Segmentation based Optical Character Recognition for Handwritten Marathi characters

Prof. M V Vaidya[1],Dr. Y V Joshi[2],Prof. M V Bhalerao[2]

Head, Department of Information Technology[1]

Department of Electronics and Telecommunication[2]

Shri Guru GobindSinghji Institute of Engg. &Tech.Vishnupuri, Nanded-431606 (M. S.), India[3]

ABSTRACT

Valuable ancient documents like historical books, old scripts etc. are available in specific regional languages. Problems occurs when those documents have to be preserve in digital form or to modify them. Optical Character Recognition is used to convert the scanned document word, notepad or any other format, so that we can easily edit that document. A complete OCR system for handwritten Devanagari document has been studied and we explore our work on segmentation technique. Literature survey on feature extraction and classification process has been completed. Finally directions for future research are suggested.

Index Terms—Devanagari Script, Preprocessing, Segmentation, feature extraction, classification.

I. INTRODUCTION

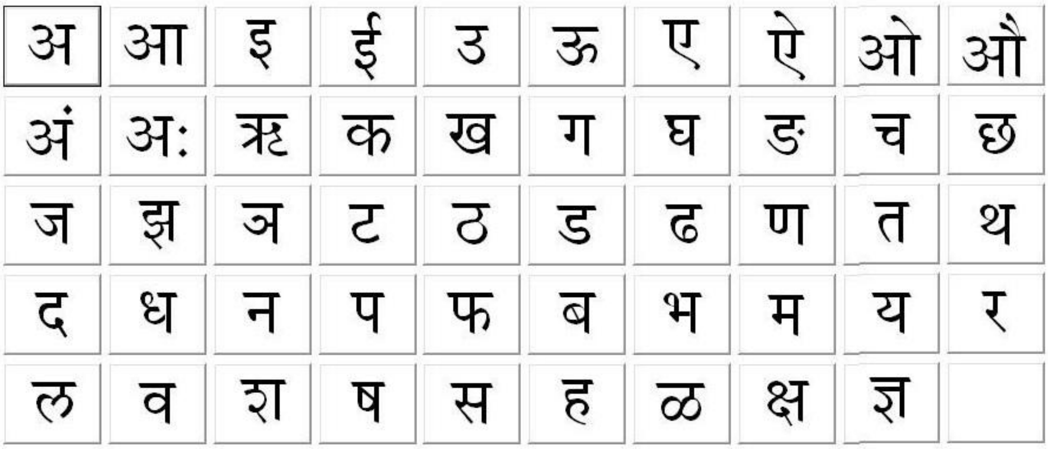
Devanagari Script is most popular in India. It has been used by more than 300 million people. Various languages use Devanagari script. Mainly Hindi, Marathi, Pali, Sanskrit, Punjabi, Bhojapuri etc. comes under the category of Devanagari script as shown in Table 1.

Sr No.	Language	Population (2001censes)	Region
1.	Marathi	84,184,806	Maharashtra, Goa,
2.	Maithili	12,179,122	Northern, Bihar, Madhesh
3.	Marwari	7,936,183	Rajasthan, Haryana, Gujarat,
4.	Konkani	2,489,015	Maharashtra
5.	Kashmiri	5,527,698	Kashmir
6.	Hindi	422,048,642	All over India
7.	Bodo	1,350,478	Assam
8.	Bihari	25,48,794	Bihar
9.	Bhojpuri	40,000,000	Bihar, UP Jharkhand etc.
10.	Awadhi	2,529,308	Ancient India
11.	Nepali	2,871,749	Nepal
12.	Newari	14,568	Ancient India
13.	Nihali	12,548	Ancient India
14.	Pali	245864	Bihar, Ancient India
15.	Punjabi	29,102,477	Punjab
16.	Rajsthani	15,65,843	Rajsthan
17.	Sanskrit	4,991,289	Vedic Language

Table 1: LANGUAGES IN INDIA

Devanagari has been used as base for many languages in India and Abroad. Origin of this script is 'Brahmi' which is purely Indian in nature. Constitutionally 22 languages are

recognized in India. Devanagari Script is different from Roman script, it is phonetic script. This script occurs in the composition of vowels and modifiers as it has 13 vowels and 36 consonants i.e. overall 49 basic characters. There exists a horizontal line at the upper part called 'Shirorekha' or header line. It divides the word into three major zone: lower strip, middle strip and upper strip. The upper strip denotes the portion above the headline, middle zone covers the portion of composition of vowels and components, lower modifiers are present in lower strip.



अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ
अं	अः	ऋ	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	ळ	क्ष	ज्ञ	

FIGURE 1: Character set in Marathi Language

Recognition of handwritten Devanagari script is more difficult than printed, because there are variation in writing style of each person. It depends upon the alignment and different forms of handwritten strokes and variations are geometrical. Handwritten characters of Devanagari script are cursive and unconstructed in nature, so segmentation and recognition is critical part. For getting high accuracy of recognition good feature extraction is one of the important factor.

This paper presents the detail study of Devanagari OCR for handwritten document. Section II describes Data Acquisition and preprocessing steps as preprocessing is needed for removing the noise from document. Section III is about the implementation of segmentation technique which we are using. Section IV discusses the Feature Extraction and Classification process from survey. The references include the most relevant papers recently published as well as some old papers, which gives a comprehensive outline of the developments in the field of research.

II. METHODOLOGY

In this approach first the image is scanned and made suitable for further processing by preprocessing. After preprocessing line word and character segmentation is performed. features of segmented characters are extracted for classification purpose.

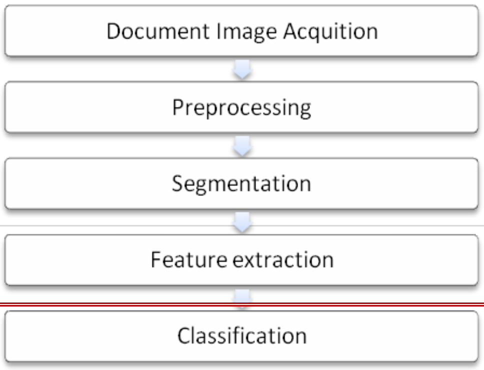
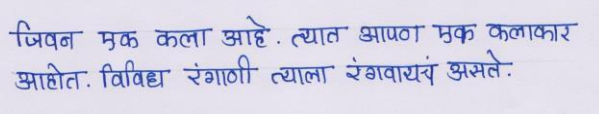


FIGURE 2: Character recognition architecture

III.PREPROCESSING

In preprocessing stage we will the input image from the scanner as shown in figure 2.



जिवन मुक कला आहे. त्यात आपण मुक कलाकार
आहोत. विविद्य रंगाठी त्याला रंगवायचं असते.

FIGURE 2: Sample document Image

There may be some noise and other things may be present in this input image rather than text of Devanagari text. Preprocessing stage is needed to remove the noise and convert the original scanned RGB image into binary form. Noise introduced by optical scanning devices or image capturing instruments cause errors in segmentation, noise reduction eliminates these imperfections as it is necessary for recognition.

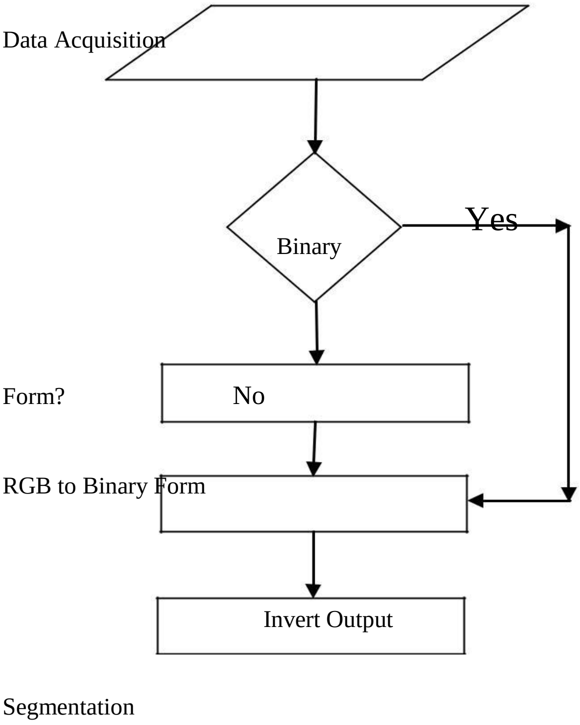
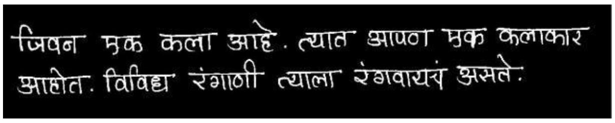


FIGURE 3: PREPROCESSING FLOWCHART

RGB image is converted into binary format, i.e., only two value matrix of image is produced. Binary image is inverted and the pixels of each character in input are represented by value 1 that is in white color as shown in figure 4. Noise reduction, Normalization of data, Compression in the amount of data to be retained are main objectives of preprocessing.



जिवन एक कला आहे. त्यात आपण एक कलाकार
आहोत. विविद्य रंगाणी त्याला रंगवायचं असते.

FIGURE 4: BINARY IMAGE

IV. SEGMENTATION

For recognition of individual characters and convert it into standard format, the separation of paragraphs, text lines, words, and characters is to be carried out for effective feature extraction. Segmentation is one of the hardest, crucial, and time consuming phase. In image processing the segmentation approaches falls broadly in Pixel classification, Edge based Segmentation, and Region based categories. For document segmentation pixel classification and region based classification are used.

Algorithm presented by Bansal and Sinha [1] for the segmentation of machine-printed composite characters into their constituent symbols used a two-pass algorithm. Two-pass algorithm extensively used structural properties of the script. Words are segmented into easily separable or composite.

characters in first pass. For the identification of whether a character box is composite, statistical information about the height and width of each separated box is used. The hypothesized composite characters are further segmented in second pass. Depending on the context the regions of image for further segmentation is selected based on statistical analysis of height or width. Kompalli et al. [9] proposed a method foe determination of the Shirorekha using projection profile and run length. Upper modifiers above the Shirorekha are isolated as ascenders. The top middle and bottom strips are identified easily after removing the Shirorekha. For feature extraction and classification each of these components is then scaled to a standard size. A fuzzy multi-factorial analysis presented by Garain and Chaudhuri [10] for identification and segmentation of touching machine-printed Devanagari characters. This technique developed a predictive algorithm for selecting possible cut columns for segmenting the touching characters effectively in the same.

A. Holistic Approach

Instead of segmenting word into characters, the holistic approach is used to recognize the entire words as single one entity. It is the segmentation-free approach as there is no attempt to identify characters invidiously and recognition is globally performed on whole image of words. This approach eliminates the issue of finding individual characters by considering words as atomic units of recognition.

B. Proposed Technique for Segmentation

We used the isolated characters approaches for segmenting individual characters from whole document, considering the Shirorekha with words. Shirorekha is used to determine the upper strip, middle strip, and lower strip but we are not eliminating the Shirorekha because the database we create also have Shirorekha in each vowels and consonants. The propose method is 100% efficient for the segmentation of line and word. The segmentation of characters from word is critical, as single word may contain composite characters i.e. combination of vowels and consonants.

i. Line segmentation

ii. Word segmentation

iii. Character segmentation

Line segmentation divides the A4 size page document into number of images equal to the lines present in page. That individual image is again segmented which is word segmentation and it gives separated words from single straight lines. Word segmentation is easier than line segmentation and character segmentation. Words are segmented by the projection based method. Each word is saved in new image and further it is segmented in characters.

Algorithm:

Input - Binary complementary image of document (A4 size page).

[1] Rotate image with 270 degree.

[2] Plot the vertical histogram projection.

[3] Find out minima‘s of histogram by taking row-wise summation.

[4] Crop image(s) for consecutive white pixels.

[5] Save the new segmented lines as separate image.

[6] Repeat above steps up to last line.

[7] Take image of separated single line.

[8] Plot the vertical projection of line.

[9] Find out column wise sum, get the minima's.

[10] Each individual word has different values in histogram obtained from it and the histogram has consecutive zero value is the blank space between two words.

[11] Crop image.

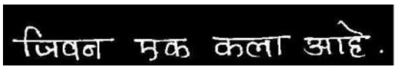
[12] Save the each word as new image.

[13] Repeat above step up-to last pixel in line.

[14] Get the new image of separated word.

[15] Repeat steps from 8 to 14 but here in histogram will have some value at the end of characters because of presence of Shirorekha.

1. Line from Document



2. Segmented Words

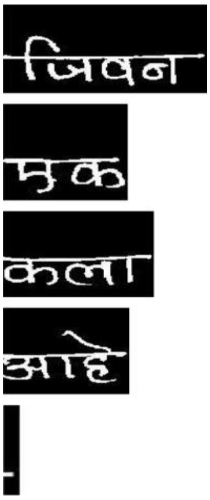
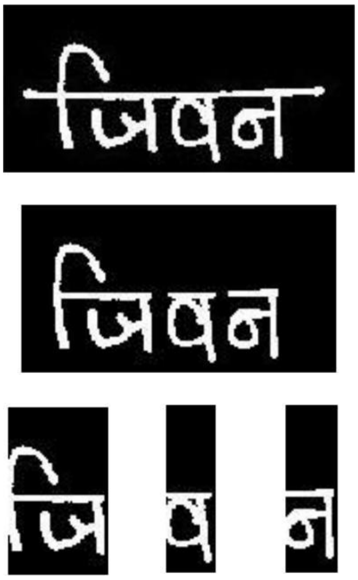


FIGURE 5: RESULT OF LINE SEGMENTATION



1. Image obtained after Preprocessing

2. Vertical Segmentation

3. Segmented Letters

FIGURE 6: RESULT OF WORD SEGMENTATION

V. FEATURE EXTRACTION AND CLASSIFICATION

Feature extraction is challenging task as it proves to be successful in one application domain may turn out not to be very useful in another domain. Feature extraction pick out the different features from single character and it helps to distinguish between various classes for further classification. Extracting features from raw data which are relevant for classification purposes and enhancing that between-class pattern variability is nothing but the feature extraction.

Sandhya Arora et al. [12], proposed a method which used four different techniques for feature extraction. They have taken the three different features by performing scaling of image of character. After performing thinning, generating one pixel wide skeleton of character image and segmenting the image into 16 segments, 32 intersection features are extracted in First type of feature. From eight octants of the character image, 16 shadow features are extracted in second type of feature. Detecting the contour points of original scaled character image and dividing the contour image into 25 segments, 200 chain code histogram features are obtained in third type of feature.

For finding Shirorekha, vertical bar (Spine) in handwritten Devanagari characters cannot be done very efficiently by simple histogram method. A two stage classification method is proposed again by S. Arora et al., in which they got 89.12% success. First stage was for structural properties like Shirorekha, spine in character and second exploits some intersection features of characters which are fed to a feed forward neural network.

Following are the various methods which can be used for extraction; it requires some mathematical computation and calculations [5]:

a) *Fourier Transforms*: Fourier transform is used to identify the position-shifted characters after discovering the magnitude spectrum in feature extraction. Magnitude spectrum of the measurement vector can be chosen as the features in an n-dimensional Euclidean space.

b)Gabor Transform: Gabor Transform vary from Fourier Transform where window is used defined by Gaussian function.

c)Wavelets: It is used get different levels of resolution of image and signal. Separated characters obtained by segmentation are represented by wavelet coefficients corresponding to various levels of resolution. Coefficients getting after extraction are then fed to a classifier for classification [3].

d)Moments: Central moments, Legendre moments, and Zernike moments, make the process of recognizing an object scale, translation, and rotation invariant. original image can be completely reconstructed from the moment coefficients as series expansion representation.

Govinda Raju et al. [9] used the gradient features computed using a Sobel operator measures the magnitude and direction of intensity changes in a small neighborhood of each pixel. A threshold is considered to map gradient feature is computed by retaining those gradient magnitudes. The gradient vectors of the constituent blocks are used to construct the feature vector. Gradient, structural, and concavity (GSC) features are used in method proposed by Kompalli et al. [8], for OCR of machine printed and multi-font Devanagari text. This feature is also used to classify the segment.

After extracting the features, the pattern is needed to be classified according to the predefined classes. Classification used in OCR falls in category of Supervise learning. A set of predefine classes need to store already and classifier attempts to identify the pattern that represent the input feature. It is totally based of the patterns or classes that have already been classified. Table II presents an abstract view of various techniques of feature extraction and classification which have been used by different researches we referred.

TABLE II: RESULTS OF FEATURE EXTRACTION AND CLASSIFICATION TECHNIQUES OBTAINED FROM STUDY.

Sr. No.	Proposed by	Feature Extraction Technique	Classifiaction Technique	Results % Accuracy
1.	Arora et. al [12]	Combined	Multi Layer Perceptrons	92.80
2.	Pal et al.[7]	Gradient & Gaussian Filter	Quadratic Classifiers	91.24
3.	Kompalli et al. [8]	GSC features	Artificial Neural Network	88.12
4.	Vaidya and Joshi [13]	Statistical Features	Decision tree	94.16
5.	Vaidya and Joshi [14]	Pixel level features	Decision tree	94.12

VI. CONCLUSION

Converting handwritten document into editable form is not a single step procedure, as it requires multiple processes like preprocessing, segmentation, feature extraction and classification. Recognition of handwritten document is more difficult task than printed one. Data acquisition can be done by scanner or any image capturing devices. For recognition of whole document, separation of character is necessary which gives the result of segmentation step. Dividing whole word into individual characters the holistic approaches can be used to avoid the segmentation of word into character. We used the histogram projection method for segmentation which gives 99% accuracy in result of segmentation of line and segmentation of word. Various techniques are listed and studied for feature extraction and classification as our future work.

REFERENCES

[1] Bansal V, Sinha R. M. K., "Integrating Knowledge Resources in Devanagari. Text recognition system", IEEE Transaction on System, Man & Cybernetics Part A: Systems & Humans. Vol. 3, No. 4, pp. 500-505, July 2000.

[2] P. S. Deshpande, Latesh Malik, "Fine Classification & Recognition of Hand Written Devnagari Characters with Regular Expressions & Minimum Edit Distance Method", Journal of Computers, Vol. 3, No. 5, pp. 11-17, May 2008.

[3] Nafiz Arica, Fatos T. Yarman-Vural, "An Overview of Character Recognition Focused on Off-Line Handwriting", IEEE Transactions on Systems, Man, and Cybernetics— Part C: Applications and Reviews, Vo. 31, issue 2, May 2001.

[4] Agnihotri, Ved Prakash, "Offline Handwritten Devanagari Script Recognition", International Journal of Information Technology and Computer Science, 2012, Vol. 4, No. 8, pp. 37-42, 2012.

[5] R. Jayadevan, Satish R. Kolhe, Pradeep M. Patil, and Umapada Pal, "Offline Recognition of Devanagari Script: A Survey", IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, Volume 41, Issue 6, November 2011.

[6] Mahesh Jangid, "Devanagari Isolated Character Recognition by using Statistical features", International Journal on Computer Science and Engineering (IJCSE), Vol. 2, No 3, pp. 2400-2407, 2011.

[7] U. Pal and B. B. Chaudhuri, "Indian script character recognition: A survey", Pattern Recognition., Volume 37, pp. 1887–1899, 2004.

[8] S. Kompalli, S. Setlur, and V. Govindaraju,"Devanagari OCR using a recognition driven segmentation framework and stochastic language models." International Journal on Document Analysis and Recognition (IJDAR), Vol. 12, no. 2, pp. 123-138, 2009.

[9] V. Govindaraju, S. Khedekar, S. Kompalli, F. Farooq, Setlur and Vemulapati, "Tools for enabling digital access to multi-lingual Indic documents", In Proc. 1st International Workshop on Document Image Analysis for Libraries, pp. 122-133, 2004.

[10] U. Garain and B. B. Chaudhuri, "Segmentation of touching characters in printed Devnagari and Bangla scripts using fuzzy multifactorial analysis", IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, Vol. 32, Issue 4, pp. 449–459, 2002.

[11] U. Pal, P. P. Roy, N. Tripathy and J. Llados, ―Multi-oriented Bangla and Devnagari text recognition", Pattern

Recognition., Volume 43, Issue 12, pp. 4124–4136, 2010.

[12] Sandhya Arora, Debotosh Bhattacharjee, Mita Nasipuri, " Combining Multiple Feature Extraction Techniquesfor Handwritten Devnagari Character Recognition", 3rd IEEE International conference on Industrial and Information Systems, pp. 1-6, 2008.

[13] M. Vaidya and Y. Joshi, "Marathi Numeral Recognition using statistical distribution features", In: Proc. IEEE conference on Information processing, pp. 586-591, 2015.

[14] M. Vaidya and Y. Joshi, "Handwritten Numeral Identification System Using Pixel Level Distribution Features", In: Proc. 2nd International Conference on Information and Communication Technology for Intelligent Systems, pp.1-9, 2017.