International Journal of Electronics Communication and Computer Engineering
Volume 3, Issue 3, ISSN 2249 –071X

# Segmentation of Devnagari Characters from Noisy Images for creation of Dataset

**Sushilkumar N. Holambe**
Department of Information Technology
College of Engineering
Osmanabad-413512 (M.S.) India
snholambe@yahoo.com

**Pravin P Kalyankar**
Department of Computer Science & Engineering
College of Engineering
Osmanabad-413512 (M.S.) India
kalyankarpp1@yahoo.com

**Dr. Ulhas B. Shinde**
Principal
SPW Engineering College
Aurangabad (M.S) India
drshindeulhas@gmail.com

***Abstract* — In this paper we propose segmentation method for noisy document. The segmentation method is proposed for Devnagari document by considering the structure of Devnagari script .We have also consider the irregularities of Devnagari writing style, Yuktaksha and numbers. We are not removing shirorekha. It separates the image text documents into lines, words and characters. We are getting 100% accuracy at line and word level but at character level it depend on the type of character, whether it is single, Yuktakshar or triakshar.**

***Key Words* — Segmentation, Thresh holding, Feature Extraction, Devanagari.**

## I. INTRODUCTION

Document image understanding and analysis means that transforms the information of a document in the from of paper into an electronic format i.e. text without manual keyboard entry. It is still challenging and interesting task to design a system which gives high recognition accuracy, without considering quality of the input document and different character font style variation. Optical Character Recognition (OCR) is the process of translating images of handwritten, printed text into a format understood by machines. The worlds information of literature, history, and other information is in hard-copy documents. OCR systems convert this information by converting the text on paper into electronic form. OCR system based on segmentation can be divided as preprocessing of given input, segmentation, feature extraction and Classification. For document analysis we have to use OCR system based on segmentation. We can avoid preprocessing if the document is noise free but its not possible because when we scan the noise appear in the document image. Segmentation is the decomposition of an image into sub images. Segmentation is dependent on local decisions with regards to shape similarity, as well as global decisions with regards to surrounding context. In the late 1960s and 1970s researchers observed that segmentation cause more errors in reading characters, whether hand or machine-printed. In the 1980's researchers give new dimension to OCR to less constrained documents [1]. Some authors have surveyed segmentation[2] [3] [4] [5] [6] [7], or document analysis [8] [9] more details can be found in [10].Many of techniques have been developed for Latin but for Devnagari Veena Bansal and R.M.K. Sinha[11][12], U. Garain, B. B. Chaudhuri [13], had done remarkable work and for document processing S. Kompalli, S. Kayak, S. Setlur and V. Govindaraj [14] had done but still we are lagging. More research in noise free, error free and distortion free segmentation technique is required for high accuracy and recognition for Devanagari document processing and OCR.

## II. THE DEVNAGARI SCRIPT

The writing style of Devanagari script is horizontal, left to right and the characters do not have any uppercase or lowercase distinction. There are about fifty basic characters in scripts. Within a word, the vowel characters often take modified shapes called modifiers. Consonant modifiers are also possible. The basic and compound characters can be attached with modifiers to generate new shapes. Apart from these, the documents printed in this script show large variations in font faces, type styles, and in character sizes. The line called shirorekha in Devanagari and is referred as headline. The neighboring characters of a word very often touch through the headline to form a connected component. There are about a thousand conjunct consonants, most of which combine two or three consonants. There are also some with four-consonant conjuncts and at least one well- known conjunct with five consonants. The details are shown in Figure 1 and some Conjuncts Consonants are shown in Figure 2.

(a) Vowels अ आ इ ई उ ऊ ऋ ए ऐ ओ औ
(b) Modifier Symbols corresponding to the vowels (the modifier symbol has also been attached to the consonant क to indicate its placing
ा ि ी ु ू ृ े ै ो ौ
का कि की कु कू कृ के कै को कौ
(c) Consonants क ख ग घ ङ च छ ज झ ञ
ट ठ ड ढ ण त थ द ध न प फ ब भ म य
र ल व श ष स ह
(d) Pure Consonants क् ख् ग् घ् ङ् च् छ् ज् झ् ञ्
ट् ठ् ड् ढ् ण् त् थ् द् ध् न् प् फ् ब् भ् म् य् र् ल् व् श् ष् स्
(e) Some Conjuncts formed by Pure Consonants modifiers when combined with character य
क्य ख्य घ्य च्य ज्य त्य थ्य ध्य न्य प्य भ्य म्य य्य ल्य व्य

Fig.1. Devnagari Vowels, Consonants, Modifier, Conjuncts & Pure Consonants. [15]

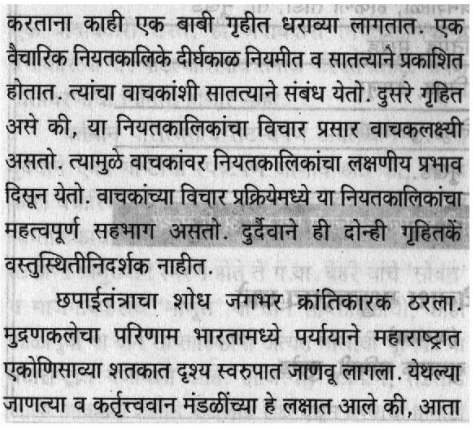

Fig.2. Conjuncts Consonants. [16]

The challenges in segmenting Devanagari documents are

1. Matras (modifiers)in the documents.
2. The combined letters (conjuncts) are called Yuktakshar (joDda_AkShar).
3. Improper different between two consecutive lines, because if the first line has ukar or lower modifier below the character then second line has top modifier above the character like eekar etc. then the distance between line is so minimum that it cannot be segmented correctly.
4. There are vowel modifiers, namely, “Anuswar”, “Visarga” and “Chandra Bindu”, which add up to the confusion.
5. There are infinite variations of handwriting of individuals.

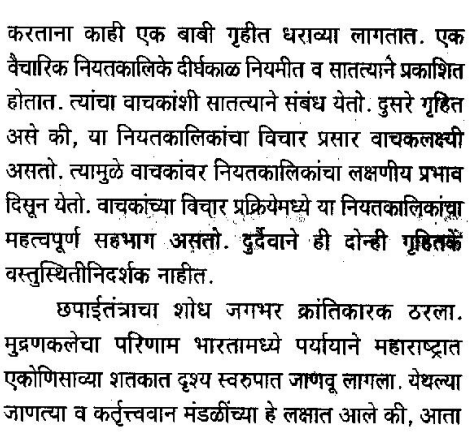
## III. PREPROCESSING

The Preprocessing phase includes the conversion of gray scale image into binary, noise removal, thinning and skew detection and correction. The scanned documents may contain noise. Digital capture of images can introduce noise from scanning devices and transmission media. Here we are minimizing scanning noise by eliminating the grayish background and then we do the inverse operation.


करताना काही एक बाबी गृहीत धराव्या लागतात. एक वैचारिक नियतकालिके दीर्घकाळ नियमीत व सातत्याने प्रकाशित होतात. त्यांचा वाचकांशी सातत्याने संबंध येतो. दुसरे गृहित असे की, या नियतकालिकांचा विचार प्रसार वाचकलक्ष्यी असतो. त्यामुळे वाचकांवर नियतकालिकांचा लक्षणीय प्रभाव दिसून येतो. वाचकांच्या विचार प्रक्रियेमध्ये या नियतकालिकांचा महत्वपूर्ण सहभाग असतो. दुर्दैवाने ही दोन्ही गृहितकें वस्तुस्थितीनिदर्शक नाहीत.

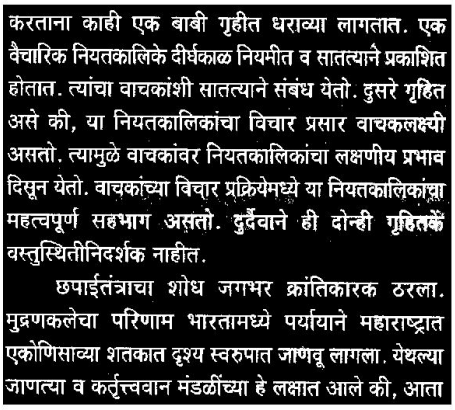
छपाईतंत्राचा शोध जगभर क्रांतिकारक ठरला. मुद्रणकलेचा परिणाम भारतामध्ये पर्यायाने महाराष्ट्रात एकोणिसाव्या शतकात दृश्य स्वरुपात जाणवू लागला. येथल्या जाणत्या व कर्तृत्त्ववान मंडळींच्या हे लक्षात आले की, आता

Fig.3. Original Document Image


करताना काही एक बाबी गृहीत धराव्या लागतात. एक वैचारिक नियतकालिके दीर्घकाळ नियमीत व सातत्याने प्रकाशित होतात. त्यांचा वाचकांशी सातत्याने संबंध येतो. दुसरे गृहित असे की, या नियतकालिकांचा विचार प्रसार वाचकलक्ष्यी असतो. त्यामुळे वाचकांवर नियतकालिकांचा लक्षणीय प्रभाव दिसून येतो. वाचकांच्या विचार प्रक्रियेमध्ये या नियतकालिकांचा महत्वपूर्ण सहभाग असतो. दुर्दैवाने ही दोन्ही गृहितकें वस्तुस्थितीनिदर्शक नाहीत.

छपाईतंत्राचा शोध जगभर क्रांतिकारक ठरला. मुद्रणकलेचा परिणाम भारतामध्ये पर्यायाने महाराष्ट्रात एकोणिसाव्या शतकात दृश्य स्वरुपात जाणवू लागला. येथल्या जाणत्या व कर्तृत्त्ववान मंडळींच्या हे लक्षात आले की, आता

Fig.4. Grayish Background Eliminated Document Image


करताना काही एक बाबी गृहीत धराव्या लागतात. एक वैचारिक नियतकालिके दीर्घकाळ नियमीत व सातत्याने प्रकाशित होतात. त्यांचा वाचकांशी सातत्याने संबंध येतो. दुसरे गृहित असे की, या नियतकालिकांचा विचार प्रसार वाचकलक्ष्यी असतो. त्यामुळे वाचकांवर नियतकालिकांचा लक्षणीय प्रभाव दिसून येतो. वाचकांच्या विचार प्रक्रियेमध्ये या नियतकालिकांचा महत्वपूर्ण सहभाग असतो. दुर्दैवाने ही दोन्ही गृहितकें वस्तुस्थितीनिदर्शक नाहीत.

छपाईतंत्राचा शोध जगभर क्रांतिकारक ठरला. मुद्रणकलेचा परिणाम भारतामध्ये पर्यायाने महाराष्ट्रात एकोणिसाव्या शतकात दृश्य स्वरुपात जाणवू लागला. येथल्या जाणत्या व कर्तृत्त्ववान मंडळींच्या हे लक्षात आले की, आता

Fig.5. Inverse (Final preprocessed) Document containing Noise

The above Figure 3 is the original document containing scanning, salt and pepper noise. We are not removing the noise but we are minimizing the same. Figure 4 shows Grayish Background Eliminated Document Image. Figure 5 is the document image which is used for segmentation which contains noise.

## IV. SEGMENTATION

In Devanagari script, the consonant and vowel modifiers may be attached / placed on top or bottom or left or right to the base character as shown in Figure 1. and Yuktakshar shown in Figure 2.

### *4.1 Proposed segmentation Algorithm*

1. Rotate the image by 270.
2. Take the row wise summation and find minima’s in this graph.
3. Separate each minima’s. You will get line separated.
4. Rotate line by 90 degree.
5. You we get the line
6. Perform step 1 to 5 for each line independently and store each line in a directory .
7. Take Each line from directory
8. Take Column wise sum to obtain the minima’s .
9. With each minima’s representing a start of a

character and the next represent the end of character.

**10.** Characters are separated and stored in directory.

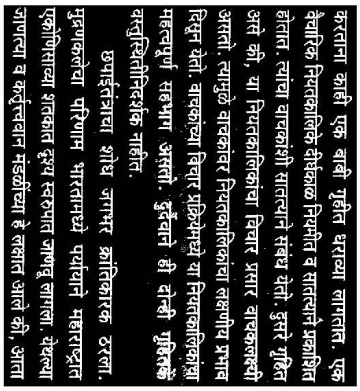


Fig.6. 270 Degree Rotated Document containing Noise

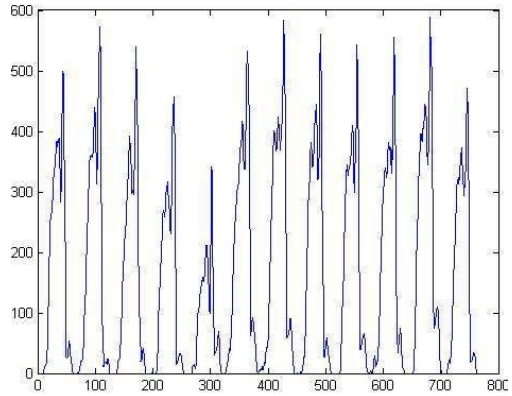



Fig.7.Graph showing minima's

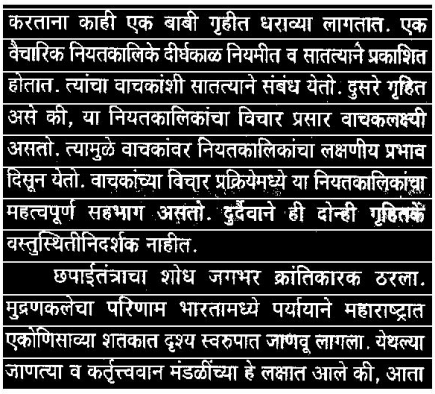



Fig.8. Document segmented into line


Fig.9. Each Separated line is segmented into characters

## V. CREATION OF DATASET

Here we have combined numbers, vowels, consonant together in the dataset. After segmentation we have separated the characters and formed the dataset. We have created dataset by using ISM software fonts for Devanagari (Marathi) Script. The font used for this purpose are DVB-TTSurekh, DVB-TTBhima, DVB-TT Chhaya, DVB-TTDhruv, DVB-TTDhruv, DVB-TTGanesh, DVB-TTRadhika, DVB-TRaghav. DVB-TTShridhar, DVB-TTYogesh. The font sizes are 16, 18, 20, 22. Which we will use in future research.

## VI. RESULTS AND DISCUSSIONS

The algorithm is implemented in MATLAB. The data set is created by us from marathi literature books. Which is scanned at 300 dpi. The algorithm is tested with several noisy document images. Some of these documents contained Numbers and Yuktakshar. Sample test results are shown in above Figures. From the experimentation it is found that the proposed method is reliable to segment noisy text documents that contain numbers, symbols and Yuktakshar. The line segmentation accuracy of 100% for good quality and for poor quality documents and the character segmentation (including characters, numbers, symbols and Yuktakshar) accuracy of 99% for good quality, 98% for poor quality documents achieved.

## REFERENCES

[1] J. Schuermann, A Reading machines, Proc. 6th Int. Conf. on Pattern Recognition, Munich, 1982.

[2] L.D. Harmon, Automatic Recognition of Print and Script, Proceedings of the IEEE, vol. 60, no. 10, pp. 1165-1177, Oct. 72.

[3] G. Dimauro, S. Impedovo and G. Pirlo, From Character to Cursive Script Recognition: Future

[4] Trends in Scientific Research, Proc. 11th Int. Conf. on Pattern Recognition, vol. II, page 516, Aug. 1992.

[5] C.E. Dunn and P.S.P. Wang, Character Segmenting Techniques for Handwritten Text - A Survey, Proc. 11th Int. Conf. on Pattern Recognition, vol. II, page 577, August 1992.

[6] E. Lecolinet and O. Baret, Cursive Word Recognition: Methods and Strategies, Fundamentals in Handwriting Recognition, S. Impedovo (Ed.), NATO ASI Series F: Computer and Systems Sciences, vol. 124, Springer Verlag, 1994, pages 235-263.

[7] G. Lorette and Y. Lecourtier, Is Recognition and Interpretation of Handwritten Text: a Scene Analysis Problem? Pre-Proceedings IWFHR III, Buffalo, page 184, May 1993.

[8] C.C. Tappert, C.Y. Suen and T. Wakahara, The State of the Art in On-line Handwriting Recognition, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 12, no. 8, page 787, Aug. 1990.

[9] D.G. Elliman and I.T. Lancaster, A Review of Segmentation and Contextual Analysis Techniques for Text Recognition, Pattern Recognition, vol. 23, no. 3/4, pp. 337-346, 1990.

[10] H. Fujisawa, Y. Nakano and K. Kurino, Segmentation methods for character recognition: from segmentation to document structure analysis, Proceedings of the IEEE, vol. 80, no. 7 pp. 1079- 1092, July 1992.

[11] R. G. Casey and E. Lecolinet, "A survey of methods and strategies in character segmentation, " IEEE Trans. Pattern Anal. Machine Intell., vol. 18, July 1996.

[12] Veena Bansal and R.M.K. Sinha, Partitioning and Searching Dictionary for Correction of Optically-Read Devanagari Character Strings, in Proceedings - Fifth International Conference on Document Analysis and Recognition, IEEE

Publication, held at Bangalore from Sep21-23, 1999, pp. 410-413.
[13] V. Bansal, R.M.K. Sinha, "Segmentation of Touching and Fused Devanagari Characters", Pattern Recognition, vol. 35, pp. 875-893, April 2002.
[14] U. Garain, B. B. Chaudhuri, "Segmentation of Touching Characters in Printed Devnagari and Bangla Scripts using Fuzzy Multifactorial Analysis", Proc. 6th ICDAR, pp. 805-809, 10-13 Sept. 2001.
[15] S. Kompalli, S. Kayak, S. Setlur and V. Govindaraj, Challenges in OCR of Devnagari Documents. Proceedings of Eighth International Conference on Document Analysis and Recognition (ICDAR'05), pp. 327- 331, Seoul, South Korea, September, 2005.
[16] V. Bansal and R. M. K. Sinha, "Integrating Knowledge Sources in Devnagri Text Recognition, " IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems

## AUTHOR'S PROFILE

**Sushilkumar N. Holambe**
Department of Information Technology College of Engineering
Osmanabad-413512 (M.S.) India
snholambe@yahoo.com

**Pravin P Kalyankar**
Department of Computer Science & Engineering
College of Engineering
Osmanabad-413512 (M.S.) India
kalyankarpp1@yahoo.com

**Dr. Ulhas B. Shinde**
Principal,
SPW Engineering College
Aurangabad (M.S) India
drshindeulhas@gmail.com