International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Survey on Optical Character Recognition Techniques

Sarika Pansare[1], Dhanshree Joshi[2]

[1]Sinhgad Academy of Engineering, Computer Engineering Department, Pune University, Maharashtra, India

[2]Professor, Sinhgad Academy of Engineering, Computer Engineering Department, Pune University, Maharashtra, India

Abstract: ***This paper presents a review on Optical Character Recognition Techniques. Optical Character recognition (OCR) is a technology that allows machines to automatically recognize the characters through an optical mechanism. OCR can be described as Mechanical or electronic conversion of scanned images where images can be handwritten, typewritten or printed text. It converts the images into machine-encoded text that can be used in machine translation, text-to-speech and text mining. Various techniques are available for character recognition in optical character recognition system. This material can be useful for the researchers who wish to work in character recognition area.***

Keywords: Character Recognition, machine-encoded, OCR, Translation, Text Mining

1. Introduction

Optical Character recognition has been a subject of research. Optical Character Recognition (OCR) systems is transforming large amount of documents, either printed alphabet or handwritten into machine encoded text without any transformation, noise, resolution variations and other factors[8].

In general, handwriting recognition is classified into two types as off-line and On-line character recognition. Off-line handwriting recognition process automatically converts the text into an image into letter codes which are usable within computer and text-processing applications. Because of different handwriting styles of persons off-line handwriting recognition is more complex and difficult. In the On-line character recognition it deals with a data stream which comes from a transducer while the user is writing. Suitable hardware is used to collect data is a digitizing tablet which is electromagnetic or pressure sensitive. Optical character recognition is needed when the information should be readable both to humans and to a machine and alternative inputs can not be predefined.

OCR consists of many phases such as Pre-processing, Segmentation, Feature Extraction, Classifications and Recognition. The task of preprocessing relates to the removal of noise and variation in handwritten. Several area where OCR used including mail different bank processing , reading document and postal address recognition require offline handwriting recognition systems, pattern recognition.

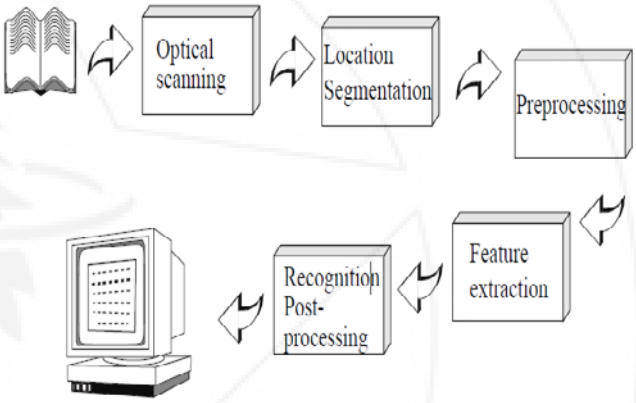

Figure 1: Components Of OCR System

1.1 Approaches used for the design of OCR systems

Matrix Matching: Matrix Matching converts each character into a pattern within a matrix, and then compares the pattern with an index of known characters. Its recognition is strongest on monotype and uniform single column pages.

Fuzzy Logic: Fuzzy logic is a multi-valued logic that allows intermediate values to be defined between conventional evaluations like yes/no, true/false, black/ white etc. An attempt is made to attribute a more human-like way of logical thinking in the programming of computers. Fuzzy logic is used when answers do not have a distinct true or false value and there is uncertainly involved.

Feature Extraction: This method defines each character by the presence or absence of key features, including height, width, density, loops, lines, stems and other character traits. Feature extraction is a perfect approach for OCR of magazines, laser print and high quality images.

Structural Analysis: Structural Analysis identifies characters by examining their sub features shape of the image, sub-vertical and horizontal histograms. Its character repair capability is great for low quality text and newsprints.

Neural Networks: This strategy simulates the way the human neural system works. It samples the pixels in each image and matches them to a known index of character pixel patterns. The ability to recognize characters through abstraction is great for faxed documents and damaged text. Neural networks are ideal for specific types of problems, such as processing stock market data or finding trends in graphical patterns.

2. Motivation for Survey

In past few years it has become necessary to digitize historic paper based documents like books, newspapers, information material of any field . The purpose behind this is to preserve the document, make them fully and easily accessible, searchable, process able in digital form.

The first step towards transforming paper based information into a digital information is to scan the documents. The next step is to apply an OCR (Optical Character Recognition) process, in this process scanned image of each document will be translated into machine process able text. After finishing the OCR process several post-processing steps are applied on image depending on the application. A post-processing step to correct these errors is a very important part of the post-processing chain. A semi-automatic post-correction system detects errors automatically and proposes corrections to human correctors who then have to choose the correct proposal. A fully-automatic post-correction system does the detection and correction of errors by its own. Thus OCR plays very important role in this type of applications and there are different OCR techniques are available to perform character recognition.

3. Literature Survey

This material serves as a guide and update for readers working in the Character Recognition area.Gur et al. (2012) [2] has discussed that text recognition and retrieval problem. Automated optical character recognition (OCR) tools lacks in supplying a complete solution for the problem and in most of cases human interaction is required. In this paper the authors suggest a novel text recognition algorithm based on usage of fuzzy logic rules relying on statistical data of the analyzed font. This new approach combines letter statistics and correlation coefficients in a set of fuzzy based rules which enables the recognition of distorted letters that may not be retrieved otherwise. The authors have focused on Rashi fonts associated with commentaries of the Bible that are actually handwritten calligraphy.

Chirag I Patel et al. [2011][3] Objective in this paper is recognize the characters in a given scanned documents and study the effects of changing the Models of Artificial Neural Network. Now a days Neural Networks are mostly used for Pattern Recognition task. This paper describes the behaviors of different Models of Neural Network used in OCR. Optical Character Recognition is widespread use of Neural Network. They have considered parameters like number of Hidden Layer, size of such hidden Layer and epochs. Multilayer Feed Forward network is used with Back propagation. In Preprocessing they have applied some basic algorithms for segmentation of characters, normalizing of such characters and De-skewing. Different Models of Neural Network are used and applied the test set on each to find the accuracy of the respective Neural Network.Vijay Laxmi Sahu et al. (January 2013)[4] This paper explains that classification methods based on learning from examples have been widely applied to character recognition from the 1990s and have brought forth significant improvements of recognition accuracies. The class of methods includes different methods like statistical methods,ANN(artificial neural networks), SVM (support vector machines), multiple classifier combination. In this paper, the characteristics of the classification methods that have been successfully applied to character recognition and remaining problems that can be potentially solved by learning methods have been discussed.

Majida Ali Abed et al. (August 2013)[5]This paper concentrate on new approach to simplify Handwritten Characters Recognition based on simulation of the behavior of schools of fish and flocks of birds that is called the Particle Swarm Optimization Approach (PSOA). This Paper presents an overview of the proposed approaches to be optimized and tested on a number of handwritten characters in the experiments. In this paper experimental results demonstrate the higher degree of performance of the proposed approaches. PSOA generally generates an optimized comparison between the input samples and database samples which improves the final recognition rate. Experimental results given in paper show that, the PSOA is convergent and more accurate in solutions that minimize the error recognition rate. Jawahar et al. (2012) [6] has proposed a recognition scheme for the Indian script of Devanagari. Character recognition accuracy of Devanagari script is not yet comparable to its Roman counterparts. This is because of the complexity of the script, different writing style. Proposed solution uses a Recurrent Neural Network known as Bidirectional Long- Short Term Memory (BLSTM). This approach does not require word to character segmentation, which is one common reason for high word error rate. Results of the paper has reported a reduction of more than 20% in word error rate and over 9% reduction in character error rate while comparing with the best available OCR system.

Dileep Kumar Patel et al. [2012] [9]In this paper, the problem of handwritten character recognition has been solved with multiresolution technique using Discrete wavelet transform (DWT) and Euclidean distance metric (EDM). The proposed technique has been tested and found to be more accurate and faster than existing one. Characters is classified into 26 pattern classes based on appropriate properties. Features are extracted from the handwritten character images using DWT with appropriate level of multi-resolution technique, and then each pattern class is characterized by a mean vector. As shown in this paper Distances from input pattern vector to all other mean vectors are computed by using Euclidean Distance Metric. The Minimum distance determines the class membership of input pattern vector. The pro- posed method provides good recognition accuracy of 90% for handwritten characters even with fewer samples.

Shalin A. Chopra et al(2014)[7]This paper presents a simple, efficient, and low cost approach to construct OCR for reading any document that has fix font size and handwritting style. Optical Character Recognition in this paper uses database to recognize English characters which makes this OCR very simple to manage which helps to achieve efficiency and less computational cost. The feature extraction step of optical character recognition is the most important. It can be used with other existing OCR methods for the purpose of English text recognition. This system offers an upper edge by having an advantage i.e. its scalability, i.e. although it is configured to read a predefined set of document formats, as proposed in this paper for English documents, it can be configured to recognize new types.

4. Conclusion

This is detailed discussion about optical character recognition techniques and includes its use in different area of character recognition by using OCR. From study of various papers we have seen that selection of relevant technique plays an important role in performance of character recognition rate. This material serves as a helpful guide and update for readers working in the Character Recognition area.

References

[1] Faisal Mohammad, Jyoti Anarase, Milan Shingote, Pratik Ghanwat,"Optical Character Recognition Implementation Using Pattern Matching "(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2088-2090
[2] Gur, Eran, and ZeevZelavsky. "Retrieval of Rashi Semi-Cursive Handwriting via Fuzzy Logic." Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on. IEEE, 2012.
[3] Chirag I Patel, Ripal Patel, Palak Patel Handwritten Character Recognition using Neural Network International Journal of Scientific & Engineering Research Volume 2, Issue 5, May-2011
[4] Vijay Laxmi Sahu, Babita Kubde "Offline Handwritten Character Recognition Techniques using Neural Network: A Review"International Journal of Science and Research (IJSR), India Online ISSN: 2319-7064 Volume 2 Issue 1, January 2013
[5] Majida Ali Abed, Hamid Ali Abed Alasadi "Simplifying Handwritten Characters Recognition Using a Particle Swarm Optimization Approach" European Academic Research, Vol. I, Issue 5/ August 2013 pp-532-552
[6] Sankaran, Naveen, and C. V. Jawahar. "Recognition of printed Devanagari text using BLSTM Neural Network." Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012.
[7] Shalin A. Chopra, Amit A. Ghadge, Onkar A. Padwal, Karan S. Punjabi, Prof. Gandhali S. Gurjar," Optical Character Recognition" International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 1, January 2014
[8] OCR-Optical Character Recognition by Line Eikvil December 1993
[9] Dileep Kumar Patel, Tanmoy Som1, Sushil Kumar Yadav, Manoj Kumar Singh," Handwritten Character Recognition Using Multiresolution Technique and Euclidean Distance Metric" JSIP 2012, 208-214
[10] V. Govindaraju and S. Setlur. Guide to OCR for Indic Scripts. 2009.

Author Profile

Sarika V. Pansare received the B.E degree in Computer Engineering from Bharti Vidyapith,s College of Engineering For Women Pune ,Maharashtra ,India in 2009 .Presently pursuing M.E degree in Computer Engineering from Sinhgad Academy of Engg., Pune, Maharashtra ,India.

Prof. D. K. Joshi received the M.E degree in Computer Engineering from D Y Patil College of Engg. Pune, Maharashtra ,India..She is associated currently with the computer engineering department of Sinhgad Academy of Engineering Kondhwa, Pune, Maharashtra, India.