Presented in the National Conference on Recent Advances in Microwave Tubes, Devices and Communication Systems, Jagannath Gupta Institute of Engineering and Technology, Jaipur, March 4-5, 2011

HANDWRITTEN ODIA CHARACTER RECOGNITION

Debasish Basa[1], Sukadev Meher[2]

Department of Electronics and Communication Engineering
National Institute of Technology, Rourkela, India
[1] debasishbasa@gmail.com
[2]sukadevmeher@gmail.com

***Abstract*—Odia is one of the oldest and popular languages of India, spoken by more than 44 million people, especially in Odisha, India. Some characters in Odia are made up of more than one connected symbols. Compound characters are written by associating modifiers with consonants, resulting in a huge number of possible combinations, running into hundreds of thousands. Therefore, systems developed for recognition of other scripts, like Roman, cannot be used directly for the Odia language. In the present work, we have proposed robust structural solution for Odia character recognition where, a given text is segmented into lines and then each line is segmented into individual words and then each word is segmented into individual characters or basic symbols. Basic symbols are identified as the fundamental units of segmentation used for recognition. Using unique structure of some characters we have found better result as comparison to other methods.**

I. INTRODUCTION

During the past thirty years, substantial research efforts have been devoted to character recognition that is used to translate human readable characters to machine readable codes. An immense effort has been spent on character recognition, because it provides a solution for processing large volumes of data automatically in a large variety of scientific and business applications. Handwriting is converted to the digital form either by scanning the written paper or by writing with special pen on an electronic surface such as a digitizer combined with a liquid crystal display [1]. The two approaches are distinguished as offline and online handwriting, respectively. In the online case, the two-dimensional co-ordinates of successive points of the handwriting as a function of time are stored in order [2]. For offline handwriting, only the completed writing is available as an image [3]. Offline systems are therefore less accurate than online systems.

Template matching, or matrix matching, is one of the most common classification methods [4],[5]. In template matching, individual image pixels are used as features. Classification is performed by comparing an input character image with a set of templates (or prototypes) from each character class. Each comparison results in a similarity measure between the input character and the template. One measure increases the amount of similarity when a pixel in the observed character is identical to the same pixel in the template image. If the pixels differ, the measure of similarity may be decreased. After all templates have been compared with the observed character image, the character's identity is assigned as the identity of the most similar template when correlation coefficient is maximum.

Template matching technique can used for a small set of postures, requires small amount of calibration, no advance learning of patterns and is quite accurate. But this technique does have limitations. The limitation is the small number of possible postures that can be recognized. If the application requires a large posture set, then template matching will not work better.

The main challenge in the handwritten character recognition involves a development of a method that can generate the description of the handwritten objects in a short period of time. In this study we propose a simple yet robust structural solution for performing Odia (the official language of Odisha) character recognition.

Rest of the paper is organized as follows. Section II describes character modelling. In Section III, character recognition using sub-structure based method is explained. The experiments and results are discussed in section IV, Finally conclusion of the paper is given in section V.

II. CHARACTER MODELLING

A. *Odia literature*

The number of characters in Odia is large. Two or more characters may combine to form compound character, as a result the total number of characters to be recognized is more than 200.

1) Properties of Odia Script: The properties of the Odia script that are useful for building the character recognition are:

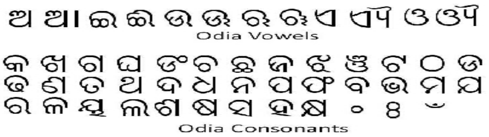

Figure 1. Set of Odia Vowels and Consonants

- The Odia basic characters consist of vowels and consonants which are shown in Fig. 1. As in other Indian scripts, the concept of upper lower case is absent here.
- The first vowel is never printed after a consonant in a word and can occur only at the beginning of a word.

ଗ୍	ଗ	ଗା	ଗି	ଗୀ	ଗୁ	ଗୂ	ଗୃ	ଗେ	ଗୈ	ଗୋ	ଗୌ
g	ga	gā	gi	gī	gu	gū	gṛ	ge	gai	go	gau

Figure 2. Modified vowels attached to the first consonant and some commonly occurring compound characters

- A vowel (other than the first one) following a consonant takes some modified shape as shown in Fig.2. Depending on the vowel, this modified shape is placed to the left, right (or on both sides), top or bottom of the consonant. The modified shapes are called modifiers or allographs. The vowel allographs do not disturb the shape of the basic characters (in the middle zone) to which they are attached.

- If the shape in the middle zone is altered by combining two or more consonants, the resultant shape is termed as compound character. In some cases, a consonant preceding or following another consonant is represented by a modifier called consonant modifier.

B. *Preprocessing*

It is necessary to perform several document analysis operations prior to recognition of text in scanned document. The common operations are:

1) Thresholding: The task of thresholding is the extraction of the foreground from the background [6]. The histogram of grayscale values of a document image typically consists of two peaks: one is corresponding to the foreground and another is corresponding to the white background. Hence, the task of determining the threshold grayscale value is the determining of an 'optimal' value in the valley between the two peaks. Two categories of thresholding are:

- Globally - picks one threshold value for the entire document image which is often based on an estimation of the background level from the intensity histogram of the image.

- Locally (Adaptive) - uses different values for each pixel according to the local area information.

2) Noise Reduction: Digital image can have noise, introduced from the scanning devices and/or transmission medium. In order to achieve an accurate result, all non-word data must be removed. There are three common type of noise in handwriting known as: background noise, shadow noise and salt and pepper noise. Smoothing operations are often used to eliminate the artifacts introduced during the image capture.
Two main approaches of noise reduction are:

a) Filter by masking.

b) Morphological Operations i.e by erosion, dilation.

3) Image Segmentation: Character Segmentation is a two stage segmentation process in which the subscripts of the word are removed first and then the individual characters are segmented. Image Segmentation plays a crucial role in Character Recognition [7]. If one views an image as depicting a scene composed of different objects, regions. Then segmentation is the decomposition of an image into these objects and regions by associating or 'labeling' each pixel with the object that it corresponds to.
There are two types of segmentation:

- Implicit Segmentation: The words are recognized entirely without segmenting them into letters. This is most effective and viable only when the set of possible words is small and known in advance, such as the recognition of bank checks and postal address.

- Explicit Segmentation: In explicit approaches one tries to identify the smallest possible word segments (primitive segments) that may be smaller than letters, but surely can-not be segmented further. Later in the recognition process these primitive segments are assembled into letters based on input from the character recognizer. The advantage of this strategy is that it is robust and quite straightforward, but is not very flexible.

a) Line Segmentation: The handwritten text must be divided first into lines.

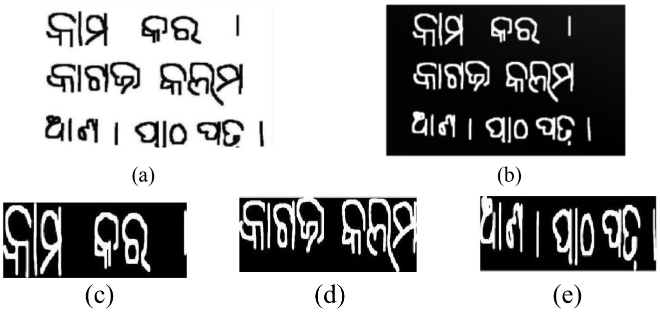

Figure 3(a) Odia handwritten text (b) Binary filter image of handwritten text (c) First line (d) Second line and (d) Third line of handwritten text after line segmentation.

b) Word Segmentation: For Odia script, spacing between the words is greater than the spacing between the characters in a word. This spacing between the words is used for word segmentation. The spacing between the words is found by taking the Vertical Connecting Pixel (VCP) of an input text line. VCP is the sum of ON pixels along every column of the image. In VCP, the width of the zero-valued valleys is more between the words in the line when compared to the width of zero-valued valleys that exists between the characters in a word. This information is used to separate words from the input text lines as shown in fig.4.

(a) (b)

Figure 4(a) First word and (b) Second word of first line

c) Character Segmentation: We know that Odia is a non-cursive script. So, spacing between the characters in a word is used for character segmentation as shown in fig.5. For word segmentation also VCP is used.

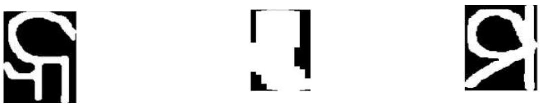
(a) (b) (c)

Figure 5(a) First character (b) Second character and (c) Third character of first word.

III. Character Recognition Using Sub-Structure Based Method

In Odia some characters contain a vertical line at the right most part. According to this property all the character are divided into two groups.

- Group I- A vertical line is not present at the right most side as shown in fig. 6(a).

- Group II- A vertical line is present at the right most side as shown in fig. 6(b).

GROUP-I CHARACTERS

ଇ ଈ ଉ ଊ ଋ ଓ ଚ ଛ ଜ ଝ ଞ ଟ ଡ଼ ଢ଼

ତ ଦ ନ ବ ଭ ର ଳ ହ

Figure 6(a)

GROUP-II CHARACTERS

ଅ ଆ ଏ ଖ ଗ ଘ ଣ ଥ ଧ ପ ମ ଯ ଶ ଷ ସ

Figure 6(b)

In Odia handwritten characters, the vertical line takes 20% out of the total width at the right most part. So in our work we have cropped only that portion of vertical line for detection. Figure 9 shows the flow chart for finding group-I and group-II characters. If all the rows of cropped image contain at least a non-zero element, than it is called ‘vertical line is present’. But most of the times vertical line is not exactly straight. So for this case we have to calculate the number of connecting rows.

If the number of connecting rows is greater than 65% from the total number of rows of the handwritten character, than it is also called ‘vertical line is present’. The pixel connectedness is checked from top to bottom of the cropped image.

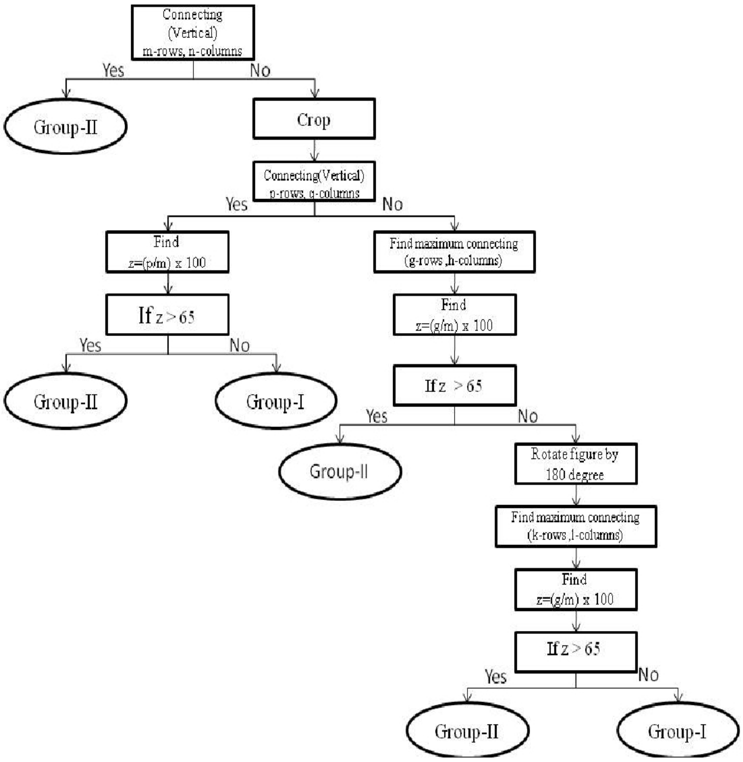
Connecting (Vertical) m-rows, n-columns
Yes
No
Group-II
Crop
Connecting(Vertical) p-rows, q-columns
Yes
No
Find z=(p/m) x 100
Find maximum connecting (g-rows ,h-columns)
If z > 65
Find z=(g/m) x 100
Yes
No
Group-II
Group-I
If z > 65
Yes
No
Group-II
Rotate figure by 180 degree
Find maximum connecting (k-rows ,l-columns)
Find z=(g/m) x 100
If z > 65
Yes
No
Group-II
Group-I

Figure 7.Flow chart for finding group-I and group-II characters.

Sometimes the connectedness is not greater than 65%. For these cases we have to rotate 180 degree and we have to calculate the connectedness. If the connectedness is not greater than 65% than it is called ‘vertical line is not present’ as shown in figure 8.

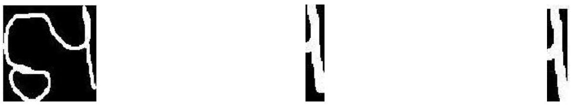
(a) (b) (c)

Figure 8(a) Odia Handwritten ‘Kho’ Character of size 100X100 (b) Cropped image of size 100X20 (c) Cropped image of only non-zero elements.

A. Recognition of Characters

Every Odia character that belongs to either group-I or group-II having at least one or more unique shape in some portion. We have to extract that portion and create a sub-image template database separately for group-I and group-II characters by using lots of handwritten samples of a particular character.

IV. Experimets And Results

When we take a test image of a character, first we have to find whether it belongs to group-I or group-II character as shown in fig.7. Then extract the unique shape as sub-image and match with either group-I sub-image template database or group-II sub-image template database.

Experiments are performed on different handwritten Odia characters. Instead of describing in detail, we are describing here only for one character which is a difficult task.

Recognition of Odia 'Kho' character: Recognition of handwritten Odia 'Kho' character is a difficult task, because it is almost similar to Odia 'Gaa' character as shown in fig. 9. The marked shown in fig. 9(a) and 9(b) is the only difference between Odia 'Kho' and 'Gaa' character.

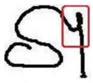

Figure 9(a)

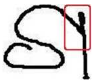

Figure 9 (b)

To distinguish between Odia 'Kho' and 'Gaa' character, two databases of the sub-image marked in fig.9 are created. Figure 10 shows the two template database where 35 samples of different handwritten characters are taken.

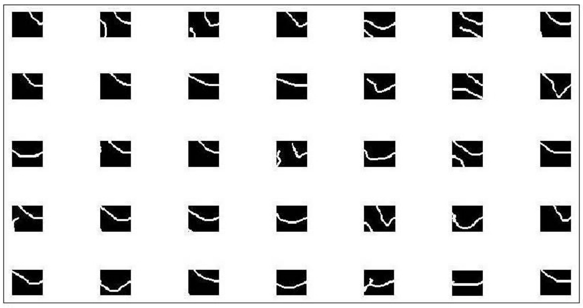

Figure 10(a). Database of the sub-image of 'Kho' character

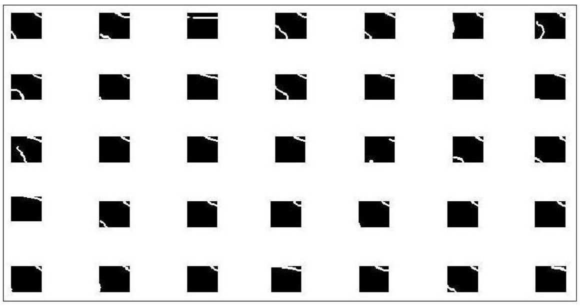

Figure 10(b). Database of the sub-image of 'Gaa' character

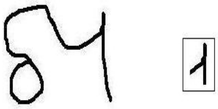

Figure 11(a) and (b)

Figure 11(a) shows the input handwritten Odia 'Kho' character. Before matching with the template databases, we have to extract the unique portion of the input handwritten character as shown in fig.11(b) that distinguish Odia 'Kho' and 'Gaa' character. Then matching is performed with the databases. Here we have calculated the correlation coefficient as similarity measure.

Figure 12(a) shows the correlation coefficients of the input sub-image with all the templates of 'Kho' character database, where we have observed the maximum value (>0.5) of correlation as compared to fig,12(b) which corresponds to 'Gaa' character.

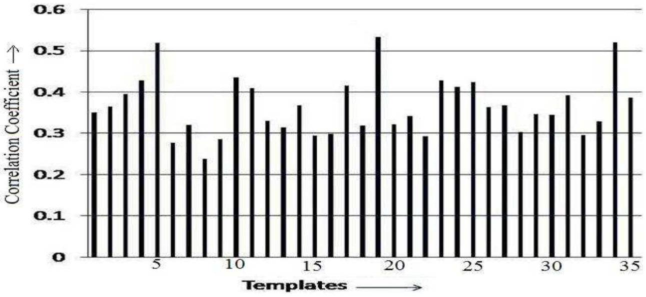

Figure 12(a)

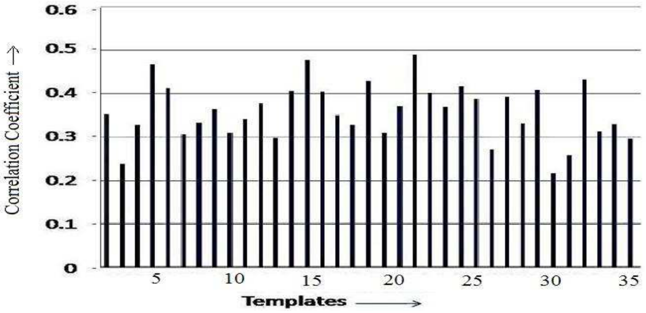

Figure 12(b)

From fig. 12, we conclude that the recognition rate of 'Kho' character is more than 'Gaa' character and hence the test image character is a 'Kho' character.

V. Conclusion

Recognition rate is highly affected by similarity of various characters. There are more similar characters which in turn degrade the recognition rate. We have treated individual image pixels as features, where each comparison results the similarity measure between the input character and the database. The comparison is performed on pixel by pixel basis.

References

[1] U. K.Roy, T.Pal and F. Kimura, "Oriya handwritten numeral recognition systems," computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata-108, India.

[2] V. S. C. H. Swethalakshmi, Anitha Jayaraman and C. C. Sekhar, "Online handwritten character recognition of devanagari and telugu characters using support vector machines," department of Computer Science and Engineering, Department of Biotechnology, Indian Institute of Technology Madras, Chennai - 600 036, India.

[3] T. W. U. Pal1 and F. Kimura2, "A system for off-line oriya handwritten character recognition using curvature feature," computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata-108, India.

[4] L. Song and Y. Lin, "Study on the vision reading algorithm based on template matching and neural network," in Proceedings of International Joint Conference on Neural Networks, ser. Orlando, Florida, USA, August 2007, pp. 12 – 17.

[5] R. S. P. Jayashree R.Prasad, Dr.U.V.Kulkarni, "Template matching algo-rithm fof gujrati character recognition," second International Conference on Emerging in Engineering and Technology,ICETET-09.

[6] R. C.Gonzalez and R. E.Woods, Digital Image Processing, 3rd ed. Pearson.

[7] B. D. Mohammad Isbat Sakib Chowdhury and M. S. Rahman, "Segmentation of printed bangla characters using structural properties of bangla script," in 5th International Conference on Electrical and Computer Engineering, ser. ICECE, 20-22 December 2008.