Survey on Character Recognition using OCR Techniques

Karishma Tyagi, Vedant Rastogi
Department of Computer Science & Engineering, IET
Alwar, Rajsthan, INDIA

***Abstract*- Optical Character Recognition has number of applications in day-to-day life. OCR has been widely used in banking, legal, health care, finance etc. Handwriting recognition has been one of the most interesting and challenging research areas in field of image processing and pattern recognition in the recent years. This paper describes the techniques for converting textual content from a paper document into machine readable form. The computer actually recognizes the characters in the document through a transforming technique called Optical Character Recognition. In this paper, Several techniques like OCR using correlation method and OCR using neural networks has been discussed.**

Keywords*- *OCR, segmentation, neural network, character recognition, hidden markov.

I. INTRODUCTION

Highlight in 1950's [1], applied throughout the spectrum of industries resulting into revolutionizing the document management process. Optical Character Recognition or OCR has enabled scanned documents to become more than just image files, turning into fully searchable documents with text content recognized by computers. Optical Character Recognition extracts the relevant information and automatically enters it into electronic database instead of the conventional way of manually retyping the text. Optical Character Recognition is a process by which we convert printed document or scanned page to ASCII character that a computer can recognize.[3] The document image itself can be either machine printed or handwritten, or the combination of two.

OCR has three processing steps, Document scanning process, Recognition process and Verifying process. In the document scanning step, a scanner is used to scan the handwritten or printed documents. The quality of the scanned document depends up on the scanner. So, a scanner with high speed and color quality is desirable. The recognizing process includes several complex algorithms and previously loaded templates and dictionary which are crosschecked with the characters in the document and the corresponding machine editable ASCII characters. The verifying is done either randomly or chronologically by human Intervention. Difference in font and sizes makes recognition task difficult if preprocessing, feature extraction and recognition are not robust. There may be noise pixels that are introduced due to scanning of the image. Besides, same font and size may also have bold face character as well as normal one. Thus, width of the stroke is also a factor that affects recognition. Therefore, a good character recognition approach must eliminate the noise after reading binary image data, smooth the image for better recognition, extract features efficiently, train the system and classify patterns.

Segmentation of a document into lines and words and of words into individual characters and symbols constitute an important task in the optical reading of texts. Presently, most recognition errors are due to character segmentation errors. Very often, adjacent characters are touching, and may exist in an overlapped. Therefore, it is a complex task to segment a given word correctly into its character components. The process of handwriting recognition involves extraction of some defined characteristics called features to classify an unknown handwritten character into one of the known classes. A typical handwriting recognition system consists of several steps, namely: preprocessing, segmentation, feature extraction, and classification, several types of decision methods, including statistical methods, neural networks, structural matching (on trees, chains, etc). The stochastic processing (Markov chains, etc.) have been used along with different types of features [5]. The advantage of HMM approach over ANN approach in optical character recognition is that it can be easily extendible to the recognition of handwritten characters.

In this paper, we will discuss how artificial neural network , genetic algorithm and fuzzy logic can be used in optical character recognition for the use of character recognition.

The remaining part of this survey paper is organized as follows:-In section II, we will discuss the literature review in the field of character recognition and in section III we describe the various techniques used for character recognition using OCR, the comparative study of techniques discussed in section III given in section IV and in section V, we will conclude the paper and give the future scope of this paper.

II. LITERATURE REVIEW

A brief description of the history of OCR is as follows. In 1929 Gustav Tauschek obtained a patent on OCR in Germany, followed by Handel who obtained a US patent on OCR in USA in 1933. In 1935 Tauschek was also granted a US patent on his method. Tauschek's machine was a mechanical device that used templates and a photo detector.

RCA engineers in 1949 worked on the first primitive computer-type OCR to help blind people for the US Veterans Administration, but instead of converting the printed characters to machine language, their device converted it to machine language and then spoke the letters. It proved far too expensive and was not pursued after testing [2] [3].

In 1978 Kurzweil Computer Products began selling a commercial version of the optical character recognition computer program. LexisNexis was one of the first customers, and bought the program to upload paper legal and news documents onto its nascent online databases. In about 1965 Reader's Digest and RCA collaborated to build an OCR Document reader designed to digitize the serial numbers on Reader's Digest coupons returned from advertisements. Two years later, Kurzweil sold his company to Xerox, which had an interest in further commercializing paper-to-computer text conversion. Kurzweil Computer Products became a subsidiary of Xerox known as Scansoft, now Nuance Communications.

Khankasikam et al [8], [9] proposed the valley sharpening techniques which restricts the histogram to the pixels with large absolute values of derivatives where as S. Wantanable et. Al [10], [11] proposed the histogram difference method, which selects threshold at the gray level with the maximum amount of difference. This method uses the information concerning neighboring pixels or edges in the original picture to modify the histogram so as to make it useful for thresholding. Another method includes directly dealing with the grey level histogram by parametric techniques. By a sum of Gaussian distributions histogram is approximated in the least sense square and a statistical decision procedures are applied [10]. These methods are tedious and they involve high computational power.

In Di gesu [12], [10] the idea of using both intensities and spatial information has been considered to take into account local information used in human perception. A number of new methodologies and strategies have been proposed over the past few years to find global as well as local solutions in nonlinear multimodal function optimization [12], [13], [14]. In addition attempts have also been made to use Fuzzy Logic [15], Artificial Neural Network [16] for optical character recognition. With the help of crowding multiple peaks can be maintained in multimodal optimization problem. Crowding method is extremely reliable in detecting the peaks on bimodal histogram.

This paper makes use of an OCR system which makes use of histogram equalization to extract images. The histogram used by the mentioned algorithm is bimodal in nature hence it can be divided into two classes [17]. Genetic algorithm is further used to select the threshold from the histogram for extracting the object from the background.

III. CHARACTER RECOGNITION TECHNIQUES

Optical Character Recognition can be applied to recognize text from any multimedia such as image, audio, video. Automatic multimedia recognition is based on the computer vision and pattern recognition application [16]. We can use image processing, character positioning, character segmentation, neural network to solve the problem of image to text recognition.

1. HMM Approach

A hidden Markov model is a doubly stochastic process, with an underlying stochastic process that is not observable (hence the word hidden), but can be observed through another stochastic process that produces the sequence of observations [11],[14],[16][17]. The hidden process consists of a set of states connected to each other by transitions with probabilities, while the observed process consists of a set of outputs or observations, each of which may be emitted by each state according to some output probability density function (PDF) [9][12]. Depending on the nature of this PDF function several kinds of HMMs can be distinguished.

2. Neural network approach

Character Recognition in the license plate recognition has important role in optical recognition system which is related directly with sucess or failure of the system. We used Back Propogation Neural Network to optically recognise the image. The basic idea of BP algorithm is the learning process is divided into two phases :

PHASE I: Forward Propogation

PHASE II: Back Propogation

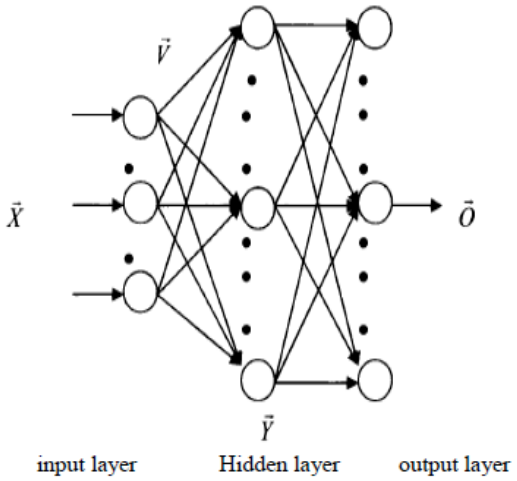

Figure 3.1: Character recognition using neural network

3. Character Normalization

It is necessary to normalize the character, letters and numbers to standard size. We can normalize the characters of different size into one fixed size to make our task easy for optical character recognition

Figure 3.2: Normalization

4. Correlation method for single character recognition

Preprocessing: The image is taken and is converted to gray scale image. The gray scale image is then converted to binary image. This process is called Digitization of image. Practically any scanner is not perfect; the scanned image may have some noise. This noise may be due to some unnecessary details present in the image. So, all the objects having pixel values less than 30 are removed. The de-noised image thus obtained is saved for further processing. Now, all the templates of the alphabets that are pre-designed are loaded into the system.

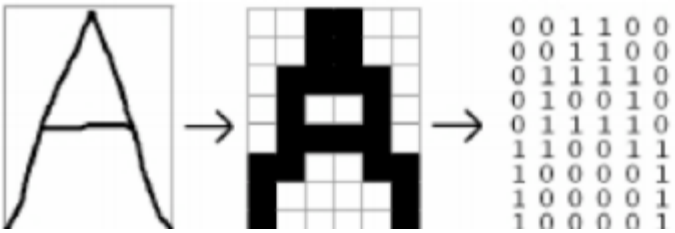

Figure 3.3: Digitized Image

Segmentation: In segmentation , the position of the object i.e., the character in the image is found out and the size of the image is cropped to that of the template size.

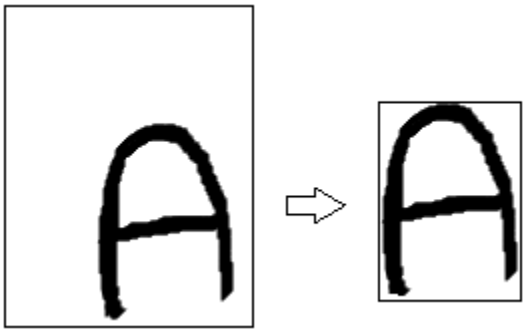

Figure 3.4: Segmented Image

Recognition: The image from the segmented stage is correlated with all the templates which are preloaded into the system. Once the correlation is completed, the template with the maximum correlated value is declared as the character present in the image.

IV. PROBLEM FORMULATION

The overall architecture of the OCR consists of three main phases- Image Preprocessing, Segmentation, Recognition and Post-processing. We explain each of these phases below.

a. *Image Preprocessing*

Pre-processing methods use a small neighborhood of a pixel in an input image to get a new brightness value in the output image. Such pre-processing operations are also called filtration.

- Image smoothing is the set of local pre-processing methods which have the aim of suppressing image noise - it uses redundancy in the image data.
- Calculation of the new value is based on averaging of brightness values in some neighborhood O.
- Smoothing poses the problem of blurring sharp edges in the image

b. *Segmentation*

Segmentation in the context of character recognition can be defined as the process of extracting from the preprocessed image the smallest possible character units which are suitable for recognition. It consist of the following steps :

- Locate the Header Line

An image is stored in the form of a two dimensional array in computer. A black pixel is represented by 1 and a white pixel by a 0. The array is scanned row by row and the number of black pixels is recorded for each row resulting in horizontal histogram. The row with the maximum number of black pixels is the position of the header line called as Shirorekha. This position is identified as*hLinePos.*

- Separate the Character boxes

Characters are present below the header line. To identify the character boxes, we make a vertical histogram of the image starting from the *hLinePos* to boundary of the word i.e. the row where there are no black pixels. The boundaries for characters are identified as the columns that have no black pixels.

- Separate the upper modifier symbols to identify the upper modifier symbols; we make a vertical histogram of the image starting from the top row of the image to *hLinePos.*
- Separate the lower modifiers we did not attempt lower modifier separation due to lack of time.

c. *Feature Extraction*

Feature extraction refers to the process of characterizing the images generated from the segmentation procedure based on certain specific parameters. We did not explore this further.

d. *Classification*

Classification involves labeling each of the symbols as one of the known characters, based on the characteristics of those symbols. Thus, each character image is mapped to a textual representation.

e. Post-processing

The output of the classification process goes through an error detection and correction phase.

In correlation method there are many unnecessary comparisons and the efficiency of recognition is same for a particular pattern and the given set of templates. However extra templates can be added to the system for providing a wide range of compatibility but doing so will increase the computational intensity of the system. Another important drawback of this method is it requires lot of memory and execution time.

V. CONCLUSION AND FUTURE SCOPE

A number of techniques that are used for optical character recognition have been discussed which uses correlation and neural networks. Much other advancement in Optical Character Recognition are being under development. The paper presents a brief survey of the applications in various fields along with experimentation into few selected fields. The proposed method is extremely efficient to extract all kinds of bimodal images including blur and illumination. The paper will act as a good literature survey for researchers starting to work in the field of optical character recognition. The reason of its complexities are its characters shapes, its top bars and end bars more over it has some modified, vowel and compound characters and also one of the important reasons for poor recognition in OCR system is the error in character recognition.

REFERENCES

[1]. U. Garain and B. B. Chaudhuri,"Segmentation of Touching Characters in Printed Devnagari and Bangla Scripts Using Fuzzy Multifactorial Analysis"2001

[2]. Zadeh, L.A. (1965). "Fuzzy sets", Information and Control 8 (3): 338– 353.

[3]. H2M: A set of MATLAB/OCTAVE functions for the EM estimation of mixture and hidden morkov model by Olivier Cappe ENST. Dpt. TSI/LTCI (CNRS-URA 820), France, August 24, 2001.

[4]. C. Rafael Gonzalez and E. Richard Woods, Digital Image Processing, Addison-Wesley Publishing Company.

[5]. A. Magdi Mohamed and Paul Gader, Generalized hidden markov models-part II: application to handwritten word recognition, On fuzzy system, IEEE Trans., 8(1) (February 2000). S. K. Parui and B. Shaw, Offline handwritten Devanagari word recognition: an hmm based approach, Proc. PReMI-2007(Springer), LNCS-4815:528-535, December 2007.

[6]. L.R. Rabiner, A tutorial on hidden markov models and selected application in speech recognition, Proceedings of the IEEE, 77(2) (February 1989), 257-286.

[7]. M. Christopher Bishop, Pattern recognition and machine learning, Information Science and Statistic Series, Springer, 423-455.

[8]. Jia Zeng and Zhi-Qiang Liu, Markov random field-based statistical character structure modeling for handwritten Chinese character recognition, On pattern analysis and machine intelligence, IEEE Trans., 30(5), May(2008).

[9]. Imtiaz, H. and Fattah, S. A. (2011) "A wavelet-domain local dominant feature select ion scheme for face recognit ion," Internat ional Journal of Computing and Business Research, Vol.3, Issue.2.

[10]. Khurana, P. and Singh, V. (2011) "A model for human cognition," International Journal of Computing and Business Research, Vol.2, Issue.3.

[11]. Kurian, C. and Balakriahnan, K. (2012) "Cont inuous speech recognition system for Malayalam language using PLP cepstral coefficient," Journal of Computing and Business Research, Vol.3, Issue.1.

[12]. Sivanandam, S.N and Deepa, S.N. (2011) "Principles of Soft Comput ing," Wiley-India publisher, 2nd edition.

[13]. S. Mori, C.Y. Suen and K. Kamamoto, "Historical review of OCR research and development," Proc. Of IEEE, vol. 80, pp. 1029-1058, July 1992.

[14]. S. Impedovo, L. Ottaviano and S.Occhinegro, "Optical character recognition", International Journal Pattern Recognition and Artificial Intelligence, Vol. 5(1-2), pp. 1- 24, 1991.

[15]. R. Plamondon and S. N. Srihari, "On-line and off- line handwritten character recognition: A comprehensive survey,"IEEE. Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 63-84, 2000.

[16]. N. Arica and F. Yarman-Vural, "An Overview of Character Recognition Focused on Off-line Handwriting", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2001, 31(2), pp. 216 – 233.