ISSN 2395-1621

Optical Character Recognition for Handwritten Marathi Script

Jitendra Kambli[#1], Viraj Kolekar[#2], Shital Hire[#3], Karishma Sehrawat[#4], Prof. V. P. Salve[#5]

[1]jitukambli@gmail.com
[2]crazycasanova67@gmail.com
[3]Shital.hire247@gmail.com
[4]karishma.sehrawat02@gmail.com
[5]veenaya.salve@gmail.com

[#1234] Department of Computer Engineering, Shatabdi Institute of Engineering, Research, Nashik, Maharashtra, India
. [#5]Professor, Department of Computer Engineering, Shatabdi Institute of Engineering, Research, Nashik, Maharashtra, India

ABSTRACT

The Optical Character Recognition is the main method of character Recognition. Many times it gives unsatisfactory rate of character segmentation. Segmentation is a primary need of every OCR system. The lines, words and characters are separates from the image text documents. The Segmentation algorithm is used for accuracy of OCR system. Segmentation of Handwritten Devnagari text is difficult when compared with Printed Devnagari or Printed English or any other Printed document its structural complexity and increased character set. The image document contains vowels, consonants. In that again characters also overlap each other. The method of profile based can segment also characters and non-overlapping lines. This paper addresses the segmentation of Handwritten Devnagari text document, this is most popular script of Indian sub-continent it is divided into lines, words and characters. The proposed algorithm is based on projection profiles. Experimental results it is observed that 100% line segmentation and about 98% character segmentation accuracy can be achieved with overlapping lines, words and characters.

***Keywords*— segmentation, projection profiles, features extraction, zoning features, projection histogram features, Euclidean Distance classifiers**.

ARTICLE INFO

Article History

Received :10st March, 2015

Received in revised form :

15th March, 2015

Accepted :17nd March, 2015

Published online :

20th March 2015

I. INTRODUCTION

Handwriting recognition (or HWR) is the ability of a computer to get and interpret intelligible handwritten input from sources such as paper documents, images, touch-screens and other devices. Depending on the manner in which data is acquired, the domain of handwritten character recognition is divided into two types.

- On-line Handwritten Recognition
- Off-line Handwritten Recognition

On-line handwriting recognition-On-line handwriting recognition involves the automatic conversion of text and it is written on a special digitizer and PDA, where a sensor picks up the pen-tip movements and pen-up or pen-down switching. That is digital ink and can be regarded as a dynamic representation of handwriting. The obtained signal is translate into letter codes.it is usable within computer and text-processing applications.

The elements of an on-line handwriting recognition interface include:

- User can write data using pen or stylus.
- To touch sensitive surface, which may be integrated with, or adjacent to, an output display.
- A software application that interprets the movements of the stylus across the writing surface.

Off-line handwriting recognition-The image of the written text may be sensed "off line" from a piece of paper by optical scanning (optical character recognition) or intelligent word recognition. For offline character recognition the following cases can be considered: recognition of one or affixed number of fonts (Fixed Font OCR and Multi font OCR), any printed font (Omni font OCR), isolated hand printed characters (Handwriting OCR) and unconstrained

handwriting (Script Recognition) However, handwritten character recognition is a challenging task because of variability of writing styles of different writers from different environment. The task becomes more difficult when the text document quality is poor and if the characters are written very close to each other. In addition, some of the Indian scripts have compound characters. that characters have similar shapes that require advanced and complex techniques for recognition.

II. OBJECTIVES

In the research scenario, the different images need to be considered as input for identify the characters by processing the feature extraction techniques using OCR. The main aim is to identify and extract the features of he handwritten Devnagari script.

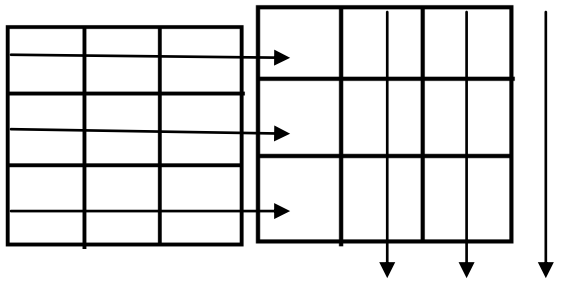
(i) To study various phases of OCR.
(ii) To study various techniques used for feature extraction.
(iii) To extract the set for handwritten Devnagari script.
(iv) Recognition of handwritten Devnagari script by using classification techniques.

III. PROPOSED METHODOLOGY

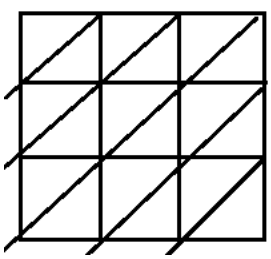
The number of steps considered for identify the features of Devnagari Scripts.

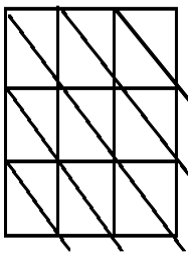
(i) **Zoning**: The frame containing the character is divided into several overlapping or non-overlapping zones and the densities of object pixels in each zone are calculated. Density is calculated by finding the many number of object pixels in each zone and dividing it by total number of pixels

(i) **Projection Histogram Features**: Calculate number Of pixels in specified direction.
a) Horizontal
b) Vertical



Horizontal Projection Vertical Projection





Diagonal-I Projection Diagonal-II Projection

c) Left Diagonal
d) Right Diagonal

(ii) **Distance based Profile Features**: Distance of No. Of pixels from Boundary Box of characters:
a) Left
b) Right
c) Top
d) Bottom

(iii) **Classifiers:**
In an OCR process classification stage assigns labels to character images on the base of features extracted and the relationships among them. In simple terms, this is part of OCR recognizes individual characters and returns the output in character processing form.

The two basic phases of any classification problem that is training and testing. In training phase the classifier learns the relationship between samples and their labels from samples that are been labelled whereas, in testing phase analysing of errors in the samples is performed in order to evaluate classifier's performance. For better performance it is desirable to have a classifier with minimal test error.

IV. General Process of Optical Character Recognition (OCR)

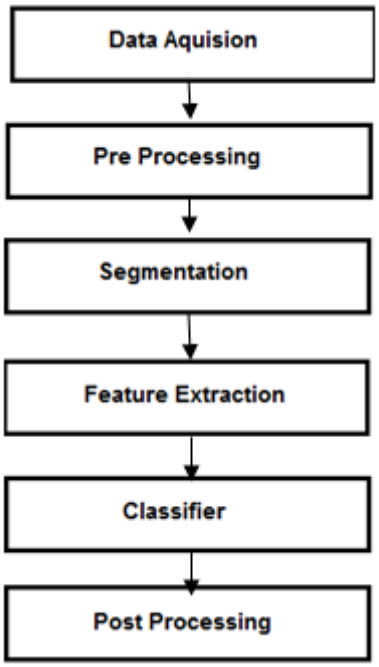


Fig. 1: General Process of OCR

Major Steps in OCR:

a) Data Acquisition
b) Pre-processing
c) Segmentation
d) Feature Extraction
e) Classification
f) Post-processing

(a) Data Acquisition

The input to the OCR system is the scanned document image. This input image have specific format such as .jpeg, .bmp etc. This image is acquired through a scanner and digital camera or any other suitable digital input device. After image acquisition, the image data goes through following processes

(b) Pre-Processing

The raw data is subjected to a number of preliminary processing steps to make it usable in the descriptive stages of character analysis. Pre-processing aims to generate data that are easy for the OCR systems to operate accurately. The main objective of pre-processing are:

2.2.1 Binarization- Document image binarization (Thresholding) refers to the conversion of a grey-scale image into a binary image.

Global binarization get one threshold value for the entire document image which is often based on an estimation of the background level from the intensity histogram of the image.

Adaptive (local) binarization uses different values for each pixel according to the local area information.

2.2.2 Noise reduction (morphological operators)- Optical scanning devices may introduce noises, e.g. disconnected line segments, gaps in line, filled loops, etc. Noise removal stage removes isolated specks and holes in the characters. Noise reduction improves the quality of the document. Two main approches used are filtering and Morphological operations.

2.2.3 Normalization-This stage removes some of the variations in the image that do not affect the identity of the input data and provides a tremendous reduction in data size. Thinning remove the shape information of the characters.

2.2.4 Skew correction- Skew Correction methods are used to align the paper document with the coordinate system of the scanner. Main approches for skew detection include correlation, projection profiles.

2.2.5 Slant removal-The slant of handwritten texts varies from user to user. One of the measurable factor of different handwriting styles is the slant angle between longest stroke in a word and a vertical direction. Slant removal methods are used to normalize the all characters to a standard form.

(c) Segmentation:

Segmentation is by far the most important aspect of the pre-processing stage. It allows the recognizer to extract features from each individual character. In the more complicated case of handwritten text, the segmentation problem becomes much more difficult as letters tend to be connected to each other, overlapped or distorted. Segmentation is done to break the single text line, single word and single character from the input document. For isolated characters or numerals, segmentation task is not that difficult. However, for joint and complex strings more advanced techniques required to be employed.

There are two types of segmentations:

1. External segmentation, that isolates various writing units such as paragraphs,
Sentences or words,
2. Internal segmentation,

(d) Feature Extraction:

In feature extraction stage, each character is represented as a feature vector, which becomes its identity. The major goal of feature extraction is to extract a set of features, which maximizes the recognition rate with the

least amount of elements and to generate similar feature set for variety of instances of the same symbol. Due to the nature of handwriting with its high degree of variability and imprecision obtaining these features, is a difficult task. Feature extraction methods analyse the input document image and select a set of features that uniquely identifies and classifies the character

(e) Classification:
Feature extraction stage gives us the feature vector that is used for classification. Classification is the decision making step in the OCR system that makes use of the features extracted from the previous stage in the process. do the classification we have a data bank to compare with many feature vectors. A classifier is needed to compare the feature vector of input and the feature vector of data bank. The selection of classifier depends upon training set and number of free parameters. There are various types of existing classical and soft computing techniques for handwritten recognition.

(f) Post-processing:
The purpose of this step is the incorporation of context and shape information in all the stages of OCR systems is necessary for meaningful improvements in recognition rates. A dictionary can be used to correct minor errors.

V. EUCLIDEAN MINIMUM DISTANCE CLASSIFIER

At this step, in order to recognize the character, a distance between various unknown or input feature vector X and the reference vectors, M from the training set it must be computed. The distance will be computed it is based on Euclidean model as:

$$d(X, M_k) = \sqrt{\sum_{i=1}^{N} (x_i - m_i^k)^2}$$

The experimental result shows in Table 1 and that classification based on multiple features that gives more accuracy than using only one feature information.

Feature Extraction	Percentage(%) of Extraction
Pixel Density(zoning)	92.77

VI. CONCLUSION

In this experiment, the proposed algorithm is tested with several document images. Some of the documents contained overlapping lines and characters. Even though it could divide all the documents in a robust way and then gave good results. But, it couldn't divide the touching lines and characters. The broken characters have been divided. Segmentation of the touching lines and characters may require some heuristic approaches

REFERENCES

[1] U. Pal, B.B. Chaudhuri. (2004): "Indian script character recognition: a survey, Pattern Recognition", 37,1887 – 1899

[2] B. Anuradhaand, Arun Agarwal and C. Raghavendra Rao. (2008): "An Overview of OCR Research in Indian Scripts", IJCSES, Vol.2, No.2.

[3] N. Otsu. (1979): "A threshold selection method from gray-level histograms", IEEE transactions on systems, Man and Cybernetics, Vol. Smc-9, No. 1.

[4] Binal M. Patel ,Intrusions Detection in Three tier Web Applications using Double Guard System

[5] C. V Lakshmi, C. Patvardhan. (2004): "An optical character recognition system for printed Telugu text, Pattern Analysis & Applications", Volume7,pp.190-204.

[6] optical character Recognition", Proceedings of the 2008 International Conference on Wavelet Analysis and Pattern Recognition

[7] R. C. Gonzalez and R. E. Wood (2004) : Digital Image processing , Person Education, Nitin Bhatia and Vandana 2010 :"Survey of nearest Neighbour Techniques IJCSIS "

