IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)
e-ISSN: 2278-2834,p- ISSN: 2278-8735.
PP 31-37
www.iosrjournals.org

Odia Offline Character Recognition using DWT Features

Bhabani Dash[1], Shibashis Pradhan[2], Debaraj Rana[3]
[1]*M.Tech Scholar, Dept. of ECE, Centurion University of Technology &Management, Odisha, INDIA*
[2,3]*Asst. Professor, Dept. of ECE, Centurion University of Technology &Management, Odisha, INDIA*

Abstract: *Optical character recognition (OCR) is the mechanical or electronic translation of images of handwritten or typewritten text (usually captured by a scanner) into machine-editable text. Many recognition techniques have been proposed so far for different languages. In this paper we have proposed a nobel method to recognize Odiya character which is the dominant language in the state of Odisha. The proposed method consists of three steps: The first two steps refer to creating a database for training using a set of documents, while the third one refers to recognition of new document images. First, a pre-processing step that includes image binarization and enhancement takes place. At a second step a segmentation approach is used in order to detect text lines, words and characters. For classification of character we have used a DWT based method for feature extraction. The proposed method is showing a good result in recognition of printed offline odia script.*

Keywords - *DWT , Enhancement, character Recognition*

I. Introduction

Character recognition [1] is one of the vital application areas in image processing. It is the process of automatic recognition of character from a document image. It is also coming under document image processing [2] to extract information inside a document. For processing a document it needs different operation which include preprocessing of document, skew correction, character segmentation, Character recognition then output generation [3].Basically the optical character recognition (OCR) focuses in English language or which are the standard optical characters.[4-6].Many researchers has done the optical character recognition for different regional languages. The most challenging part of character recognition is the handwritten character recognition, which is more crucial than OCR. In odia language also different technique has been developed for recognition of Odiya character [7-9] which is the official language of Odisha We have developed a more accurate technique to extract and recognize odiya character which is based on DWT based feature extraction [10-12].

The most application of character recognition is to read the printed character and make the document for editable and machine readable. The character recognition is quiet challenging because most of characters are of similar type. So to make it easier to recognize we have developed a DWT based feature extraction and recognition. Character recognition basically classified into two category one is offline character recognition and the other is online character recognition [13-15]. Offline recognition deals with typewritten characters where image formatted character converted into text readable character. Where as in case of online character recognition which involve direct capture of text image and recognize the characters automatically simultaneously. For recognition purpose some OCR technique classified into again two category, one is template based and the other is feature based. In case of template based template for every character are used for matching with the extracted character and after matching it will be recognize with the help of template sample [16]. In case feature based technique the features are extracted based on certain criteria, and the features are used to recognize the text [17].

Odia is one of the scheduled approved languages of India. In Odisha it is the mostly spoken language by millions of people as well as it is the official language for official documents in Odisha. Odia language is rich in literary history. Odia script in stone engravings, palm-leaf manuscripts shows its antiquity. Modern Odia script, like Devanagari script is a descendant of Brahmi script. But unlike Devanagari the characters have got a circular look, possibly under influence of Dravidian writing system and to avoid horizontal lines to be drawn on palm-leaves used as writ ing material in the earlier t imes.

There are 12 vowels, 37 simple consonants, 10 numerical digits and near about 116 composite characters in Oriya alphabets. One of the major characteristics of Oriya elementary characters is that most their upper one third is circular and a subset of them have a vertical straight line at their rightmost part.

II. Discrete Wavelet Transform

Wavelets have many advantages over other mathematical transforms such as the DFT or DCT. Functions with discontinuities and functions with sharp spikes usually take substantially fewer wavelet basis functions than sine-cosine functions to achieve a comparable approximation. Wavelets ability to provide spatial and frequency representations of the image simultaneously motivates its use for feature extraction The Haar wavelet transform is a widely used technique that has an established name as a simple and powerful technique for the multi-resolution decomposition of time series. An original image of size N x N is first of all pass through a filter horizontally and vertically as shown in figure 1.

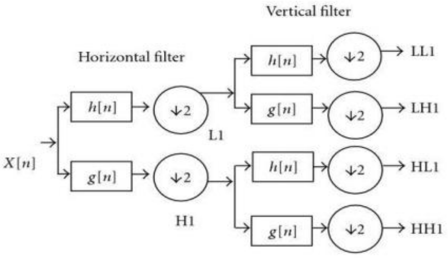


Fig. 1 wavelet Transform

The low pass filtering in horizontal direction and high pass filtering in vertical direction gives rise to LH component, likewise filtering gives rise to four components LL, LH, HL and HH during first level of decomposition [3, 17]. The LL component which represents the approximate coefficients of the decomposition is used to produce next level of decomposition.

The sub band HL represents major facial expression features. The sub band LH, theoretical features of outline and nose are clearer than its horizontal features, depicts face pose features. [18] The sub band HH is the unstable band in all sub bands because it is easily disturbed by noises, expressions and poses. And the sub band LL will be the most stable sub band. Here an image and its detail and approximate coefficients are shown in figure 2.

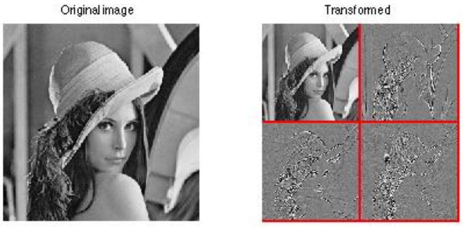
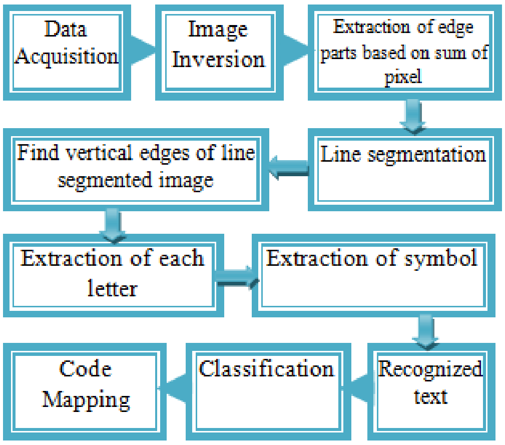


Fig.2 Single level DWT output

III. Proposed Method

In the proposed method we are aiming to extract the odia character from the scan document. After recognition the labeling has been done to the extracted character. The overall process we have divided into the three fold. The first approach is based on segmentation where the scan image has to be undergone for a segmentation process to extract the character from the document. In the second phase the extracted character has to be classify using a set of prototype character set which based on template matching


Data Acquisition
Image Inversion
Extraction of edge parts based on sum of pixel
Line segmentation
Find vertical edges of line segmented image
Extraction of each letter
Extraction of symbol
Recognized text
Classification
Code Mapping

[**Fig 3** Flow Diagram Proposed method]

1.1 Segmentation and Character Extraction

The function of segmentation is to segment the image line from paragraph and individual character from line respectively. Then the process leads to line segmentation and character segmentation. In line segmentation the Inverted image has to be taken as input image and the sum of pixel has to be calculated as row wise. The sum pixel has to calculated as

$$SR(i) = \sum_{j=1}^{n} imn(i, j)$$

Where SR is the sum of row pixel and *imn* is the inverted image.

After extracting the line from the scanned image, each line has been extracted. Then it undergone for sum of column operation using following equation

$$SC(j) = \sum_{i=1}^{m} Ln(i, j)$$

Where SC is the sum of column pixel and *Ln* is the line of text after line segmentation. Again following the same procedure which has been applied to row, now it has to apply for the column pixel.

1.2 Recognition and Classification

For recognition purpose, first we have to create a database by manually cropping the text sets. The database images are of dimension 45 x 45. Then the database set has to be made as template to utilize for classification.

For that purpose the database has to be represented in matrix form as shown below

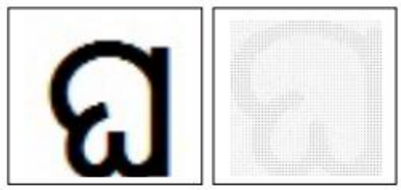


Fig 4 Character in the form of Matrix

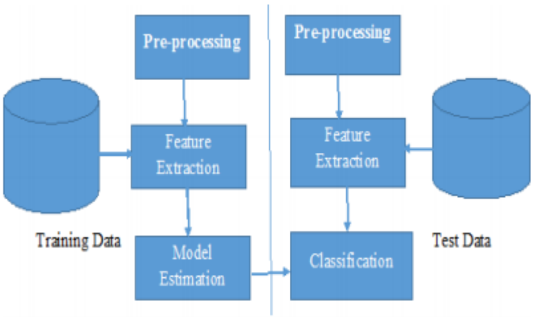


Fig 5 Classification Process

Then for classification we are finding the DWT of each set and store the features. So the database feature has to be extracted and classification done by extracting the text and their feature as shown below.

1.3 Mapping

After classification the mapping has been done to make the text in normal English pronunciation. For that initially we have to assign the some English pronunciation to every letter as well as the symbol used in odiya script.

At the end we display the recognize text with its pronunciation in English language. And then the accuracy rate is calculated based on (No. of Character Found correctly/total number of patterns) x 100.

IV. Result Analysis

In the proposed method we are recognizing the odia characters, for future preservation of odia ancient valuable document. The proposed methodology followed through five phase. Start with skew detection and correction of capture text. Then it passes through a process of segmentation to segment the line and character. In third phase the character and the symbol are extracted. In the preceding phase the extracted text are mapped through template matching process. And finally the recognized characters are labeled with English language. So the overall process is based on template matching process. Due to that it needs to create a database of odia character and symbols.

The whole processing is implemented on MATLAB 7.8 (R2009a) with system configuration Intel I3 with 2.93 GHz clock frequency. The database is created with each letter of dimension 45x45 in .JPG format. The sample of database has been shown below figure which includes 47 odia letter 8 symbols.

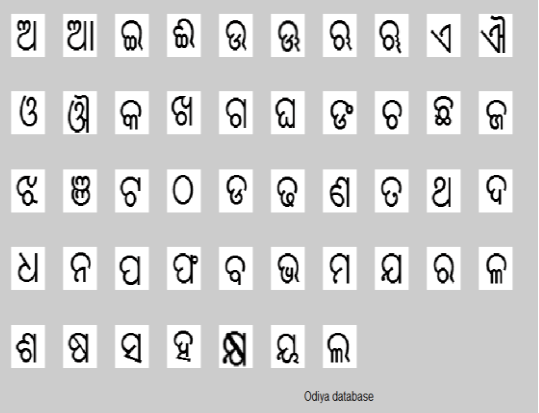


Fig. 6 Odia character database

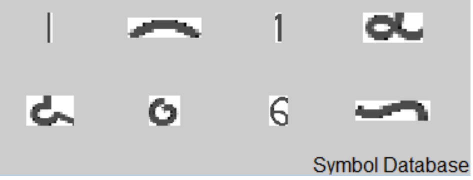


Fig .7 Symbol database

The input image has be scan from text book and taken as input to the proposed system. In the first approach we are correcting the skew error. During capturing of image, due to inappropriate setting of object or camera skew may occur, that has been removed and the processed result has been shown below figure.

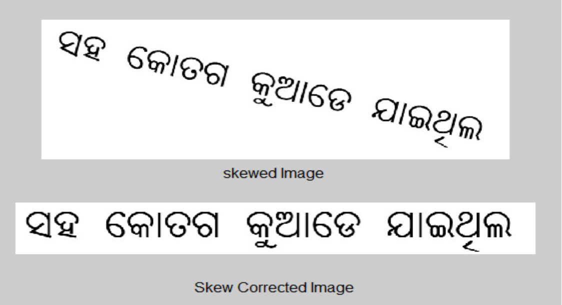


Fig. 8 Skew corrected image

After skew correction the input image is taken and inverts to make the text color while and background color as black, so it can help in further processing of text segmentation.

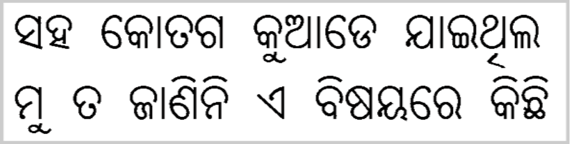

ସହ କୋତଗ କୁଆଡେ ଯାଇଥିଲ
ମୁ ତ ଜାଣିନି ଏ ବିଷୟରେ କିଛି

Fig.9 Inverted image

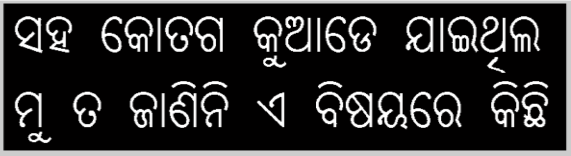

ସହ କୋତଗ କୁଆଡେ ଯାଇଥିଲ
ମୁ ତ ଜାଣିନି ଏ ବିଷୟରେ କିଛି

Fig.10 Line segmented image

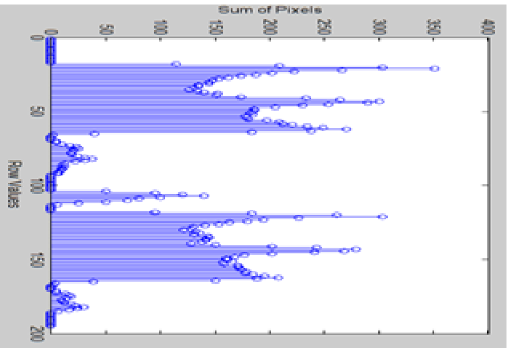


Fig. 11 Sum of row pixel of image

In the next process we are moving towards segmentation, where we are performing line segmentation and character segmentation. After inverting the original text image we have calculated the sum of pixel of pixel values row wise and display it with stem command. It shows that where ever the texts are present there exist a sum value but where there is no text the sum value is nil. From that result we have found out the start line (row) of text. The sums of row pixel are shown below.

In line segmentation we found the text which are segmented line wise and the here it has been shown the segmented text line, before the segment line are shown. The line segmentation result is shown below.

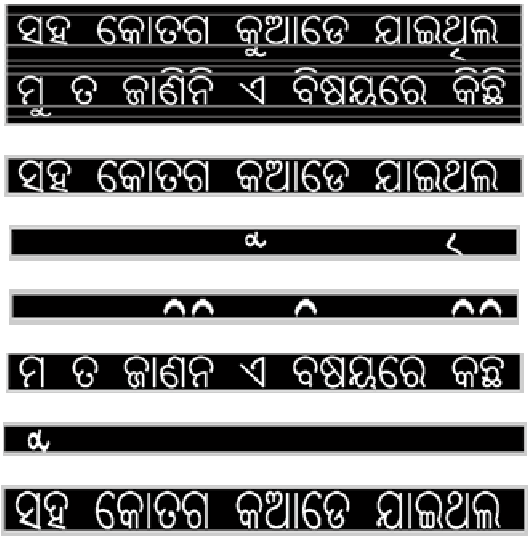


Fig. 12 Step by step line segmented image

In the next process the segmented text line are selected for character segmentation. During this processing we have implemented the same process as before, but instead sum of row pixel we have taken sum of column pixel. The sum of pixel shown below from where, the segmented characters are extracted.

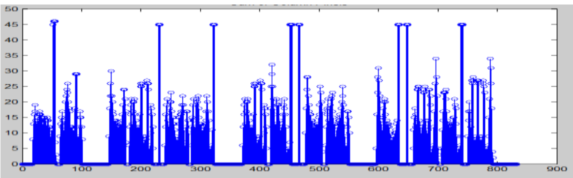
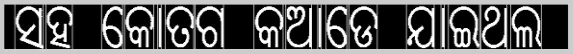


Fig. 13 Sum of column pixel after line segmentation



[**Fig. 14** character segmented image]

The below figure show the extracted text from the scan text image

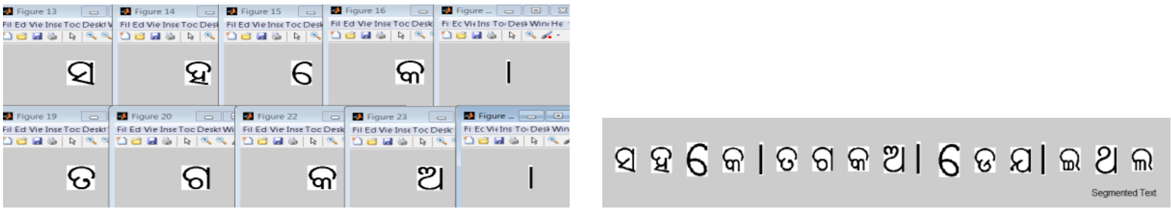
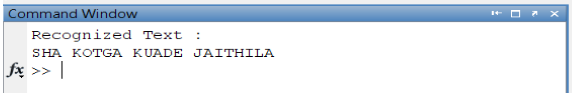


Fig. 15 Segmented text

The extracted text recognized by extracting the DWT features of the stored characters as well as the extracted character. Then the features are compared based on Euclidean distance method. The distance between feature pattern of each prototype is computed and where ever it find best matches that class will be assigned to that. After classification the character assigned with English text label, according to that it automatically assigned the English text for each character. After that the English translated odiya text character are resulted.



```
Recognized Text :
SHA KOTGA KUADE JAITHILA
fx >> |
```

Fig. 16 The mapped Result

In the proposed OCR system, the recognition rate depends on quality of input capture text image. Here we have performed the proposed method to different text set and obtained around 90% accuracy rate. The % of accuracy is calculated as= (No. of Character Found correctly/total number of patterns)x100.

V. Conclusion

In the proposed method we have implement a simple and noble technique that extract the odiya character efficiently from scan document, and represented in English language for reading purpose of non odiya people. Here a simple approach has been developed which first correct the skew error then gone through a segmentation process which include line as well as character segmentation. In third phase each character are extracted and classified through a stored prototype by DWT feature extraction method. The method showing a good recognition rate of 90% towards odiya character recognition process. By doing this we can preserve the old ancient valuable document, script in future.

The work is limited to printed character, and certainly not applicable for handwritten character. So in future we have planned to design robust methods which can successfully recognize optical as well as handwritten characters.

References

[1] Kim, K., et al., “Automatic Cell Classification in Human’s Peripheral Blood Images Based on Morphological Image Processing, in Advances in Artificial IntelligenceSpringer Berlin Heidelberg. p. 225-236.

[2] Venkatalakshmi, B. and K. Thilagavathi. Automatic red blood cell counting using Hough transform., 2013 IEEE Conference on Information & Communication Technologies (ICT). 2013.

[3] Yazan Met.al ,” Automatic Detection and Quantification of WBCs and RBCs Using Iterative Structured Circle Detection Algorithm” Computational and Mathematical Methods in Medicine, Hindawi Publishing Corporation.

[4] Habibzadeh, M., A. Krzyżak, and T. Fevens, “Comparative study of shape, intensity and texture features and support vector machine for white blood cell classification”. Journal of Theoretical and Applied Computer Science, 2013. 7(1): p. 20-35.

[5] E. A. Mohammed, M. M. Mohamed, B. H. Far, and C. Naugler, "Peripheral blood smear image analysis: A comprehensive review," Journal of pathology informatics, vol. 5, 2014.

[6] M. Hamghalam and A. Ayatollahi, "Automatic counting of leukocytes in giemsa-stained images of peripheral blood smear," in Digital Image Processing, 2009 International Conference on, 2009, pp. 13-16.

[7] J. Angulo and G. Flandrin, "Automated detection of working area of peripheral blood smears using mathematical morphology," Analytical cellular pathology, vol. 25, pp. 37-49, 2003.

[8] J. M. Sharif, M. Miswan, M. Ngadi, M. S. H. Salam, and M. Mahadi bin Abdul Jamil, "Red blood cell segmentation using masking and watershed algorithm: A preliminary study," in Biomedical Engineering (ICoBE), 2012 International Conference on, 2012, pp. 258-262. S. S. Adagale and S. S. Pawar, "Image segmentation using PCNN and template matching for blood cell counting," in Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference on, 2013, pp. 1-5.

[9] C. Di Rubeto, A. Dempster, S. Khan, and B. Jarra, "Segmentation of blood images using morphological operators," in Pattern Recognition, 2000. Proceedings. 15th International Conference on, 2000, pp. 397-400 vol.3

[10] Neural Network: Areview from stastical perspective , Bing Cheng, DM Titterington, Stastical science, vol9(1) page2-54, 1994.

[11] Acta Chimica Slovenica, “Introduction to Artificial Neural Network (ANN) Methods”: 41/3/1994, pp. 327-352

[12] J.S.R. Jang, C.T. Sun and E. Mizutani," Neuro-fuzzy and soft computing: a computational approach to learning and machine intellicence", PEARSON Education, Low Price Edition.1997

[13] N. Otsu, "A Threshold Selection Method from Grat-Level Histograms", IEEE Transaction on System, Man and Cybernetics, vol. 9(1), pp. 62-66, 1979.

[14] R.C. Gonzalez and R.E. Woods,"Digital Image Processing", 2nd Edition, Pearson Education 2007