www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issue 12 Dec. 2016, Page No. 19354-19357

Word Recognition System Using ANN

Dr. J.Kejiya Rani,
Asst.Professor,
Dept.Of.CS&T,
S.K.University, Ananthapuramu.

Abstract: This paper deals with the concept of Optical Character Recognition (OCR) and its various stages like pre-processing, segmentation, feature extraction, classification and post-processing. The main objective of this paper is:

1. To study the existing techniques for recognition of handwritten Hindi text.
2. To develop and apply the new segmentation technique on the handwritten Hindi text collected from different users. .
3. To design a new feature set for the handwritten Hindi text.
4. Recognition of handwritten Hindi text.

Developments of OCR systems for the Indian language scripts are gaining importance because of their large market potential. With the advancement of information technology there has been a dramatic increase of research in this field since the beginning of 1980. In areas of automatic document analysis and recognition, the correct interpretation of characters is very important. Automatic recognition of handwritten characters is difficult due to several reasons, including varying writing styles of different persons and different writing devices. All this leads to characters of different sizes and skews, and strokes that vary in width and shape. The performance of any system for handwriting recognition can be evaluated by several factors, such as recognition rate, independence of the writing style, and speed of recognition.

The problem of printed character recognition is divided into three parts:

1. To segment the text into individual characters.
2. To extract features form the characters that are size, font and slant invariant.
3. To classify the characters using classifier.

Keywords: Document processing, Optical Character Recognition, Preprocessing, Segmentation,

1. Introduction:

The advent of computers has revolutionized stmade many new facilities creep in like shopping with ATM, net banking etc.. Now, time can be saved by ordering almost anything on net andthe technical development in 21century. Themaking the bill payments online. Thoughcomputations which used to take many minutes are solved within milliseconds. The letters were full of errors when typewritten. The use of computers changed the whole scenario. The errors could not only be corrected, but lot of effort was saved when the once entered material could be copied and pasted into another file. Use of carbon for making the copies of the letter has almost vanished. The speed and accuracy with which the computer performed the day to day tasks almost bewildered everybody. The work which required hours of effort was now done at a much faster pace. This made computer enter almost every sphere of life. Banking, offices, shops, all wore a new look with computer on each desk. This computer could do so many services, yet man remains unsatisfied. we wants more. In business applications a lot of time is wasted in entering data through keyboard. Now the need is felt to have something with which one doesn't have to input everything by typing through keyboard. A solution to this problem is the development of Optical Character Recognizer that can recognize the printed documents. It not only save the typing efforts but can also make the work paperless. Due to large potential of document processing in office automation, banks, library automation, postal services, it is gaining popularity in daily life. A way of making the computer recognize the printed material is the utmost need of the time.

1.1 Optical Character Recognition

Recognition of printed text with the help of computer is called Optical Character Recognition. Due to its large applications in different areas like post offices, banks, libraries and other document processing fields, computer simulation of handwritten or printed text is very

essential. Fast processing of image documents with the help of computer is the

main goal of OCR. In the past, lot of research has been done on Roman script. A lot of research has also been done on various Indian scripts like Bangla, Oriya, Devanagari etc.. [1][2][3][4][5]. Devanagari script is one of the major scripts of India and many languages like Hindi, Marathi, Nepali, Sanskrit etc. are written in it. Hindi is the national language of India that is based on Devanagari script.

Text recognition systems are categorized based on level as well as type of text they recognize

(Figure1.1).

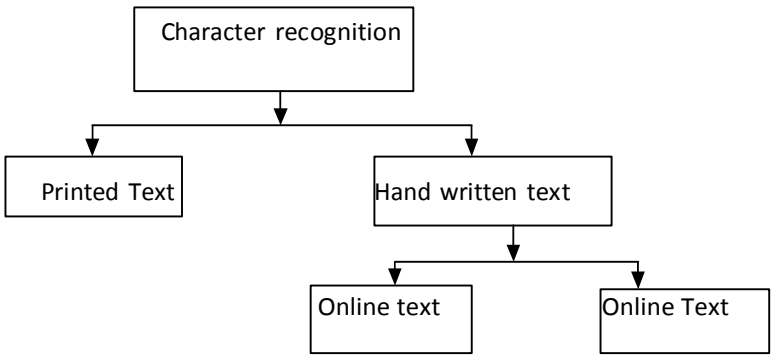

Figure 1.1: Different types of character recognition systems

In the past, researchers tried to develop the OCR for printed Hindi text. Many research reports are available on printed Hindi text. Efforts to develop the OCR for handwritten Hindi text are still going on. Due to different writing styles of people the recognition of handwritten text is a tedious task. The approach to recognize the text may be categorized as :

1) Holistic approach
2) Segmentation approach

The holistic approach recognizes the word as a whole without segmentation. The overall main features of the word are used for recognition. It is useful if the character set is small. The segmentation approach segments the word into individual characters before recognition. Then features are extracted from individual characters to recognize it. It is very useful in handwritten Hindi text due to large character set of the language.

1.2 Stages of Optical Character Recognition

The document is to be converted into electronic form before recognition. The original document is scanned and electronic form of the document in the form of bitmap image is produced. This process of converting the document into electronic form is called digitization. The changes in light intensity reflected from the document are stored as matrix of dots. The value of each dot is stored in binary digits (0 or

1). The digitized image goes through some processing steps. Most of the methods used for document processing include the following stages: -

1. Preprocessing
2. Segmentation
3. Feature Extraction4. Classification
5. Post processing

1.2.1 Preprocessing

Preprocessing is the first stage of any recognition system. The preprocessing is a collection of operations applied on the digitized or raw image for smoothing, enhancing the image to make further steps of character recognition simple and accurate.[6]. Preprocessing includes noise reduction, skew removal, slant removal, and size normalization, skeletonization (thinning) etc. Noise removal removes the unwanted bits which do not have any impact on the output. Skewness is the tilt in the image that occurs during scanning, if the paper is not fed straight into the scanner. Size normalization is done to make all the characters to a uniform size so that algorithms can be applied easily. Skeletonization removes the width of the image from much pixel width to a single pixel width. The preprocessed image is used as input in further stages after removing the above mentioned imperfections.

1.2.2 Segmentation

Segmentation is very important stage of any recognition system. Segmentation includes separation of text into text lines, text lines into words and the separation of characters from their neighbors within the word.[7]. So, image consisting of sequence of characters is separated into individual symbols before recognition. Handwritten text has lot of problems like touching of characters which results in improper segmentation. Errors in segmentation can reduce the recognition rate. So efforts should be made to develop good segmentation techniques.

1.2.3 Feature Extraction

Feature extraction is done before recognition of any character. Features distinguish one class of characters

from the others in a meaningful way. Therefore it is very important to select the meaningful features. Feature extraction is the process of extracting the useful information from the text that can be used for recognition purpose. Recognition accuracy of any OCR system directly depends upon the accuracy of feature extraction[8][9]. As handwritten characters vary largely in size and slant, so efforts should be made to select the size and slant invariant features. Statistical and structural features are most commonly used features for character recognition. Selection of type of features and their extraction from the characters is very crucial step.

1.2.4 Classification

The classification stage is the main decision making stage of any OCR system. It classifies unknown character into different classes on the basis of features extracted. A class is a feature space or region in which the particular character falls. The different approaches which are used to classify the characters are pixel based, structural, statistical and neural network based.

1.2.5 Post Processing

The output of classification process goes through an error detection and correction phase. Post processing includes dictionary look up and apply language specific information on the unrecognized words. Some of the characters that cannot be segmented properly in a word can be recognized during post processing and word as a whole can be interpreted. It is the main stage for correction of segmentation and classification errors.

1.3 Significance of Study

This investigation is an attempt to study the existing segmentation and feature extraction techniques and to explore the possibility to improve the recognition rate of handwritten Hindi text. This will help in the development of Hindi OCR. The Hindi OCR can be used for the reading of addresses on envelops and reading of ancient documents. This can also be used for the recognition of hand filled forms.

1.4 Future enhancement:

1. A detailed literature survey about segmentation, feature extraction and classification has to be done.
2. New algorithms to be proposed for line segmentation, half character segmentation, segmentation of touching modifier from consonant in middle region and lower modifier segmentation of printed Hindi text.
3. A new feature set has to be developed for recognition of printed Hindi characters. Topological features (shape based) have been used for feature extraction. Efforts are made to select size and shape independent features.
4. Different classifiers like SVM and rule based classifiers to be been considered for classification and recognition.

References:

[1]S.Khedekar, V.Ramanaprasad, S.Setlur, and V.Govindaraju, “Text 4 Image Separation in DevanāgarīDocuments”, Proc.Seventh International Conference on Document Analysis and Recognition, 2003, pp.126541269. [2]R.Bajaj,L.Dey,andS.Chaudhury, “Devnagaricharacterrecognitionbycombiningde cisionofmultipleconnectionist classifiers”,Sadhana , 27(1), 2002, pp. 59–72.

[3]S.AntananiandL.Agnihotri, “Gujarati characterrecognition”,Proc.Fifth International Conference on Document Analysis and Recognition , 1999, pp. 418–421.

[4]V.BansalandR.M.K.Sinha,“ADevanāgarīOCRan da brief review of OCR research for Indian scripts”

[5]P. Chaudhuri and U. Pal, “An OCR system to readtwo Indian language scripts: Bangla and Devanāgarī”,Proc. Fourth IEEE International Conference on Document Analysis and Recognition , 1997, pp. 1011–1015.

[6]Y. Suganuma, “Learning structures of visual patterns from single instances”,
Artificial Intelligence, 50(1),1991,pp.1–36. [7]B.B. Chaudhuri andU. Pal, “Skew angledetectionofdigitizedIndian Scriptdocuments”,IEEE Transactions
on Pattern Analysis and Machine Intelligence, 19(2),1997,pp.182–186.
[8]M.Hanmandlu,K.R.M.Mohan,S.Chakraborty, S.Goyal and D. Roy Choudhury, “Unconstrainedhandwritten character recognition based on fuzzylogic”,Pattern Recognition ,36(3),2003,pp.6034623.
[9]M. Hanmandlu, M.H.M. Yusof. And V.K.Madasu,“Off4linesignature verification and forgery detection using fuzzy modeling”,Pattern Recognition ,38(3),2005,pp.3414356