Application for Handwritten Devnagari Optical Character Recognition

Prasad Chavan, Suyog Sankpal, Akshay Sonawane, Shahid Shaikh

Abstract—

Recently research towards Indian handwritten character recognition has increased.Many such approaches have been proposed by the researchers towards handwritten Indian character recognition and many such similar kind of systems are available in the literature. In OCR domain, a unique feature extraction method and unique classification algorithm is not efficient. As a result an extraction approach i.e. based on a structural analysis for recognition of offline handwritten Marathi character is proposed.

Index Terms—**Preprocessing, Image binarization, Morphological operators, Dilation, Skew Correction, Segmentation, Feature extraction**

I. INTRODUCTION

Handwritten character recognition is used frequently to describe the ability of computer to translate human writing into text. The technical challenges in character recognition arise mainly from three sources. First the symbols: the set of idealized shapes that can occur, often in a hierarchy where simple symbols are assembled into more complex ones, at several levels of organization. Second is deformation: the range of shape variations that each symbol is allowed to undergo, including
geometric transformation (translation, rotation, scaling, Stretching, etc.) and more complex or time dependent distortion. Third is an image defect:
the imperfections in image due to printing, optics, scanning, quantization, binarization [1], etc. Handwriting and machine print demand a different approach.
Handwriting consists of elongated strokes, whereas the machine print consists of regularly spaced blobs. Approx. 500 million people around the world use Devnagari script. It provides written form to over forty languages which includes Hindi, Konkani and Marathi. It's a logical composition of its constituent symbols in two dimensions. It has a horizontal line drawn above all characters.A marked distinction in Devnagari script from the scripts of Roman genre is the fact that a character represents a syllabic sound. While most work has been published, less is reported for handwritten Devnagari script. One of the first such attempts for handprinted characters has been by and for typed Devnagari script by Sinha and Bansal.Sinha and Bansal divided the typed words in three different strips and separated it in top strip, core strip and bottom strip and achieved 93.6% performance on individual characters. Recognition of Devnagari text in Sanskrit manuscript 'Saddharmapundarika' is achieved with an accuracy of 98.09% using structural features and neural networks for classification. Database evaluation methods are given in and database for Devnagari numerals has been collected from mail addresses and job application forms. Machine recognition of online handwritten Devnagari characters has been reported in with 84-87% accuracy. In online Devnagari script recognition is attempted with 86.5% accuracy on a database of 20 writers. A combination of on-line and offline features has been used. In Binary Wavelet transform it is used for feature extraction of handwritten Devnagari characters. In a survey of different techniques used for feature extraction in OCR of different scripts is given. Recently in, Quadratic classifier based method is proposed with 83% accuracy.

II. HE DEVNAGARI CHARACTER SET

The basic character set consists of 48 characters in which there are 12 vowels (swar) and 36 consonants (Vyanjan). A unique property of Devnagari script is the formation of conjuncts (Yuktakshar) that combines two (bi-consonantal) or three (tri-consonantal) consonants.About 176 bi-consonantal and 24 tri-consonantal conjuncts can be formed. Their formation is by the simple rules and restrictions of the language of application. Each consonant and the conjuncts can be further modified by vowel modifier (Matra), likewise to that of the formation of conjuncts, combinations of vowels, alongwith the nasal sounds, gives rise to combines in vowels also.There are as many characters in the Devnagari script similar to the syllables in the spoken language. The 45 characters of the basic handwritten character set for experimentation is based on their present-

Manuscript received March 07, 2014.
Shahid Shaikh, 4th Year Undergraduate Student, Department of CSE, TCOER, Pune, India (Email-shahidsk444@gmail.com)
Suyog Sankpal, 4th Year Undergraduate Student, Department of CSE, TCOER, Pune, India (Email-suyog5835@gmail.com)
Akshay Sonawane, 4th Year Undergraduate Student, Department of CSE, TCOER, Pune, India (Email-s23aksh@gmail.com)
Prasad Chavan, 4th Year Undergraduate Student, Department of CSE, TCOER, Pune, India (Email-chavan.prasad7001@gmail.com)

day usage. The handwritten character set is shown in the following Fig. 2.

Fig. 2: Basic character set (handwritten).

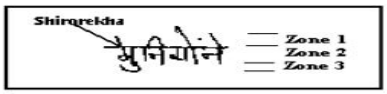
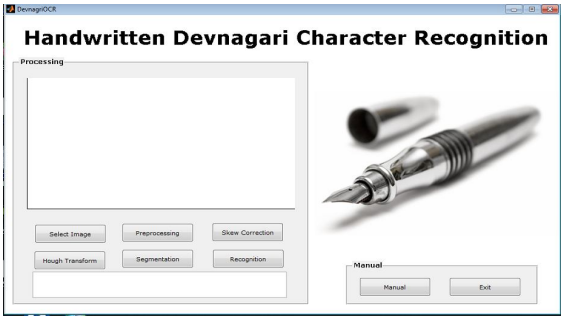


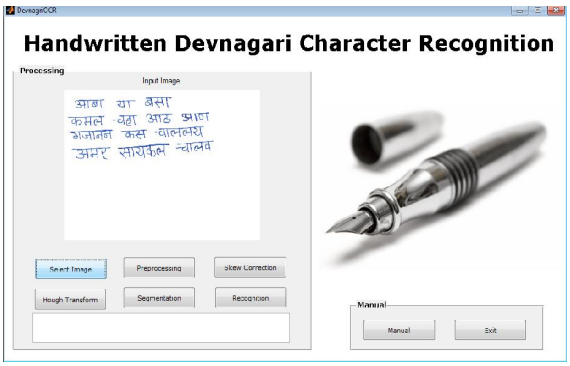
Fig. 3: 'Devnagari' word with modifiers and zones.

There is a horizontal header line present on the top of each character, called the 'Shirorekha' shown in Fig. 3. This line serves as a reference to divide the character into two distinct portions: Head and Body or zone 1 and 2, if the top modifier (matra) is present. The lower modifier occupies zone 3. The basic character width ranges from very small to large going through lots of medium sizes.

III. PROPOSED SYSTEM

In the proposed system, we aim at recognizing handwritten Marathi compound characters. This is done by employing multiple feature extraction and classification stages. At first, the character is pre-classified into one of 24 classes based upon the structural features. The handwritten Devanagari script is scanned. This two stage structural classification is followed by character normalization. Three different features are extracted from the normalized character and applied to three neural networks built for each structural class. It includes Preprocessing, Segmentation and Feature Extraction of Image. The final recognition is decided based upon Training and majority voting criterion. If the outputs from all the networks differ, then the output from the network with modified wavelet approximation features is selected. The neural networks are trained prior to testing, with sufficient amount of training samples. This is done by applying features to the networks that are extracted from pre-classified characters in the database. In the testing phase, the weights and biases, fixed during training, are used to get the recognition result as shown in the following fig. The next section discusses the proposed system in detail.





Preprocessing

The pre-processing stage handles steps, Image Binarization and Noise Removal.

Image binarization

In this process, a grey scale image is converted to binary image containing only pixel values of ones and zeros. This separates the area of interest from the background information which is essential. The method to binarization[1] is to change the pixel values of area of interest (here greatly characters or text) to higher value i.e. to one and to make all other part zero by fixing a threshold value. The pixels values below this threshold are to be converted to zero and above this threshold to one. The method for binarization applies morphological operator along with thresholding.

Morphological operators

Morphology is a term that refers to the description of shape and area of an object. Morphological operators usually process the objects in input image according to its shape. These operators apply a structural element on the objects of the image and produce an output image of the same size. The found value for each pixel in the output image of morphological operations is based on the comparison of the corresponding pixels and its neighboring pixels.

The basic operations using morphological operators are dilation and erosion. The former makes the boundary of the image undergo an expansion by addition of pixels to the boundary region and the later results in boundary shrinkage due to the removal of pixels from the boundary region. The number of pixels included or removed from the objects in an image is dependent on the size and shape of the element used to process the image.

Dilation

The value of the o/p pixel is *max* value of all pixels in input pixel's neighborhood. As regards to the binary image, if any of the pixels is set to the value 1, then the o/p pixel is set also to 1. The dilation of a gray scale image is illustrated in Figure.

The structuring element defines the related pixel of interest, which is circled. The morphological dilation function sets the value of the output pixel to 1.

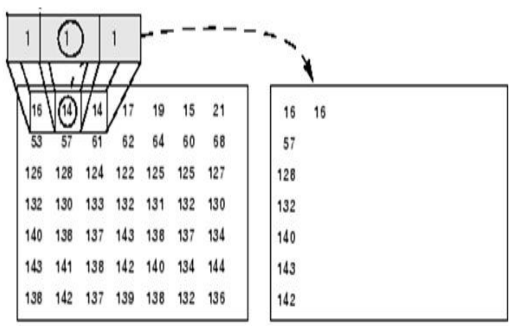
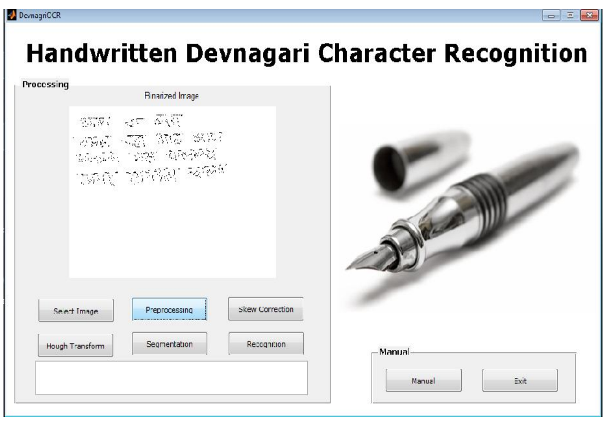


Figure : *Dilation of a Gray Scale Image*



Segmentation

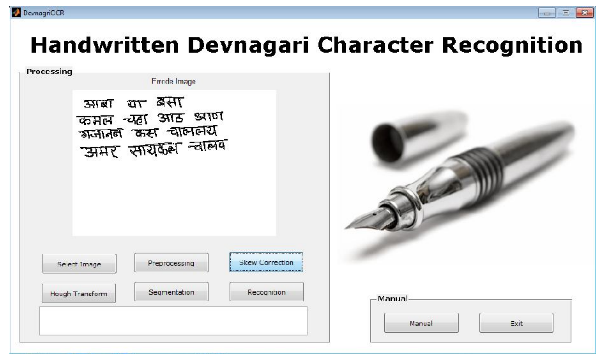
Image segmentation is the process of partitioning a digital image into various multiple segments. The main aim of segmentation is to simplification and changing the representation of an image into something i.e. much more meaningful and easier to analyze. Image segmentating is basically used to locate objects and a boundary (lines, curves, etc.) within images.It is the process of assigning a label to every pixel in an image such that pixels with the similar label share certain visual characteristics.

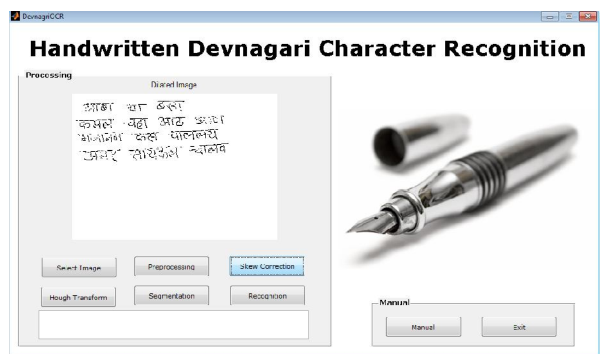
The result of segmentation is a set of segments that in all cover the entire image, or set of contours which are extracted from the image (see edge detection). Each pixel in a region is similar to some computed property, such as color, intensity, or texture. Adjacent regions are differing with respect to the similar characteristics.

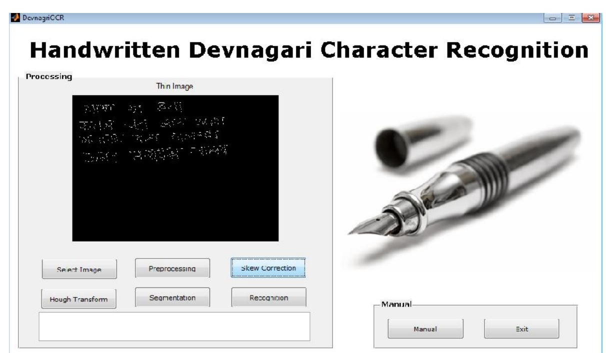
Skew Correction

On detection of the words from a given text, we can run our algorithm on the individual words and angular skewness [3] of words can be removed.

In this method we use *Hough* transform [4] to find straight lines within the words. In Hindi the headline is the longest straight line, using this property of Hindi language the longest straight line is selected which we obtained by using Hough's transform. We calculate the angle of inclination of the line corresponding to the x-axis. This gives the skew angle of the word, rotating the word by the skew angle removes the angular skewness of the word.







Feature extraction

It plays a major role in improving the recognition accuracy. They are extracted from
binary characters. Thus, the characteristics are used for classification depends on the shape variations. Many characters are same in shape or slight variation in the individuals writing style may result into misclassification. The features selected should handle and overcome all these problems. A single feature extraction and classification recognize a character which may not be

recognized by other systems. Hence there is a need of implementation of a hybrid system that can recognize the characters over a wide range of varying conditions. The cropped characters which are present in each structural class are resized to a fixed size before extracting the features. Three different features are extracted for applying to the three separate neural networks. The features which are extracted are pixel density features, Euclidean distance [5] features and modified approximation wavelet features.

Why MATLAB?

MATLAB provides a comprehensive set of standard algorithms and graphical tools for image processing process.You can restore noisy or degraded images, enhance images for improved intelligibility, extract features, analyze shapes and textures, and register two images. Most of the toolbox functions are written in the open MATLAB language, giving you the ability to inspect the algorithms, modify the source code, and create your own custom functions.

MATLAB provides a number of features to document and lets you share your work. You can integrate your MATLAB code with other languages and applications, and distribute your MATLAB algorithms and applications.

Key Features

High-level language for technical computing

Development environment for managing code, files, and data. Interactive tools for iterative exploration, design, and problem solving Mathematical functions for linear algebra, statistics, Fourier analysis, filtering, optimization, and numerical integration[1].

IV. CONCLUSION

This paper presents a system for handwritten devnagri character recognition for Devnagri script. A huge character dataset is collected from various writers and used for database creation for neural network training and testing. The recognition of characters is done using multistage multi-feature hybrid recognition scheme.

ACKNOWLEDGMENT

Authors are thankful to the Department of Computer Engg, Trinity College of Engineering, Pune, Maharashtra, India, for providing the necessary facilities for carrying out this work. Authors gratefully acknowledge the support given by University Of Pune, Pune, Maharashtra, India, for carrying out this research work. Authors also grateful to Prof. Anup Raut, Trinity College of Engineering, Pune, Maharashtra for their time to time guidance and support towards carrying out the research work. Authors are also thankful to Mrs.Sonali Pathare and Ms.Sampada Pingale for influencing us to take forth this research concept and work towards its implementaion and also the anonymous reviewers for their valuable suggestions towards improving the paper.

REFERENCES

[1] N.ARICA, F.T.Y. VURAL, "AN OVERVIEW OF CHARACTER RECOGNITION FOCUSED ON OFFLINE HANDWRITING", IEEE TRANS ON SYSTEM ,MAN,CYBERNATICS-PARTC, VOL 31,N0.2(2001)

[2] OVIND TRIER, ANIL JAIN AND TORFINN TAXT," A FEATURE EXTRACTION METHODS FOR CHARACTER RECOGNITION-A SURVEY", PATTERN RECOGNITION, VOL 29, NO-4, AND PP 641-662, 1996.[3] U.PAL AND B.B. CHAUDHURI," AN IMPROVED DOCUMENT SKEW ANGLE ESTIMATION TECHNIQUES", PATTERN RECOGNITION LETTERS 17:899-904, 1996.

[3] B.B. CHAUDHRURI AND U.PAL, "A COMPLETE PRINTED OCR", PATTERN RECOGNITION,(5):531-549, 1998.

[4] REJEAN PLAMONDON AND SARGUR N. SRIHARI, "ON-LINE AND OFF-LINE HANDWRITTEN RECOGNITION" A COMPREHENSIVE SURVEY", IEEE PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL 22, NO. 1, JANUARY 2000.

[5] U. PAL, T. WAKABAYASHI, F. KIMURA, "COMPARATIVE STUDY OF DEVNAGARI HANDWRITTEN CHARACTER RECOGNITION USING DIFFERENT FEATURE AND CLASSIFIERS", 10TH INTERNATIONAL CONFERENCE ON DOCUMENT ANALYSIS AND RECOGNITION 2009.

[6] ZHIYI ZHANG, LIANWEN JIN, KAI DING, XUE GAO,"CHARACTERSIFT: A NOVEL FEATURE FOR OFFLINE HANDWRITTEN CHINESE CHARACTER RECOGNITION" 10TH INTERNATIONAL CONFERENCE ON DOCUMENT ANALYSIS AND RECOGNITION, 2009

[7] T.V.ASWIN AND P S SASTRY, "A FONT AND SIZE-INDEPENDENT OCR SYSTEM FOR PRINTED KANNADA DOCUMENTS USING SUPPORT VECTOR MACHINES", SADHANA VOL.27.PART I, PP.35-58, FEBRUARY 2002.

[8] T.V.ASWIN, "A FONT INDEPENDENT OCR FOR PRINTED KANNADA USING SVM", MASTER THESIS, INDIAN INSTITUTE OF SCIENCE, BANGALORE, 2000.

[9] VEENA BANSAL AND R. M. K. SINHA, "A DEVANAGARI OCR AND A BRIEF OVERVIEW OF OCR RESEARCH FOR INDIAN SCRIPTS", PROCEEDINGS OF STRANS01, IIT KANPUR 2001

Prasad Chavan, 4th Year Undergraduate Student, Department of CSE, TCOER, Pune, India
Email-chavan.prasad7001@gmail.com

Shahid Shaikh, 4th Year Undergraduate Student, Department of CSE, TCOER, Pune, India
Email-shahidsk444@gmail.com

Akshay Sonawane, 4th Year Undergraduate Student, Department of CSE, TCOER, Pune, India
Email-s23aksh@gmail.com

Suyog Sankpal, 4th Year Undergraduate Student, Department of CSE, TCOER, Pune, India
Email-suyog5835@gmail.com