ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 7, July 2014

To Extract Feature of Handwritten Devnagri Script

Ajay Garg[1], Simpel Jindal[2]

Student, CSE, YCOE, Talwandi Sabo, India[1]

Assistant Professor, CSE, YCOE, Talwandi Sabo, India[2]

Abstract: This paper is about the study of OCR technique using Feature Extraction of Handwritten Devnagri script and provides the approach to recognize the half and full characters of Devnagri script. Devnagri script is the popular script of India or national script of India. This script is required for write Hindi, Nepali and Marathi Language. This paper will work on feature extraction of Hindi. Hindi Language consist of vowels, constants and various modifiers. Proper Feature Extraction technique on Hindi character is challenging. This paper will use Feature Extraction techniques to recognise handwritten Devnagri script document. The Feature Extraction in any handwritten script is very important part of OCR. If the efficient feature extraction technique is used, this will gives the efficient recognition of individual character easy. In this paper, the character will be divides in three parts horizontally as well as vertically to extract the feature of each divided part and store its value for provide the training to the feature extraction technique.

Keywords: Feature Extraction, Character Recognition, Devnagri Script.

I. INTRODUCTION

Optical character recognition is for recognizing the handwritten text as well as printed text by a computer. The scanning process converts the text document in image. For character recognition, the character codes are required and can be translated from text image. In OCR implementation consists of a number of steps followed by the actual recognition. Recognition of handwritten characters has been good research area for many years because of it has many applications in all fields. Devnagri script is the script for writing Hindi language. Hindi is the official language of India. Offline handwritten Hindi text recognition is need of the hour due to large number of application of Hindi OCR. Recognition of handwritten Optical character recognition is very difficult due to different writing styles of the different person. The techniques which we used for recognition of printed characters cannot be directly applied on handwritten text. Due to large number of characters and presence of half characters and some confusing characters makes the recognition process even more complex. Initially, there is need of data from many users. This has been found that the writing style of every user is different. So, the recognition the character is very difficult. In this, there is need of the recognition of character in Hindi by using some Feature Extraction Technique. There are some vowels in Hindi language, when they are written before or after the consonant they are known as the modifiers. In Hindi all the characters have a horizontal line at the upper part known as Shirorekha. These all are the features extracted by some techniques. The data is required from many people for check the writing style and has been founded that the style is very different for each and every person. Extract the feature of character and recognised.

Devnagri Script is used primarily for writing Hindi language, which is the world's 3rd most widely spoken language. Following are the properties of Devnagri Script:

(i) Writing style is from left to right.
(ii) No concept of upper and lower case characters.
(iii) There are 11 vowels, 33 consonants, compound characters, half characters and some special characters in the Language.

Devnagri Script has following challenges:

(i) Variability of writing style, both between different writers and between separate examples from the same writer overtime.
(ii) Similarity of some characters.
(iii) Low quality of text images
(iv) Unavoidable presence of background noise and various kinds of distortions

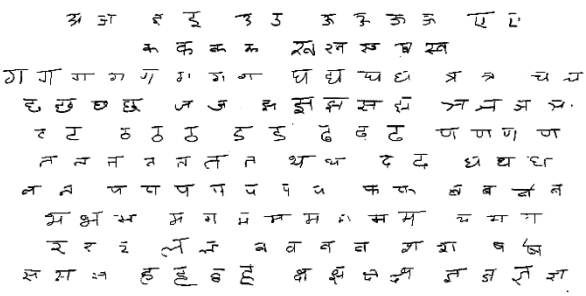

Fig 1 Handwritten character by different users

II. FEATURE EXTRACTION

Automatic reading of numerical fields has been attempted in several application areas such as online handwritten recognition on computer tablets; recognize zip codes on mail for postal address sorting, processing bank check amounts, and numeric entries in forms filled up by hand (e.g. Tax forms) and so on. While solving this domain of handwritten recognition many challenges are faced. As the handwritten digits are not always of same size, thickness, or orientation and position relative to the margins, many handwritten versions are even hard to recognize. Handwritten recognition is the ability of a computer to receive and interpret intelligible handwritten input from

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 7, July 2014

sources such as document, photographs, touch-screens & other devices.

The OCR consists of five stages out of these feature extraction and classification are very important. Feature extraction is the name given to a family of procedures for measuring the relevant shape information contained in a pattern so that the task of classifying the pattern is made easy by a formal procedure. The feature extraction stage analyses a text segment and selects a set of features that can be used to uniquely identify the text segment. The selection of a stable and representative set of features is the heart of pattern recognition system design. Among the different design issues involved in building an OCR system, perhaps the most significant one is the selection of the type and set of features. The main problem in OCR system is the large variation in shapes within a class of character. This variation depends on font styles, document noise, photometric effect, document skew and poor image quality. The large variation in shapes makes it difficult to determine the number of features that are convenient prior to model building.

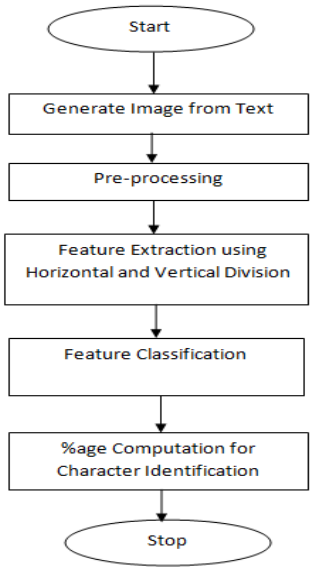

Fig 2 OCR Feature Extraction Steps

Though many kinds of features have been developed and their test performances on standard database have been reported, there is still room to improve the recognition rate by developing an improved feature.

III. LITERATURE REVIEW

In today's Research scenarios, there are different techniques, which have been discussed for character Recognition from the image. A multistage scheme for the recognition of handwritten Bengali characters has been considered. An analysis of the Bengali character set has been carried out to isolate specific high-level features that can help in forming smaller sub-groups within the character set. This analysis demonstrates how detection of these various high-level features might help formulate successful multistage OCR design [2]. The recent research is on optical character recognition (OCR) systems for Indian language scripts. The survey gives us knowledge on developing OCRs for Indian scripts and gives about the emerging status about the field. Peculiarities in Indian scripts, present status of the OCRs for Indian scripts, techniques used in them for recognition accuracies [4].

Author presented a robust and font independent Gurumukhi OCR system, which performs on old document in this the OCR is based on four classifiers operating in serial and parallel mode for combining the result of the classifier operating in parallel mode, a corpus based weighted voting method is used. The problem of broken character, which frequently appears in old documents, has also been tackled using a structural feature based algorithm [5].

A new representation method for recognition of handwritten characters, called LLF (Local Line Fitting), is presented. The method, based on simple geometric operations, is very efficient and yields a relatively low-dimensional and distortion invariant representation. An important feature of the approach is that no pre-processing of the input image is required. A black & white or gray-scale pixel representation is directly used without thinning, contour following, binary conversion etc. Therefore, high recognition speed can be achieved.

Experiments using this parameterization method and several classification procedures on handwritten digits and letters are reported [10]. Handwritten Hindi text recognition is latest research area in the field of optical character recognition. The segmentation based approach is used to recognize the text. The offline handwritten text is segmented into lines, lines into words and words into character for recognition [8].

IV. 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 the handwritten Devnagri script.

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

V. PROPOSED METHODOLOGY

The number of steps needs to be considered for identify the features of Devnagri 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 number of object pixels in each zone and dividing it by total number of pixels

Density= No. of Object Pixels in Each Zone/ Total No. of Pixels

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 7, July 2014

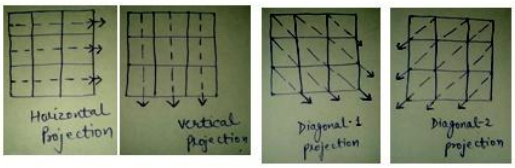

Fig 3 Feature Extraction Types

(i) Projection Histogram Features: Calculate no. of pixels in specified direction.
   a) Horizontal
   b) Vertical
   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 part of OCR recognizes individual characters and returns the output in character processing form.

The two basic phases of any classification problem are 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 analyzing 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.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have been proposed the Feature Extraction technique for identify the Handwritten characters of Devnagri script. The proposed Algorithm will be implemented in MATLAB and can read handwritten characters of Devnagri Script.

The implementation part will be covered in the next paper, which will demonstrate the real working of proposed algorithm.

REFERNCES

[1] G. S. Lehal, Chandan Singh(2000), "A Gurmukhi Script Recognition System", International conference on pattern recognition(ICPR'00),1051-4651/00.
[2] A.F.R. Rahman, R. Rahman and M.C. Fairhurst (2002), "Recognition of handwritten Bengali characters: a novel multistage approach", ScienceDirect, pp. 997–1006, Received 17 July.
[3] Marisa R. De Giusti, Maria Marta Vila and Gonzalo Lujan Villarreal (2006), "Manuscript Character Recognition Overview of features for the Feature Vector", JCS&T ,Vol. 6 No. 2, October.
[4] B.Anuradha Srinivas, Arun Agarwal and C. Raghavendra Rao (2008), "An Overview of OCR Research in Indian Scripts", International Journal of Computer Sciences and Engineering Systems, Vol.2, No.2, April.
[5] G. S. Lehal (2009), "Optical Character Recognition of Gurmukhi Script using Multiple Classifiers", in Proceedings of the International Workshop on Multilingual OCR.
[6] M. K. Jindal, R. K. Sharma and G. S. Lehal (2009), "Segmentation of Touching Characters in Upper Zone in Printed Gurmukhi Script" in Compute,Jan 9, 10, Bangalore, Karnataka, India.
[7] Pritpal Singh and Sumit Budhiraja(2011), "Feature Extraction and Classification Techniques in O.C.R. Systems for Handwritten Gurmukhi Script, in International Journal of Engineering Research and Application , Vol. 1, No. 4, pp. 1736-1739,Dec.
[8] N. K. Garg, L. Kaur and M. K. Jindal (2013), "Recognition of Offline Handwritten Hindi Text Using SVM" in International Journal of Image Processing (IJIP), Vol. 7, No.4.
[9] H. Swethalakshmi, Anitha Jayaraman, V. Srinivasa Chakravarthy and C. Chandra Sekhar,"Online Handwritten Character Recognition of Devanagari and Telugu Characters using Support Vector Machines" Indian Institute of Technology Madras, Chennai - 600 036, India.
[10] Juan-Carlos Perez, Enrique Vidal and Lourdes Sanchez, "Simple and Effective Feature Extraction for Optical Character Recognition" in CICYT project TIC93-0633-C02.