ISSN: 2395-7639

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

(A Monthly, Peer Reviewed Online Journal)

Visit: www.ijmrsetm.com

Volume 3, Issue 7, July 2016

Optical Character Recognition System For Devanagari Script

Asmita Kunkari

Dept. of I.T, SGGS I E&T, Nanded, India

ABSTRACT: Today, Everything is done with the help of computers. So, Restoration of documentation such as digital form, historical books, handwritten materials, letters, holistic books are important. Optical character recognition is a method of taking input as images or photos or typewritten scripts and convert them into information that machine can understand. The Government applications are filled with Offline applications and directly entered into a structured database, if handwriting recognition system is perfectly standardized and can be ubiquitously available in all computer devices. In this paper we focused on pre-processing, segmentation and feature extraction and literature survey of classification has been studied.

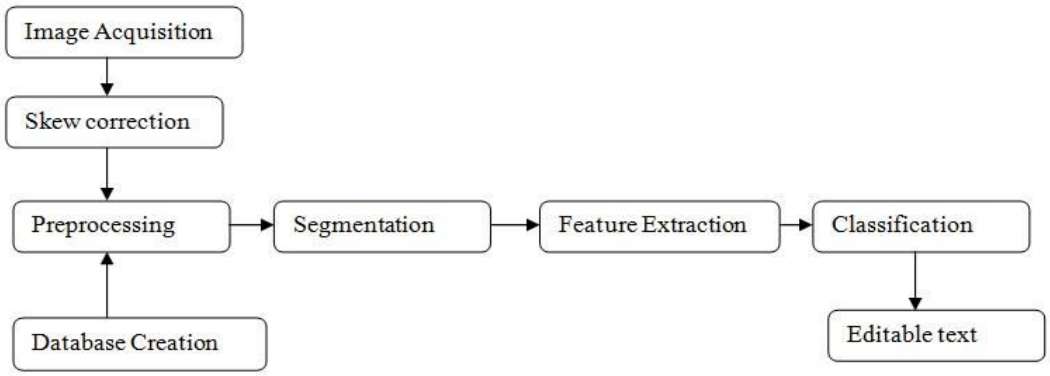
KEYWORDS: Optical character recognition, Devanagari script, Pre-processing, Segmentation, Feature extraction, Gabor filter, Classification.

I. INTRODUCTION

The making of handwriting character recognition (HCR) system is a most interesting area in pattern recognition. Our system is most useful because it uses the marathi language document for recognition. As days passes the paper quality documents like Historical books, handwritten material, letters, holistic materials decreases so restoring documentation in digital form is important. The digital memory is very cheap to compare with real time storage spaces. OCR is process of scanning text data and converting it into margin editable text form and can be save as text files in word or notepad. There are lots of application of OCR such as text detection of address on envelops and signatures on bank cheques, archiving of data in companies and ancient script retrieval, aid to blinds, creating digital libraries More research has been done on roman scripts like English while it still remains less research on devanagari script.

Today, there is big demand for OCR related research for Indian languages, for that there are numberless technical problems as well as insufficient commercial market. The use of computers in organizations as well as in homes drastically increase and automatic processing of paper documents is rapidly gaining denotation in India [2].Several other Indian languages Panjabi, bengali, Gujarati having same features like devanagari script.

Optical character recognition system is shown below. The necessary steps for character recognition are segmentation and classification. Segmented output is recognition unit in OCR [1].



Our paper is structured as follows, section II is about to discuss the literature survey. Next section III is describes the

ISSN: 2395-7639

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

(A Monthly, Peer Reviewed Online Journal)

Visit: www.ijmrsetm.com

Volume 3, Issue 7, July 2016

pre-processing i.e image binarization, noise removal of image and skew correction of document if present. In section IV segmentation algorithm is discussed. Section V details of Feature extraction of Devanagari characters. In next section VI discuss the survey of different classification techniques for character recognition. Conclusion is given in section VII.

II. LITERATURE SURVEY

In [1] the knowledge sources we use are mostly statistical in nature or in the form of a word dictionary tailored specifically for optical character recognition (OCR). We do not perform any reasoning on these. However, we explore their relative importance and role in the hierarchy. Some of the knowledge sources are acquired a priori by an automated training process while others are extracted from the text as it is processed. A complete Devanagari OCR system has been designed and tested with real-life printed documents of varying size and font. In [2] the process of segmentation process introduced. Devanagari is mostly useful Script in India for number of officials and banking applications. Gabor filters [4] have been widely used in Pattern Recognition, Computer Vision and Document Analysis applications. With their insemination from the classic paper1, they have been successfully applied in different applications of Image Processing like texture analysis, number plate recognition, object tracking and biometric systems of iris, face, palm and finger-print recognition. The proposed work present in [6] recognition of handwritten Devnagari characters free from normalization there by giving flexibility and allowing size variation. The project of converting Indian language document to Bharti Braille script has many challenges. The illumination invariant character recognition is one of such challenge which is addressed in this paper. The Gabor features provide illumination invariance up to certain extend, but in recent developments such as local binary pattern and binarizing the directional filter's response and then computing features from them have made feature highly tolerant to lighting changes. In [7] paper proposes an approach for feature extraction wing a CNN Gabor filter and an orientation map. [8] Presents the experiment is to compare the performances of different feature extraction methods including the proposed method in different classifiers. Mamatha Hosalli Ramappa and Srikantamurthy Krishnamurthy [9] the recognition results of different classifiers and feature extraction methods and to provide a new benchmark for future research a comparative study of handwritten Kannada numeral recognition with eight different type of features and nine different classifiers has been reported. N. Arica and F. T. Y. Vural [10] serves as a guide and update for readers working in the CR area. First, the historical evolution of CR systems is presented. Then, the available CR techniques, with their superiorities and weaknesses, are reviewed. The [11] paper describes a correction method for optically read Devanagari character strings which uses a partitioned word dictionary. The word dictionary is partitioned in order to reduce the search space besides preventing forced match to incorrect word.

III. PROPOSED METHOD

1. Preprocessing

Preprocessing is process of increasing the image quality which is required to allows the steps followed by it to deliver accurate results. Scanned image or document is a input to pre-processing steps , during scanning these may be noise present in input image so, it shows be remove during preprocessing step . The input image is Gray image and in preprocessing, the image gets transforms into binary format is called as binarization that is only 0 and 1's format. For binarization, we have used thresholding and then converting the original image to binary image.

2. Segmentation

In OCR segmentation plays very important role, so correct segmentation brings about better recognition. Segmentation is almost difficult task in OCR for Devanagari script and it gets to be more difficult with handwritten scripts due to varieties of writing style of different people. Veena basal [1] introduces two pass algorithm in which structural properties and projection profile based segmentation .Richard Casey & Eric Leolinet proposed 4 different techniques for segmentation i.e Classical, Region based segmentation, Holistic and hybrid. Dharam veer proposed horizontal and vertical projection profile for simple Gurumukhi script segmentation.

ISSN: 2395-7639

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

(A Monthly, Peer Reviewed Online Journal)

Visit: www.ijmrsetm.com

Volume 3, Issue 7, July 2016

A. Segmentation Methodology

The process of segmentation includes isolating line, word, Individual characters from input image. Profile based method used for this process.

a. line segmentation

Algorithm

{
Start
Read preprocessed image
Rotate image by appropriate angle
Plot vertical histogram projection
Find row wise summation.
Crop the image for successive white pixels
Repeat the procedure till the last pixel
Save image.
}

b. Word segmentation

Algorithm

{
Start
Read image
Plot vertical histogram projection
Find column wise summation
Crop image Repeat the steps
Save images
}

c. Character segmentation

Repeat the process of word segmentation for separating each character from word

3. Feature extraction

Feature extraction includes finding the set of such features which defines the shape of character precisely and uniquely. Each character has set of feature which is represented by feature vector which because of its identity. Sinha and Mahabala describe the system for printed Devanagari characters stores structural descriptive features for each character in the script. Statistical features like horizontal zero crossings, moments, vertex points are considered by Veena Bansal and Sinha. Govindaraju et al. introduces features such as gradient features for the Devanagari script. The sobel operator uses the intensity changes in a small neighborhood of each pixel the for measuring the magnitude as well as direction, gradient features are computed. For above threshold values of magnitude of gradient feature ,a gradient map is computed . In 1946 D. Gabor introduces unique strategy for using signal in that time and frequency are equally important. For our work, we use 2D Gabor filter for feature extraction for character recognition.

A. Ease of use

Different feature extraction techniques are used for recognition among them the directional feature is very popularly used they are very efficient for character recognition .Wang et al compared the Gabor features with contour direction (chain code features).Gabor filters are used for detecting local and structural patterns of image, and widely used for analysis of text and object recognition. The gabor filter and gradient features having few common properties both are applicable to binary as well as gray scale images and not effected by noise. In Devanagari script , there are presence of half characters, more than one shape of single character ,different characters are created by combining multiple characters and this all make recognition task more difficult. The Gabor filter is used to increase the recognition rate specially for pattern recognition.

B. Gabor filter computation

2D Gabor filter selects of both direction and orientation. Its Gaussian envelope with impulse response function (IRF) is a sinusoidal wave with oscillating orientation and frequency f0 is used for feature extraction. Apply Gabor filter function to every part of image and also calculate some regional features like, area, orientation, pixel density to produce the feature vector of every character. Some mathematical computation has to be done for Gabor feature extraction.

ISSN: 2395-7639

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

(A Monthly, Peer Reviewed Online Journal)

Visit: www.ijmrsetm.com

Volume 3, Issue 7, July 2016

Impulse response function represented mathematically as,

$$g(x,y;f_0,\theta)=\frac{1}{2\pi\sigma_x\sigma_y}.\exp^{-\left[\frac{x'^2}{2\sigma x^2}+\frac{y'^2}{2\sigma y^2}\right]}.\exp^{i2\pi(u_0x+v_0y)}$$

$$=\frac{1}{2\pi\sigma_x\sigma_y}.\exp^{-\left[\frac{x'^2}{2\sigma x^2}+\frac{y'^2}{2\sigma y^2}\right]}.\exp^{(i2\pi f_0x')}$$

C. Oscillating Frequency

Oscillating Frequency can be calculated as,

The oscillating frequency is calculated as,

$f_{ij}= v^j f_{max,}$, $j=\{0,\ldots\ldots,n-1\}, v=\{\sqrt{2}, 2, 2\sqrt{2}\}$

where

v=scaling factor

n=total no. of frequency

D. Orientation

Orientation

We sample the orientation between $[0,\pi]$ uniformly,

$\theta_k=k\,\pi/n$

where

n=no.of orientation

k={0,……n-1}

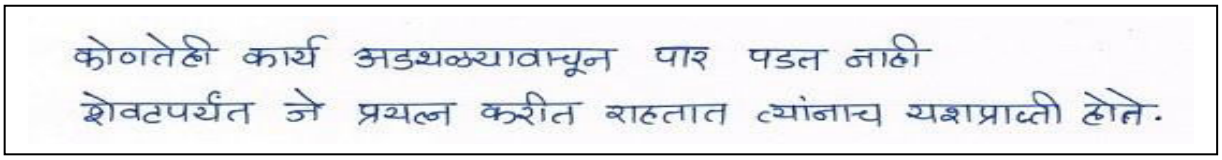
IV. CLASSIFICATION

For Devanagari character recognition ,classifiers like ANN , HMM, SVM,, MQDF etc., have been used .For training of these techniques, patterns i.e features applied to the system and the classifier rectify itself to reduce rate of errors of wrong classification errors for applied patterns. These types of fully prepared systems with desired features are used for the classification of different unknown testing patterns which is input to recognition system. ANN is example of such system , for better classification of patterns it adjusts the weights of its links from the training patterns .For the machine text having different font Meshesha et al. and Kompalli et al.[13] mention two techniques. SVM is used by Jawahar et al. for classifying machine printed script.

Name of feature extraction Methods	Size of feature vector	Classifier	Recognition Rate
Radon features	703	K-nn	90.4
Directional features	72	Euclidean distance	94
		K-means	96
		K-medoids	97
		K-nn Linear classifier	94
Zoning	144	AIS	98.5
DFT	72	K-nn	89.33

5. RESULTS

1. Original image



ISSN: 2395-7639

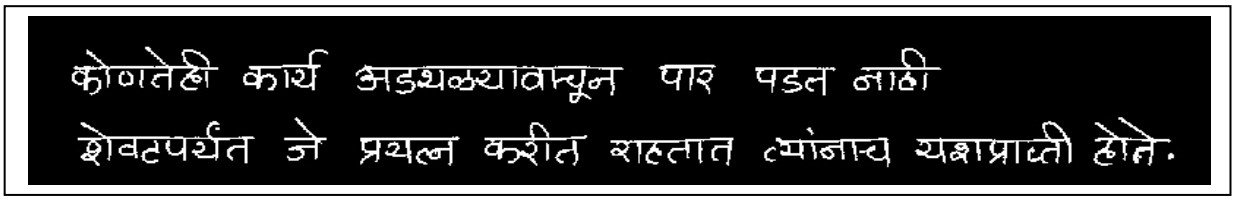
International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

(A Monthly, Peer Reviewed Online Journal)

Visit: www.ijmrsetm.com

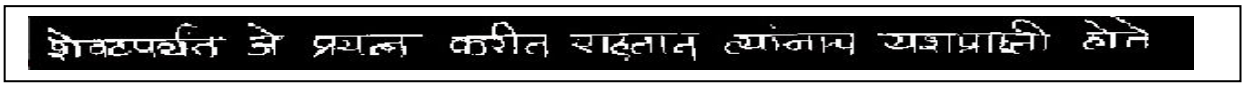
Volume 3, Issue 7, July 2016

2. Pre-processing

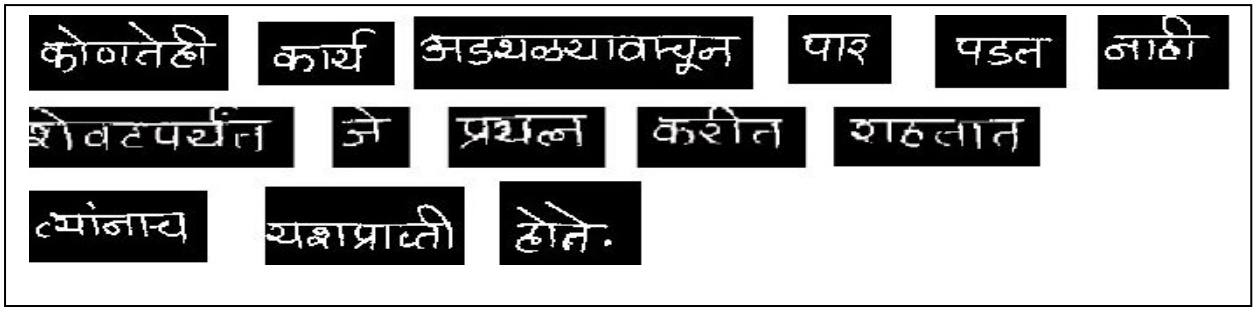


3. Segmentation

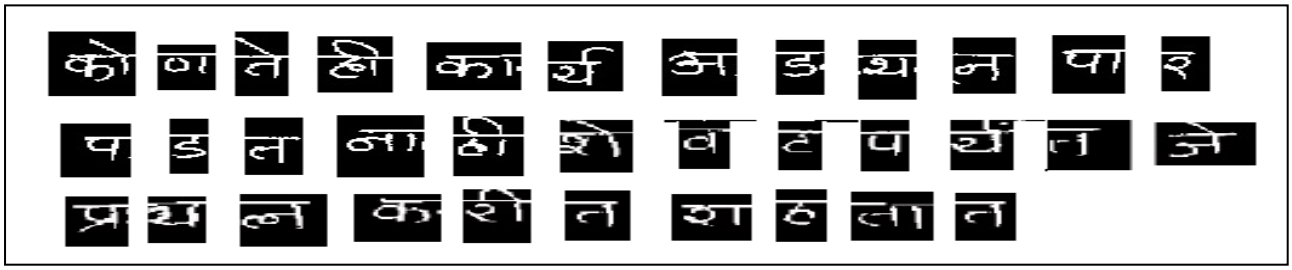
a) Line Segmentation



b) Word Segmentation



C) Characters segmentation



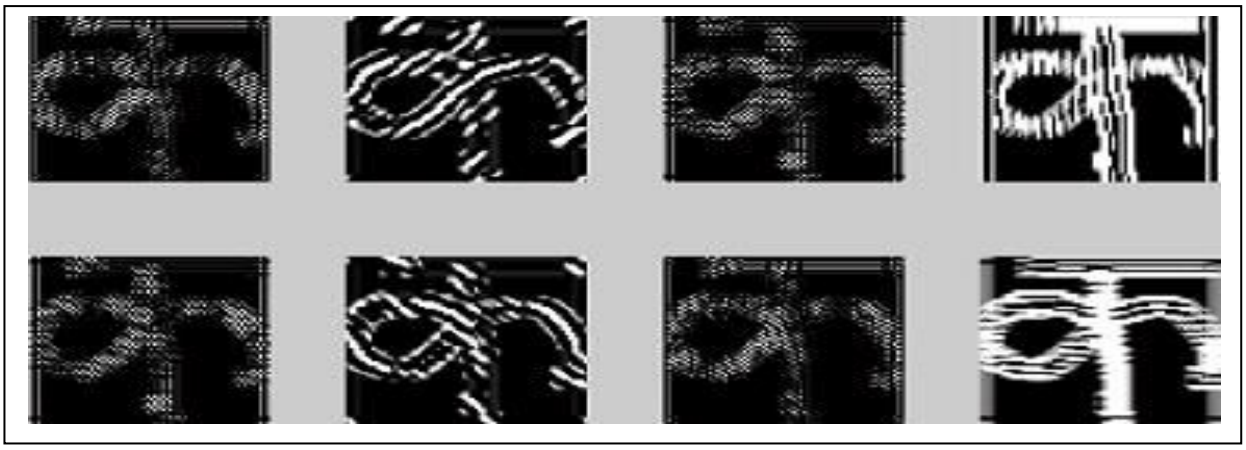
ISSN: 2395-7639

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

(A Monthly, Peer Reviewed Online Journal)

Visit: www.ijmrsetm.com

Volume 3, Issue 7, July 2016



V. CONCLUSION

Optical character recognition involves pre-processing, segmentation, feature extraction, classification and recognition of Script. As compare to Roman script Devanagari character recognition is more difficult. Input to the system is scan image and output becomes editable text. So, we can save the output in the form of text file. Some times during scanning document get titled for that we used skew correction algorithm. After that for recognition the individual character is input here we segment the document in to line, words and then in characters. For feature extraction we use Gabor filter ,but here rather than using Gabor filter for whole image we firstly divide image into dual parts then segment it. And finally brief survey of classification is studied for our future work.

REFERENCES

[1]Veena Bansal and R. M. K. Sinha , *"Integrating Knowledge Sources in Devanagari Text Recognition System,"* IEEE Transaction on system,man, and cy-bernetics—Part A ,System and Humans , Vol. 30, No. 4, July 2000.

[2]U. Bhattacharya and B. B. Chaudhuri , *"Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals,"* IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 3, pp. 444–457, Mar. 2009.

[3]Ashwin S Ramteke, Milind E Rane , *"Offline Hand-written Devanagari Script Segmentation,"* Interna-tional Journal Of Scientific & Technology Research Volume 1, Issue 4, MAY 2012.

[4]Gabor D. Theory of communication, Part 1: The analysis of information. Journal of the Institution of Electrical Engineers-Part III: Radio and Communi-cation Engineering. 1946; 93(26):429–41

[5]Urolagin S, Prema KV, Reddy NVS, *"Illumination invariant character recognition using binarized Gabor features,"* IEEE International Conference on Computational Intelligence and Multimedia Applications. 2007; 2:423–7

[6]Vedat Tavsanoglu and Ertugrul Saatci , "*Feature Ex-traction for Character Recognition Using Gabor-type Filters Implemented by Cellular Neural Networks,"* IEEE International Workshop on Cellular Neural Networks and Their Applications proceedings,2000.

[7]P. Bao-Chang,W. Si-Chang, and Y. Guang-Yi *,"A method of recognizing hand printed characters,"* in Computer Recognition and Human Production of Handwriting, OH: World Scientific, 1989, pp. 37–60.

[8]Mamatha Hosalli , Ramappa and Srikantamurthy Kr-ishnamurthy , *"A Comparative Study of Different Feature Extraction and Classification Methods for Recognition of Handwritten Kannada Numerals,"* International Journal of Database Theory and Application Vol. 6, No. 4, August, 2013 [9]N. Arica and F. T. Y. Vural ,"*An overview of character recognition focused on off-line handwriting,"* IEEE Trans. Syst., Man, Cybern. C: Appl. Rev., vol. 31, no. 2, pp. 216–233, May 2001.

[10]V. Bansal and R. M. K. Sinha , "*Partitioning and searching dictionary for correction of optically read Devanagari character strings,"* Int. J. Document Anal. Recognition., vol. 4, pp. 269–280, 2002.