Text Recognition from Images

Mr. Pratik Madhukar Manwatkar
Department of Computer Technology,
YCCE, Nagpur (M.S.), 441 110, India.
pratikmm@ymail.com

Mr. Shashank H. Yadav
Department of Computer Technology,
YCCE, Nagpur (M.S.), 441 110, India.
Shashank.y89@gmail.com

***Abstract*— Text recognition in images is a research area which attempts to develop a computer system with the ability to automatically read the text from images. These days there is a huge demand in storing the information available in paper documents format in to a computer storage disk and then later reusing this information by searching process. One simple way to store information from 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 challenges involved in this the font characteristics of the characters in paper documents and quality of images. Due to these challenges, computer is unable to recognize the characters while reading them. Thus there is a need of character recognition mechanisms to perform Document Image Analysis (DIA) which transforms documents in paper format to electronic format. In this paper we have discuss method for text recognition from images. The objective of this paper is to recognition of text from image for better understanding of the reader by using particular sequence of different processing module.**

***Keywords*: Document Image Analysis (DIA), electronic format, text recognition, font characteristics.**

I. Introduction

Now-a-days, 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. Whenever we scan the documents through the scanner, the documents are stored as images format in the computer system. These images containing text cannot be edited by the user. But to reuse this information it is very difficult for computer system 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. This process is also called Document Image Analysis (DIA). Thus our need is to develop some text recognition algorithm to perform Document Image Analysis which transforms documents in paper format to electronic format.

The paper is organized as follows: in Section 2, we discuss the related work done in field of image to text recognition. Section 3, overview of text recognition system. Section 4, we discuss about experimental results of this system. Section 5, discusses about the applications of text recognition and Section 6, finally, conclusion is given.

II. Literature Review

As discussed earlier text recognition from images is still an active research in the field of pattern recognition. To address the issues related to text recognition many researchers have proposed different technologies, each approach or technology tries to address the issues in different why. In forthcoming section we present a detailed survey of approaches proposed to handle the issues related to text recognition.

Yang et al.[1] has proposed a novel adaptive binarization method based on wavelet filter is proposed. This approach was processes faster, so that it is more suitable for real-time processing and applicable for mobile devices. They evaluated this adaptive method on complex scene images of ICDAR 2005 database. Sankaran et al. [2] has proposed a novel recognition approach that result in a 15% decrease in word error rate on heavily degraded Indian language document images.

Gur et al. [3] has discussed some problems in text recognition and retrieval. Automated optical character recognition(OCR) tools do not supply a complete solution and in most cases human inspection is required. They suggest a novel text recognition algorithm based on usage of fuzzy logic

978-1-4799-6818-3/15/$31.00 © 2015 IEEE

rules relying on statistical data of the analyzed font. The new approach combines letter statistics and correlation coefficients in a set of fuzzy based rules, enabling the recognition of distorted letters that may not be retrieved otherwise. They focused on rashi fonts associated with commentaries of the bible that are actually handwritten calligraphy.

Rhead et al. [4] has considered real world UK number plates and relates these to ANPR. It considers aspects of the relevant legislation and standards when applying them to real world number plates. The varied manufacturing techniques and varied specifications of component parts are also noted. The varied fixing methodologies and fixing locations are discussed as well as the impact on image capture.

Badawy, W. et al. [5] has discussed the Automatic license plate recognition (ALPR) is the extraction of vehicle license plate information from an image or a sequence of images. The extracted information can be used with or without a database in many applications, such as electronic payment systems (toll payment, parking fee payment), and freeway and arterial monitoring systems for traffic surveillance. The ALPR uses either a color, black and white, or infrared camera to take images.

Jawahar et al. [6] has proposed a recognition scheme for the indian script of devanagari. They used approach does not require word to character segmentation, which is one of the most common reason for high word error rate. They have been 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.

Ntirogiannis et al. [7] has studied that the document image binarization is of great importance in the document image analysis and recognition pipeline since it affects further stages of the recognition process. The evaluation of a binarization method aids in studying its algorithmic behavior, as well as verifying its effectiveness, by providing qualitative and quantitative indication of its performance. They proposed a pixel-based binarization evaluation methodology for historical handwritten/machine-printed document images.

Malakar et al. [8] has described that extraction of text lines from document images is one of the important steps in the process of an Optical Character Recognition (OCR) system. In case of handwritten document images, presence of skewed, touching or overlapping text line(s) makes this process a real challenge to the researcher.

III. TEXT RECOGNITION SYSTEM RELATED WORK

In this section we describe the overall architecture of Text recognition system. A Text recognition system receives an input in the form of image which contains some text information. The output of this system is in electronic format i.e. text information in image are stored in computer readable form. Our text recognition system divided in following module:

A. Pre-processing Module
B. System Training Module.
C. Text Recognition Module
D. Post-processing Module

The overall architecture is depicted in figure 1.

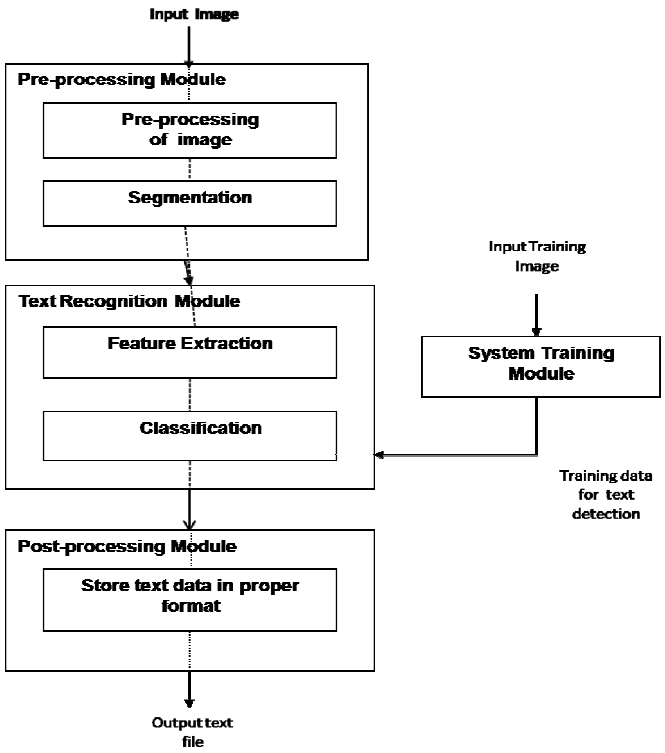

Fig.1: Architecture of text recognition.

A. Pre-processing Module

The Paper document is generally scanned by the optical scanner and is converted in to the form of a picture. A picture is the combinations of picture elements which are also known as pixels. The pixels contain basically two values ON and OFF. The ON value points that's the pixel is visible and the OFF value points that's the pixel is not visible. At this stage we have the data in the form of image and this image can be further analyzed so that's the important information can be retrieved. So to improve quality of the input image and make it suitable for further analysis, We perform some operation on it such as Grayscale conversion, Binary image conversion and the most important is segmentation. In this we perform some operation on scan image such as:

1) Pre-processing of images

This performs certain activities such as scanning documents, storing them as images. The module supports the following services:-

a. Scanning printed documents and storing the documents as snapshots or images.

b. Processing those image-based documents, Converting these image-based documents into proper format(also called structured documents) such as Greyscale and Binary format.

2) Segmentation:

The segmentation is the most important process in text recognition. Segmentation is done to make the separation between the individual characters of an image. Segmentation is one of the most important phases in this project. The performance of of this project is depending on segmentation. Segmentation subdivides an image into its constituent regions or objects. Basically in segmentation, we try to extract basic constituent of the script, which are certainly characters. This is needed because our classifier recognizes these characters only. In this project ,We perform the segmentation of character from image by applying Line detection and Character detection algorithm which are discuss as follows:

Algorithm: ***Line detection from image***

Step 1 :Start scanning the image horizontally from the topmost left corner row by row.
Step 2:If any black pixel is encountered in a row make the row status as '0'.
Step 3: If no black pixel in encountered in a row while tracing it then marks the row status as '1'.
Step 4:By counting and following the total numbers of continuous '0' from row status vector number and position of lines can be obtained

Algorithm: ***Character detection from the line***

Step 1: Take a single line under consideration.
Step 2:Start scanning the image vertically from the topmost left corner column by column.
Step3: If any black pixel is encountered in a column mark the column status to '0'.
Step 4:If no black pixel in encountered in a column while tracing it then marks the column status as '1'.
Step 5: By counting and following the total numbers of continuous '0' from column status vector number and position of lines can be obtained.

B. System Training Module

This module can be used to train the system for text recognition. Before converting the printed documents in to editable and searchable documents, the first and the mandatory step is providing training to the system. Here training in the sense the font followed in the scanned document should be identified by the user. Then the user types all the characters that are required for recognition from the scanned document as an image file. This image file should be provided as an input during the training process.

C. Text Recognition Module

This module can be used for text recognition in output image of pre-processing model and give output data which are in computer understandable form. Hence in this module following techniques are used.

1) Feature Extraction

Feature extraction is the process to retrieve the most important data from the raw data. The most important data means that's on the basis of that's the characters can be represented accurately. To store the different features of a character, the different classes are made. There are many technique used for feature extraction like Principle Component Analysis (PCA), Linear Discriminate Analysis (LDA), Independent Component Analysis (ICA), Chain Code (CC), zoning, Gradient Based features, Histogram etc. .

In this we use matrix feature extraction method. In this method first we convert the image to binary matrix i.e. black and white image convert to matrix form, it may look like as shown in figure 2. in the above figure text image is converted in to the matrix of 0's and 1's.from this matrix data we was extract text character line by line and word by word by using above segmentation method. After that segmented characters data are normalized and store in fixed dimension as a feature of that character which can be shown in above figure 3.

2) Classification

The classification is the process of identifying each character and assigning to it the correct character class, so that texts in images are converted in to computer understandable form. This process used extracted feature of text image for classification i.e. input to this stage is output of the feature extraction process. Classifiers compare the input feature with stored pattern and find out best matching class for input. There are many technique used for classification such as Artificial Neural Network (ANN), Template Matching, Support Vector Matching (SVM) *etc*.

In this we use Artificial Neural Network (ANN) for classification because neural network can get itself trained automatically on the basis of efficient tools for learning large databases and examples. This approach is non algorithmic and trainable. There are the different types of neural networks which can be used for the classification from which we used Kohonen neural network.

Kohonen Neural Network:

The Kohonen neural network works differently than the feed forward neural network. The Kohonen neural network contains only an input and output layer of neurons. There is no hidden layer in a Kohonen neural network. First we will examine the input and output to a Kohonen neural network.

The input to a Kohonen neural network is given to the neural network using the input neurons. These input neurons are each given the floating point numbers that make up the input pattern to the network. A Kohonen neural network requires that these inputs be normalized to the range between -1 and 1. Presenting an input pattern to the network will cause a reaction from the output neurons.

The output of a Kohonen neural network is very different from the output of a feed forward neural network. If we had a neural network with five output neurons we would be given an output that consisted of five values. This is not the case with the Kohonen neural network. In a Kohonen neural network only one of the output neurons actually produces a value. Additionally, this single value is either true or false. When the pattern is presented to the Kohonen neural network, one single output neuron is chosen as the output neuron. Therefore, the output from the Kohonen neural network is usually the index of the neuron (i.e. Neuron #5) that fired. The structure of a typical Kohonen neural network is shown in Figure 2.

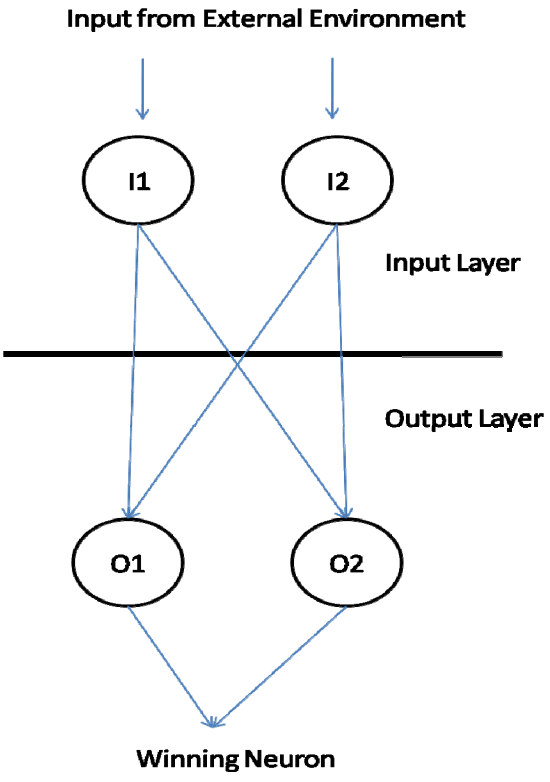

Fig.2: The structure of a typical Kohonen neural network

D. Post-processing Module

The output of Text Recognition Module is in the form text data which is understand by computer, So there need to store it in to some proper format(i.e. txt or MS-Word)for farther use such as Editing or Searching in that data.

IV. EXPERIMENTAL RESULT

The Paper document is generally scanned by the optical scanner and is converted in to the form of a picture. A picture is the combinations of picture elements which are also known as pixels. At this stage we have the data in the form of image and this image can be further analyzed so that's the important information can be retrieved. So, we apply our method of text recognition which disused in this paper and output results are shown in the form of following images.

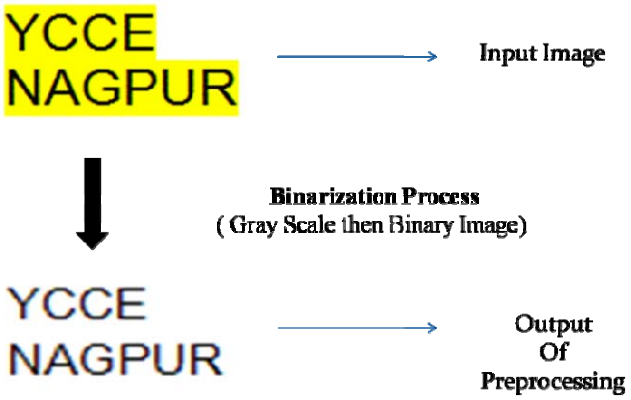

Fig.3: Output of preprocessing.

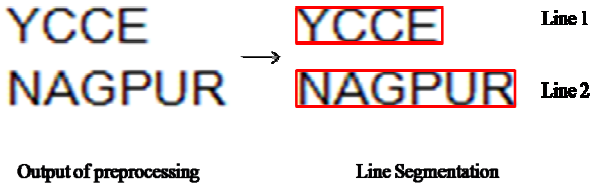

Fig.4: Output of Line segmentation.

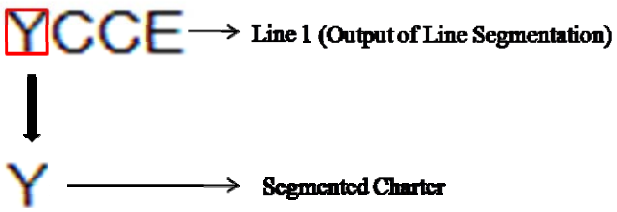

Fig.5: Output of Character segmentation.

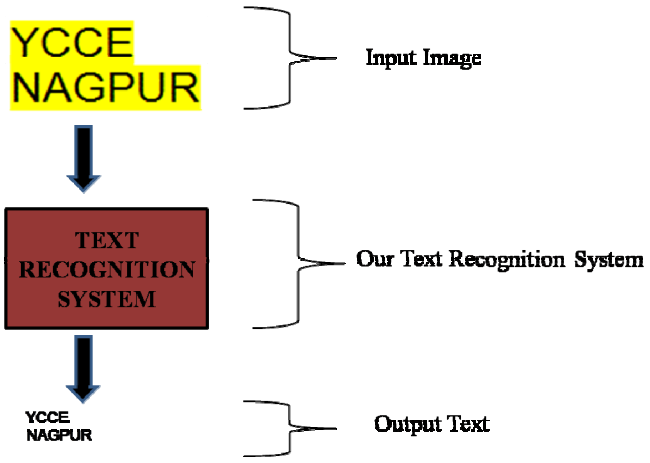

Fig.6: Output of Text Recognition System.

V. APPLICATION

Text recognition technology may be apply throughout the entire spectrum of industries, revolutionizing the document management process. This technology enable scan documents to become more than just image files, turning into fully searchable documents with text content that is recognized by computers. With the help of this technology, people no longer need to manually retype important documents when entering them into electronic databases. Instead, Text recognition system extracts relevant information and enters it automatically. The result is accurate, efficient information processing in less time. In the following, we overview some applications of text recognition system

A. Banking[18]

The uses of image text recognition vary across different fields. One widely known application is in banking, it is used to process checks without human involvement. A check can be inserted into a machine, the writing on it is scanned instantly, and the correct amount of money is transferred. This technology has nearly been perfected for printed checks, and is fairly accurate for handwritten checks as well, though it occasionally requires manual confirmation. Overall, this reduces wait times in many banks.

B. Legal[18]

In the legal industry, there has also been a significant movement to digitize paper documents. In order to save space and eliminate the need to sift through boxes of paper files, documents are being scanned and entered into computer databases. Image text recognition further simplifies the process by making documents text-searchable, so that they are easier to locate and work with once in the database. Legal professionals now have fast, easy access to a huge library of documents in electronic format, which they can find simply by typing in a few keywords.

C. Healthcare[18]

Healthcare also use of image text recognition technology to process paperwork. Healthcare professionals always have to deal with large volumes of forms for each patient, including insurance forms as well as general health forms. To keep up with all of this information, it is useful to input relevant data into an electronic database that can be accessed as necessary. By using image recognition technology they are able to extract information from forms and put it into databases, so that every patient's data is promptly recorded. As a result, healthcare providers can focus on delivering the best possible service to every patient.

D. Image text recognition in Other Industries[18]

Image text recognition technology is widely used in many other fields, including education, finance, and government agencies. This technology has made countless texts available online, saving money for students and allowing knowledge to be shared. Invoice imaging applications are used in many businesses to keep track of financial records and prevent a backlog of payments from piling up. In government agencies and independent organizations, image text recognition technology simplifies data collection and analysis, among other processes.

As the technology continues to develop, more and more applications are found for technology, including increased use of handwriting recognition.

VI. CONCLUSION

In this paper, we proposed and discussed method text recognition. The OCR is a wide area for researcher in pattern recognition. A lot of research work has been done and is still being done in OCR for various languages. More and more researchers are attracted to this challenging field. Each stage of optical character recognition has its own significance and should be designed properly for better results.

.

REFERENCE

[1] Yang, Jufeng, Kai Wang, Jiaofeng Li, Jiao Jiao, and Jing Xu, "*A fast adaptive binarization method for complex scene images*," 19th IEEE International Conference on Image Processing (ICIP), 2012.

[2] Shrey Dutta, Naveen Sankaran, PramodSankar K., C.V. Jawahar, "*Robust Recognition of Degraded Documents Using Character N-Grams*," IEEE, 2012.

[3] Gur, Eran, and ZeevZelavsky, "*Retrieval of Rashi Semi-Cursive Handwriting via Fuzzy Logic,"* IEEE International Conference on Frontiers in Handwriting Recognition (ICFHR), 2012.

[4] Rhead, Mke, "*Accuracy of automatic number plate recognition (ANPR) and real world UK number plate problems."* IEEE International Carnahan Conference on Security Technology (ICCST), 2012.

[5] Badawy, W. *"Automatic License Plate Recognition (ALPR): A State of the Art Review."* IEEE International Conference on Document Analysis and Recognition, 2012.

[6] Naveen Sankaran and C.V Jawahar, *"Recognition of Printed Devanagari Text Using BLSTM Neural Network,"* IEEE, 2012.

[7] Ntirogiannis, Konstantinos, Basilis Gatos, and Ioannis Pratikakis. "*A Performance Evaluation Methodology for Historical Document Image Binarization.,"* IEEE International Conference on Document Analysis and Recognition, 2013.

[8] Malakar, Samir, et al. *"Text line extraction from handwritten document pages using spiral run length smearing algorithm,"* IEEE International Conference on Communications, Devices and Intelligent Systems (CODIS), 2012.

[9] Application of OCR, from http://www.cvisiontech.com.