line positions and valleys in between successive text lines shown in Figure 6.

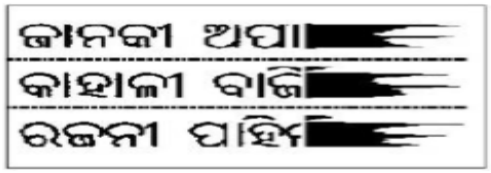

Fig6. Horizontal Projection profile of rows in Oriya text lines.

3.2.2 Zone detection

After line Segmentation, the zones in each line are identified by detecting the base and mean lines. We consider a set of horizontal lines passing through the uppermost and lowermost points of the characters of Oriya text line. The imaginary horizontal line that passes through the maximum number of uppermost points is known as the mean line. The imaginary horizontal line that passes through the maximum number of lowermost points is known as the base line.

3.2.3 Word and Character Segmentation

After a text line is segmented, it is scanned vertically to construct a vertical projection profile. Words and characters of a text line are segmented by finding the valleys of the vertical projection profile computed by counting the number of black pixels in each column. If in the profile there exists at least kl (threshold value) consecutive 0s then that is considered as a word boundary. The value of k1 is taken as 2/3 of the text line height (text line height is the normal distance between the mean line and the base line). To segment each word into individual characters, we consider only the middle zone of the word. Once again, the vertical projection profile of this zone is considered. A column that does not contain any black pixel marks the boundary between two characters. The vertical projection profile will have peaks at text positions and valleys in between characters shown in Figure 7.

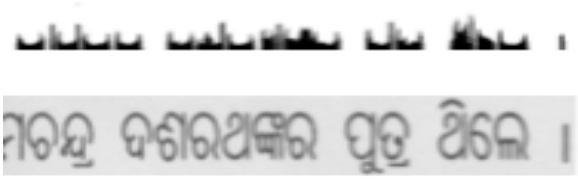

Fig7. Vertical Projection profile of columns in Oriya text line

3.3 Feature Extraction

Feature extraction is the identification of appropriate measures to characterize the components image distinctly. The two essential sub-stages of recognition phase are feature extraction and classification. The feature extraction stage analyzes a text segment and selects a set of features that can be used to uniquely identify the text. The aim of feature extraction is to identify minimum numbers of features that are effective in discriminating pattern classes. There are many popular methods to extract features. In our work, we used a HOG (Histogram of oriented gradients) features extraction technique. To extract HOG feature, an images is first divided into smaller patches and feature extraction procedure is applied in every patches separately. The orientation of gradient of each pixel within a patch is then quantized into histogram bins, while each histogram bin represents an angle range. After that, the histogram of each patch is normalized and concatenated together to form a feature vector. The derived features are used as an input to the character classifier (SVM).

3.4 Support Vector Machine

Next step is to classify or recognize characters using Support Vector Machine. The classification stage is the main decision stage of an OCR system and uses the extracted features as input to identifying the text segment. Performance of the system largely depends upon the type of the classifier used. In this, we use SVM (Support Vector Machine) Classifier. SVM classifier is a binary classifier which looks for an optimal hyperplan as a decision function. Classification is usually accomplished by comparing the feature vectors corresponding to the input text/character with the representatives of each character class, using a distance metric.

4. EXPERIMENTAL RESULT

A database of about 75 classes with 150 characters per class. The top recognition accuracy achieves is 97.2%.

Training data (in %)	Test data (in %)	BlockSize (in cells)	Recognition Rate (in %)
90	10	2 by 2	96.6
		2 by 3	97.5
		2 by 4	96.9

80	20	2 by 2	96.3
		2 by 3	97.1
		2 by 4	97.0
70	30	2 by 2	96.8
		2 by 2	97.4
		2 by 4	96.8
60	40	2 by 2	96.9
		2 by 3	96.9
		2 by 4	96.8

5. CONCLUSION

In this we proposed an approach to recognize characters of printed Oriya script. This approach is able to distinguishing between characters that have very similar shapes, but more work needs to be done. Our character segmentation method also needs to be improved so that it can handle a larger variety of touching characters, like overlapped merged characters.

6. REFERENCES

[1] Su Liang, M. Ahmadi, M. Shirdhar "Segmentation of Touching Characters in Printed Document Recognition" Proceedings of the Second International Conference on Document Analysis and Recognition, IEEE CONFERENCE PUBLICATIONS / ICDAR, pp 569-572, 2001

[2] Yi Lu "On the Segmentation of Touching Characters " Proceedings of the Second International Conference on Document Analysis and Recognition, IEEE CONFERENCE PUBLICATIONS/ICDAR pp 440 – 443, 2000

[3] B.B. Chaudhuri, U. pal, M. Mitra "Automatic recognition of Printed Oriya script" Proceedings of the Sixth International Conference on Document Analysis and Recognition @2001 IEEE

[4] Tadashi Hyuga, Hirotaka Wada, Tomoyoshi Aizawa, Yoshihisa Ijiri and Masato Kawade " Deformed and Touched Characters Recognition" 2013 Second IAPR Asian Conference on Pattern Recognition, IEEE, pp 744 – 745,2013

[5] T. Bayer, U. Krebel, M. Hammelsbeck "Segmenting Merged Characters" Vol.II. Conference B:Pattern Recognition Methodology and Systems, Proceedings, 11th IAPR International Conference on Pattern Recognition ©2002 IEEE,DOI-30 Aug-3 Sep 2002, Page(s): 346 - 349