IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 2, Ver. II (Mar – Apr. 2015), PP 22-26
www.iosrjournals.org

Optical Character Recognition (OCR) System

Najib Ali Mohamed Isheawy And Habibul Hasan

Abstract: *In the running world, there is growing demand for the software systems to recognize characters in computer system when information is scanned through paper documents as we know that we have number of newspapers and books which are in printed format related to different subjects. These days there is a huge demand in "storing the information available in these paper documents in to a computer storage disk and then later reusing this information by searching process". One simple way to store information in these paper documents in to computer system is to first scan the documents and then store them as IMAGES. But to reuse this information it is very difficult to read the individual contents and searching the contents form these documents line-by-line and word-by-word. The reason for this difficulty is the font characteristics of the characters in paper documents are different to font of the characters in computer system. As a result, computer is unable to recognize the characters while reading them. This concept of storing the contents of paper documents in computer storage place and then reading and searching the content is called DOCUMENT PROCESSING. Sometimes in this document processing we need to process the information that is related to languages other than the English in the world. For this document processing we need a software system called CHARCATER RECOGNITION SYSTEM. This process is also called DOCUMENT IMAGE ANALYSIS (DIA).*

Keywords: *Optical Character Recognition, Neural Network, Fuzzy Logic*

I. Introduction

Optical character recognition refers to the branch of computer science that involves reading text from paper and translating the images into a form that the computer can manipulate (for example, into ASCII codes). An OCR system enables you to take a book or a magazine article, feed it directly into an electronic computer file, and then edit the file using a word processor.All OCR systems include an optical scanner for reading text, and sophisticated software for analyzing images. Most OCR systems use a combination of hardware (specialized circuit boards) and software to recognize characters, although some inexpensive systems do it entirely through software. Advanced OCR systems can read text in large variety of fonts, but they still have difficulty with handwritten text.It is the mechanical or electronic translation of scanned images of handwritten, typewritten or printed text into machine-encoded text. It is widely used to convert books and documents into electronic files, to computerize a record-keeping system in an office, or to publish the text on a website.

OCR makes it possible to edit the text, search for a word or phrase, store it more compactly, display or print a copy free of scanning artifacts, and apply techniques such as machine translation, text-to-speech and text mining to it. OCR is a field of research in pattern recognition, artificial intelligence and computer vision.

OCR systems require calibration to read a specific font; early versions needed to be programmed with images of each character, and worked on one font at a time. "Intelligent" systems with a high degree of recognition accuracy for most fonts are now common.Some systems are capable of reproducing formatted output that closely approximates the original scanned page including images, columns and other non-textual components.

Optical character recognition (OCR) is one of the most popular areas of research in pattern recognition because of its immense application potential. However, most of the available methods deal with the recognition of the Roman script and some of the oriental scripts like Kanji, Kana, etc

The purpose of this OCR system is to take handwritten English characters as input, process the character, train the neural network algorithm, to recognize the pattern and modify the character to a beautified version of the input.This work is restricted to English characters and numerals only. It is also helpful in recognizing special characters. It can be further developed to recognize the characters of different languages. One of the primary means by which computers are endowed with humanlike abilities is through the use of a neural network. Neural networks are particularly useful for solving problems that cannot be expressed as a series of steps, such as recognizing patterns, classifying them into groups, series prediction and data mining.

II. Background

Offline and Online OCR

OCR can be implemented both off-line and on-line. In the off-line recognition, the writing is usually captured optically by a scanner and the completed writing is available as an image. But, in the on-line system the two dimensional coordinates of successive points are represented as a function of time. And the orders of

strokes made by the writer are also available. The on-line methods have been shown to be superior to their off-line counterparts in recognizing handwritten characters due to the temporal information available with the former.The input for the OCR problem is pages of scanned text. To perform the character recognition, our application has to go through three important steps.

1-Segmentation: Given input image, identify individual glyphs (basic units representing one or more characters, usually contiguous).

2-Feature Extraction: From each glyph image, extract features to be used as input of ANN. This is the most critical part of this approach.

3-Classification: Train the ANN using training sample. Then given new glyph, classify it.

The second step is the most difficult in the sense that there is no obvious way to obtain these features.

To gain better knowledge, techniques and solutions regarding the procedures that we want to follow, we studied the various re-search papers on existing OCR systems. All these study helped us with clarifying our target goals.

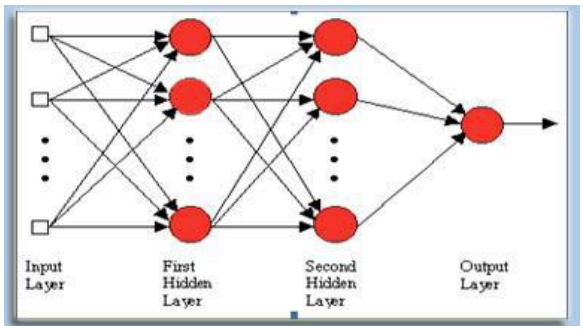
The basic steps involved in Optical Character Recognition are:-

1. Image Acquisition
2. Preprocessing
3. Document Page Analysis
4. Feature Extraction
5. Training and Recognition
6. Post Processing

Neural Networks

Pattern recognition is perhaps the most common use of neural networks. The neural network is presented with a target vector and also a vector which contains the pattern information, this could be an image and hand written data. The neural network then attempts to determine if the input data matches a pattern that the neural network has memorized.

A neural network trained for classification is designed to take input samples and classify them into groups. These groups may be fuzzy, without clearly defined boundaries. This project concerns detecting free handwritten characters.



III. Related Work

A number of researches have been proposed over the years for character recognition.A noble idea was conceptualized way back in 1999 by (Sural and Das). The ability of a software or application like this one needs the ability to apply pattern recognition, pattern interpretation and learning. This sought the use of multi-layer perception model which is a feed forward model. They used the model to develop an OCR(Optical Character Recognition) for an Indian Language named Bengali. The model that was used belong to a class of artificial neural network. The advantage of using perception model is its efficiency of learning. It can be used to train and recognize any kind of data set than initially defined. Moreover, their approach left a great scope of developing OCR that target any language other than Bengali.Another work comes from DeepayanSarkar from University of Wisconsin. He implemented OCR as an add on package for a MATLAB-like programming environment called R.

Although results were not that good, they were not bad either suggesting this technique is not flawed.More training data would improve robustness and accuracy. Moreover, he pointed out that the speed may need to be improved using C code.The authors have divided each character into a number of predetermined rectangular zones and extracted a 13-element vector comprising of the pixel values in those zones. A neural network classifier has been used to recognize the 26 alphabets of English language.

IV. Proposed System

Our proposed OCR system is based on grid infrastructure, which is a character recognition system that supports recognition of the characters of multiple languages. This feature is what we call grid infrastructure which eliminates the problem of heterogeneous character recognition and supports multiple functionalities to be performed on the document. The multiple functionalities include editing and searching too where as the existing system supports only editing of the document. In this context, Grid infrastructure means the infrastructure that supports group of specific set of languages. Thus OCR on a grid infrastructure is multi-lingual.

The main purpose of Optical Character Recognition (OCR) system based on a grid infrastructure is to perform Document Image Analysis, document processing of electronic document formats converted from paper formats more effectively and efficiently. This improves the accuracy of recognizing the characters during document processing compared to various existing available character recognition methods. Here OCR technique derives the meaning of the characters, their font properties from their bit-mapped images.

The primary objective is to speed up the process of character recognition in document processing. As a result the system can process huge number of documents with-in less time and hence saves the time.Since our character recognition is based on a grid infrastructure, it aims to recognize multiple heterogeneous characters that belong to different universal languages with different font properties and alignments.

Module Description:

Our software system Optical Character Recognition on a grid infrastructure can be divided into five modules based on its functionality.The modules classified are as follows:-

- Document Processing Module
- System Training Module.
- Document Recognition Module.
- Document Editing Module and
- Document Searching Module.

Document Processing Module

- Scanning printed documents.
- Storing the documents as snapshots or images.
- Processing those image-based documents.
- Converting these image-based documents into e-documents(also called structured documents).
- Recognizing the characters in documents.
- Generating grid infrastructure datastructure.

System Training Module

- Training the system with the pre-defined fonts.
- Training the system with the new fonts that are not present in the system and that cannot be identified by the system.

Document Recognition Module

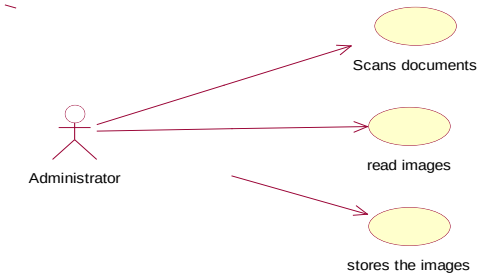
- Converts the document into specific format
- Recognizes the characters
- Heterogeneous character Recognition

Document Editing Module

- Addition of specific content to the documents
- Deletion of certain content from documents
- Any other modification of documents.

Document Searching Module

This module can be accessed by both the administrator and the end-user during the search of the user required document to implement the character recogniiton process on it. The user requests the system to search for a particular document. Then the system finds the documents based on OCR methodology and returns the result of the search to the user.



V. Results And Discussion

The main purpose of Optical Character Recognition (OCR) system based on a grid infrastructure is to perform Document Image Analysis, document processing of electronic document formats converted from paper formats more effectively and efficiently. This improves the accuracy of recognizing the characters during document processing compared to various existing available character recognition methods. Here OCR technique derives the meaning of the characters, their font properties from their bit-mapped images.

VI. Conclusion

The Grid infrastructure used in the implementation of Optical Character Recognition system can be efficiently used to speed up the translation of image based documents into structured documents that are currently easy to discover, search and process.

The network has been trained and tested for a number of widely used fonts.The recognition of new font characters by the system is very easy and quick. We can edit the information of the documents more conveniently and we can reuse the edited information as and when required.

References

[1]. ShamikSural,P.K.Das,Recognition of an Indian Script using Multilayer Perceptrons and Fuzzy Features Sixth International Conference on Document Analysis and Recognition (ICDAR2001), Seattle, 2001, pp. 1120-1124.
[2]. MamtaMaloo, Dr. K.V. Kale, Gujarati Script Recognition: A Review, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, July 2011 ISSN (Online): 1694-0814
[3]. Sujata S. Magare and Ratnadeep R. Deshmukh, Offline Handwritten Sanskrit Character Recognition Using Hough Transform and Euclidean Distance, International Journal of Innovation and Scientific Research ISSN 2351-8014 Vol. 10 No. 2 Oct. 2014, pp. 2-302
[4]. Rajiv Kapoor, AmitDhamija, A New Method for Identification of Partially Similar Indian Scripts, International Journal of Image Processing (IJIP), Volume (6) : Issue (2) : 2012
[5]. Swapnil A. Vaidya, Balaji R. Bombade, A Novel Approach of Handwritten Character Recognition using Positional Feature Extraction, IJCSMC, Vol. 2, Issue. 6, June 2013, pg.179 – 186
[6]. ImanYousif, Adnan Shaout, Off-Line Handwriting Arabic Text Recognition: A Survey, Volume 4, Issue 9, September 2014 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering
[7]. Aradhana A Malanker , Prof. Mitul M Patel , Handwritten Devanagari Script Recognition: A Survey, IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE) e-ISSN: 2278-1676,p-ISSN: 2320-3331, Volume 9, Issue 2 Ver. II (Mar – Apr. 2014), PP 80-87
[8]. PrachiMukherji, PrachiMukherji, Shape Feature and Fuzzy Logic Based Offline Devnagari Handwritten Optical Character Recognition, Journal of Pattern Recognition Research 4 (2009) 52-68
[9]. Devendra Singh Kaushal, Yunus Khan & Dr. SunitaVarma, Handwritten Urdu Character Recognition Using Zernike MI's Feature Extraction and Support Vector Machine Classifier, International Journal of Research (IJR) Vol-1, Issue-7, August 2014 ISSN 2348-6848
[10]. Pratibha Singh, Ajay Verma, Narendra S. Chaudhari, Classification of Hindi numeral using Fuzzy Zoning and SVM, Proc. of the International Conference on Advanced Computing and Communication Technologies (ACCT 2011), ISBN: 978-981-08-7932-7
[11]. SandhyaArora, DebotoshBhattacharjee, MitaNasipuri, Dipak Kumar Basu*, MahantapasKundu, Multiple Classifier Combination for Off-line Handwritten Devnagari Character Recognition, Multiple Classifier Combination for Off-line Handwritten Devnagari Character Recognition
[12]. Swati Mukherjee, recognition of handwritten bengali character based on character features
[13]. Gaurav Y. Tawde , Mrs. Jayashree M. Kundargi, An Overview of Feature Extraction Techniques in OCR for Indian Scripts Focused on Offline Handwriting, Vol. 3, Issue 1, January -February 2013, pp.919-926
[14]. SushamaShelke, ShailaApte, A Multistage Handwritten Marathi Compound Character Recognition Scheme using Neural Networks and Wavelet Features, International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 4, No. 1, Marc
[15]. ShamikSural, P.K.Das, An MLP using Hough transform based fuzzy feature extraction for Bengali script recognition

Optical character recognition is the process of recognizing optically scanned characters. Character recognition has two types: Offline and Online. One of the challenging problems in pattern recognition is Offline character recognition. Offline character recognition takes scanned image of required document paper. Offline character recognition can be done in two ways: Handwritten and Printed.Handwritten character recognition is abbreviated as HCR; handwritten characters have number of variations as different people have different writing styles.

HCR can recognize offline character and online characters. Offline HCR takes input from scanned image of paper document and Online HCR takes input from digital pen. There are many handwritten historical documents exist in electronic form, HCR is used to recognize such documents.