Text Recognition System From Images

A project paper submitted in partial fulfillment of the requirements for the Degree of Bachelor of Science (Engg.) in Information and Communication Technology

SUBMITTED BY

Md Iftekhar Mahmud Mozomder

Roll No: 11709013

Registration No: 11709013

Session: 2016-17

Department of Information and Communication Technology

Comilla University, Bangladesh

Submission: July, 2022

ABSTRACT

The goal of text recognition in image is to create a computer system that can automatically interpret text from pictures. Nowadays, there is a significant need for saving information that is now only available in paper documents on a computer storage drive so that it can subsequently be used by a search procedure. Scan the papers, then store the images as a straightforward method of storing data from these paper documents in computer systems. However, it is highly challenging to read the individual contents and search the contents of these documents line-by-line and word-by-word in order to reuse this information. The font qualities of the characters in printed texts and the image quality provide difficulties. These difficulties prevent the computer from reading the characters correctly. In order to execute Document Image Analysis (DIA), which converts paper documents into electronic format, character recognition algorithms are required. This study discusses a technique for extracting text from photographs. This paper's goal is to recognize text from images for the reader's better understanding by applying a specific order of several processing modules.

Keywords : OCR, Tesseract, OpenCV, python.

Chapter 1

Introduction

Since its introduction into a wide range of applications, such as the automatic scanning of license plates and signboards, text recognition has grown significantly in popularity. Nowadays, everything and everyone has gone digital. Most Currently, photographs or digital documents are used to communicate a large amount of information. There are a great deal of simultaneous storing and accessing of information in the realm of text recognition, the desire to maintain the data contained within and have access to the documents in a simpler and faster manner. An example of techniques that are convenient for transferring information from paper or books is to scan them, which transforms the data into a picture, barring its being used again. One of the practical methods for transferring information from paper or books is to have them scanned, which turns the data into a picture, preventing the scanned data from being used again in what a text looks like. Consequently, it's essential to create tools to make them editable by converting them. The objective of this essay is to research the different steps in the text recognition process using an aim to create editable documents from text images. Among the well-liked methods for text recognition is Character Recognition Through Optics (OCR). It alters scanned data. converting text images into editable format.

Text recognition begins with the capture of the image of the required document, followed by preprocessing to obtain the desired portion and segmentation to extract the text content. The challenge of text recognition from photographs is broken down into various phases in this paper.

1.1 Purpose

The primary goal of an Optical Character Recognition (OCR) system built on a grid architecture is to process electronic document formats that were previously only available in paper formats more effectively and efficiently. Compared to other available character recognition techniques, this increases the accuracy of character recognition during document processing. Here, OCR technology uses the bit-mapped representations of the characters to deduce their meaning and font characteristics.

The main goal is to faster the character recognition stage of document processing. As a result, the system can handle a large number of documents quickly, saving time.

Our character recognition is grid-based, therefore it seeks to distinguish various heterogeneous infrastructure characters from a variety of universal languages with various font characteristics and orientations.

1.2 Project Scope

The goal of our product, which performs optical character recognition on a grid infrastructure, is to give users an effective and improved software tool for document processing by reading and identifying characters in research, academic, governmental, and commercial organizations that have a large pool of documented, scanned images. Regardless of the size or kind of characters in a document, the product recognizes them, searches for them, and processes them more quickly in accordance with environmental requirements.

1.3 Current System

To ensure the security of their data, customers are increasingly requesting that printed documents be converted into electronic documents in today's business environment. Thus, the fundamental OCR system was developed to transform the data present on paper into computer-processable documents, allowing for the editing and reuse of the documents. There is no grid functionality in the current

system or the prior system of OCR on a grid architecture. The way the current system handles homogenous character recognition or character recognition of single languages is as described above.

1.4 Drawback Of Current System

Early OCR systems had the limitation of only being able to convert and recognize documents written in English or a certain language. The older OCR system is hence monolingual.

1.5 Proposed System

Our suggested solution is a character recognition system called OCR on a grid infrastructure, which can recognize characters from various languages. This function, which we refer to as grid infrastructure, solves the issue of heterogeneous character recognition and enables many capabilities to be applied to the document. While the current system just provides document editing, the multiple functionalities also support editing and searching. In this context, "Grid infrastructure" refers to the system that supports a certain group of languages. OCR is multilingual on a grid infrastructure because of this.

1.6 Advantage Of Proposed System

The advantage of the proposed system over the current system is that it enables numerous functionalities, including editing and searching. Additionally, it offers the ability to recognize characters of different types.

1.7 Architecture Of Proposed System

The three primary parts make up the architecture of the optical character recognition system on a grid infrastructure. They are as follows:

a. Scanner
b. OCR Hardware or Software
c. Output Interface

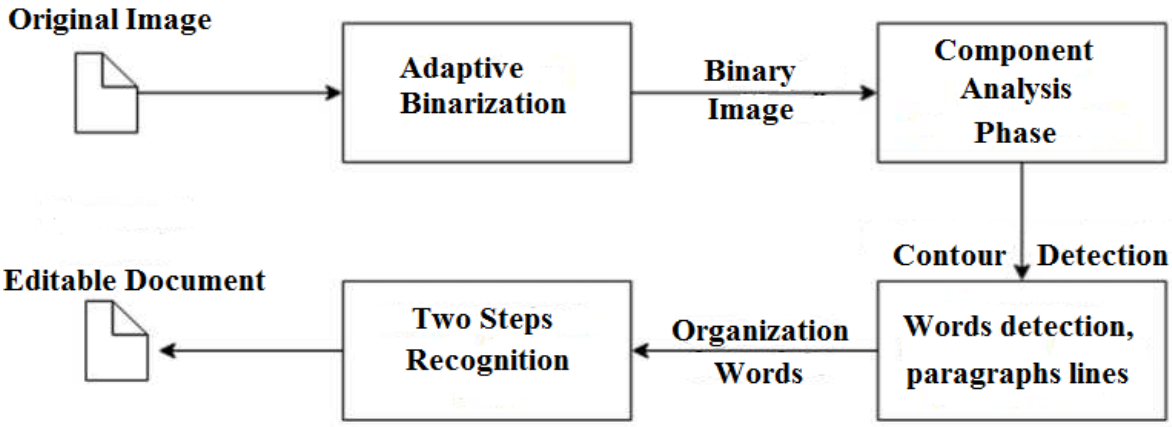

Fig 1.1 : Architecture of OCR

1.8 Application

The document management process could be revolutionized by the application of text recognition technologies across the board. Containing the aid of this technology, scanned documents can be converted into fully searchable documents with computer-readable text rather than just image files. This method eliminates the need for manual retyping of critical documents when putting them into electronic databases. Instead, a text recognition system pulls the necessary data and automatically inputs it. As a result, information is processed accurately and effectively in less time. The following is an overview of some text recognition system applications.

A. Banking

Image text recognition has applications in a variety of industries. One well-known use is in banking, where it is applied to the automated processing of checks. The writing on a check can be inserted into a device, quickly scanned, and the right amount of money sent. Although occasionally requiring manual confirmation, this technology is almost perfect for printed checks and is also fairly accurate for handwritten checks. Overall, this shortens lines at numerous banks.

B. Legal

There has also been a substantial push to digitize paper records in the legal sector. Documents are being scanned and entered into computer databases in order to free up space and remove the need to comb through bins of paper files. By making documents text-searchable so that they are easier to find and use once in the database, image text recognition further streamlines the process. Legal practitioners now have quick and simple access to a vast collection of electronic documents that may be found by merely entering in a few keywords.

C. Healthcare

The processing of paperwork in the healthcare industry also uses image text recognition technologies. Healthcare personnel constantly have to deal with a lot of paperwork for each patient, including general health and insurance documents. It is helpful to enter pertinent data into an electronic database that can be accessed as needed to keep up with all of this information. They may extract data from forms and enter it into databases using image recognition technology, ensuring that each patient's data is promptly logged. Healthcare providers can thus concentrate on giving each patient the best care possible.

D. Recognition of Image Text in Other Industries

The usage of image text recognition technology is widespread in a variety of other industries, such as government, banking, and education. Numerous texts are now easily accessible online thanks to this technology, saving students money and promoting knowledge sharing. Many organizations use invoice imaging software to manage their financial records and avoid building up a backlog of unpaid invoices. Image text recognition technology, among other things, makes data collection and analysis easier for both independent and government entities. As technology advances, more and more uses are discovered for it, including a rise in the use of handwriting recognition.

Chapter 2

Literature Review

As was previously noted, pattern recognition research is still being done on text recognition from photos. Many researchers have presented many methods to address the problems with text recognition; each strategy or technique seeks to solve the problems in a different way. We provide a thorough overview of the methods suggested to address the problems with text recognition in the part that will be published soon.

A brand-new adaptive binarization technique based on wavelet filter is proposed by Yang et al. [1]. This method was developed to process information more quickly, making it more suitable for real-time processing and mobile devices. On ICDAR 2005 database photographs of complicated scenes, they tested this adaptive methodology. On images of badly degraded Indian language document, Sankaran et al. [2] developed a novel recognition strategy that resulted in a 15% reduction in word error rate.

Some issues with text recognition and retrieval have been highlighted by Gur et al. [3]. In most instances, human inspection is necessary because automated optical character recognition (OCR) methods do not provide a comprehensive answer. On the basis of statistical information about the examined typeface, they propose a novel fuzzy logic-based text recognition system. With the use of a set of fuzzy based criteria, the novel method combines letter statistics and correlation coefficients to recognize altered letters that could not otherwise be retrieved. They concentrated on handwritten calligraphy employed in the rashi typefaces used in biblical commentary.

Real-world UK license plates were taken into consideration by Rhead et al. [4] who connected them to ANPR. When applying them to actual number plates, it takes certain features of the applicable laws and standards into account. Additionally included are the various manufacturing processes and component part specs. The various fixing techniques and locations, as well as their effects on image capture, are examined.

According to Badawy, W. et al. [5], automatic license plate recognition (ALPR) is the process of extracting information about a vehicle's license plate from an image or a series of photos. Many applications, including electronic payment systems (toll payment, parking fee payment), freeway and arterial monitoring systems for traffic surveillance, can use the extracted information with or without a database. A color, black-and-white, or infrared camera is used by the ALPR to capture images.

Devanagari, an Indian script, has a recognition technique presented by Jawahar et al. [6]. One of the most frequent causes of a high word mistake rate, word to character segmentation, is not necessary using the approach they adopted. When compared to the best OCR system currently available, they have reportedly experienced a reduction of over 20% in word mistake rate and over 9% in character error rate.

According to research by Ntirogiannis et al. [7], the document image binarization stage of the document image analysis and recognition pipeline is crucial because it influences subsequent phases of the recognition process. By providing qualitative and quantitative indications of its performance, the evaluation of a binarizationmethod assists in understanding its algorithmic behavior and confirming its efficacy. They suggested a mechanism for pixel-based binarization evaluation for old handwritten and machine printed document pictures.

According to Malakar et al. [8], one of the crucial tasks in the development of an Optical Character Recognition (OCR) system is the extraction of text lines from document images. Because of the possibility of touching, overlapping, or warped text lines in handwritten document photographs, this procedure presents a significant difficulty to the researcher.

2.1 Features

Increased productivity

OCR software enables quicker data retrieval when needed, which increases productivity for enterprises. The time and effort that the employees previously had to use to retrieve pertinent data can now be directed toward essential tasks. Employees may obtain the necessary documents without leaving their desks, eliminating the need for multiple journeys to the central records room.

Cost cutting

One of the most significant advantages of OCR data input technologies is that they enable organizations to avoid hiring specialists to perform data extraction. Additionally, this instrument aids in reducing several other expenses like copying, printing, shipping, etc. As a result, OCR eliminates the expense of misplaced or lost papers and provides further savings by reclaiming office space that would otherwise be required for archiving paper documents.

High Efficiency

Inefficiency is one of the main problems with data entering. OCR data input is one example of an automated data entry solution that reduces errors and inaccuracies for effective data entry. OCR data entry can also be effective in solving issues like data loss. Since there is no labor involved, problems like mistakenly entering incorrect information can be avoided.

Larger Storage Space

OCR is able to scan, record, and catalog data from paper documents used throughout an entire organization. This simply means that there is no longer a need to keep massive paper files because the data can now be saved in servers in an electronic format. Thus, one of the strongest tools for implementing a "Paperless" approach throughout the firm is OCR data entry.

Enhanced Data Security

For any firm, data security is of the highest significance. Paper records are readily misplaced or destroyed. Papers are susceptible to loss, theft, and natural disasters including fire, dampness, and pest infestation. With data that is scanned, examined, and saved in digital formats, however, this is not the case. To avoid improper handling of the digitized material, it is also possible to restrict access to these digital documents.

All Documents Are Text Searchable

OCR data processing has a lot of benefits, one of which is that it makes the digitized documents fully text searchable. This makes it easier for experts to seek up names, numbers, and other identifying information about the document they are searching for fast.

Enables Document Editing
Most of the time, scanned documents need to be edited, especially if some information needs to be updated. OCR translates data into any desired formats, including editable Word and other formats. When there are contents that need to be updated frequently or changed, this can be quite helpful.

Emergency Recovery
One of the main advantages of using OCR for data entry is disaster recovery. Data is kept safe even in emergency scenarios when it is stored electronically in secure servers and distributed networks. The digitized data can be promptly retrieved in the event of a natural disaster or sudden fire outbreak to ensure business continuity.

2.2 Benefits of Online Converter

1. Data Conversion and Use Ease

Users can easily alter their documents with the help of these online text converters. Additionally, consumers will be able to utilise these crucial facts digitally thanks to the data gathered from this program.
This converter can be used, for instance, to change the old manual input into an useful digital format if a user converts last year's manual input to a fresh digital file.

2. Text Extraction for Media Businesses

Media businesses frequently take vital information from enormous photos. As a result, these businesses can quickly and easily extract crucial information from big photographs using image-to-text converters.

3. Enhanced Compliance with Audits

In many businesses, documents and files are checked daily. Data that is inaccurate and jumbled up might make auditing very difficult. Additionally, there may be severe repercussions if compliance criteria are not met.
Businesses may handle their physical files, photos, and printed material more effectively by transferring accurate data from them into their systems using image-to-text converters.

4. High rates of accuracy

Using OCR technology lowers the possibility of human error. In order to cover more material in less time, people typically enter stuff faster.
For instance, it could be a major issue for your company if you are creating a lengthy policy paper and you forget to mention a policy. You need to use OCR more accurately in light of this. It fixes every issue. It successfully extracts text from photos.

5. Affordable

Because OCR technology is cost-effective, firms can employ OCR-powered web converters to increase their operational efficiency. I appreciate the converter's assistance in letting me use the software to control the utilities.

6. Saving time

By performing this operation manually, I'm wasting time and energy. Multiple files can be converted simultaneously with no issues using JPG to text Converter. As a result, more time is freed up for other crucial tasks.

7. High quality

These photos are typically hazy because of the camera's poor quality. However, there are numerous instances in which text can be extracted from low-resolution and blurry photographs using jpg to text converter. It offers high-resolution text in that case.

8. Sustainability

OCR produces no trash that harms the environment. However, you can considerably lower office wasteful paper. Because paper makes up half of office waste, this technology can be reduced.
In addition, the paper industry requires a lot of water and billions of trees in the process of producing paper. Businesses can employ image-to-text solutions to preserve their ecosystem rather than using these vital resources and harming the environment.

2.3 Cons of Online Converter

The quality of the documents produced by OCRed is one of the main drawbacks of the technology. The quality of the input image that OCR is given determines how well it performs. This implies that OCR will struggle to extract text from an image if it contains any flaws. Since the user frequently needs to remedy the OCR problems before reprocessing with OCR, OCR errors might be considerably more challenging to fix.

The possibility of errors being introduced that could affect the document's value is one of optical character recognition's key drawbacks.
OCR has the potential to make mistakes, like misclassifying a character as a word or line break. When an OCR system converts text to text and misidentifies one character as another, this is called a character recognition error. For instance, the OCR might identify "N" and transform it to "E." This frequently occurs in texts that use non-English characters.

Chapter 3

Requirement Specification

3.1 Software

Python is a popular general-purpose, high-level, interpreted, dynamic programming language. The language offers structures designed to support both small- and large-scale, concise programs. Python supports a variety of programming paradigms, including imperative, functional, and object-oriented or procedural programming. It offers a huge and thorough standard library, a dynamic type system, intelligent memory management, and more. By constructing straightforward and intelligible Python routines, it was actually very beneficial in the digital transformation of the stock photographs.

Operating System

- Windows

Hardware components

- Personal Computer

3.2 System functional requirements:

These functional criteria have been categorized as follows:

- Selecting the preferred text image
- Understanding the text
- Text duplication in a different location

Chapter 4

Design Specification

This section provides a quick overview of the text recognition system's overall design, which is depicted in figure 1. An image containing some text information is the input that a text recognition system receives. This technology produces text information in images that are saved in computer-readable form and are output in electronic format. The following modules can be used to partition the text recognition system:

(A) Pre-processing

(B) Text recognition

(C) Post-processing.

A. Pre-processing Module

The paper document is typically scanned using an optical scanner and then transformed into a picture. A picture is made up of various visual components, sometimes referred to as pixels. We currently have data in the form of an image, which may be further processed in order to extract the most crucial information. Therefore, a few operations are carried out to enhance the image, such as noise reduction, normalization, and binarization, in order to increase the quality of the input image.

B. Text Recognition

This module can be used to recognize text in pre-processing model output images and provide output data that can be understood by computers. Consequently, the following strategies are employed in this module.

Segmentation

The segmentation method is the most crucial step in the text recognition module. To distinguish between an image's individual characters, segmentation is used.

Extraction of Features

The technique of extracting the most crucial information from raw data is known as feature extraction. The most crucial information is that which establishes the validity of the characters' proper representation. The various classes are created to store the various characteristics of a character. Principle Component Analysis (PCA), Linear Discriminate Analysis (LDA), Independent Component Analysis (ICA), Chain Code (CC), zoning, Gradient Based Features, Histogram, and other techniques are used for feature extraction.

Classification

To translate writings in photographs into computer-understandable form, each character is identified and given the appropriate character class through the classification process. This procedure utilised extracted text image features for classification; therefore, the output of the feature extraction phase was the input to this stage. Classifiers choose the best matching class for input by comparing the input feature with the stored pattern. There are various techniques used for classification, including Support Vector Matching (SVM), Artificial Neural Networks (ANN), and Template Matching.

C. Post-processing Module

The output of the text recognition module is text data that the computer can understand, thus it must be stored in a suitable format (such as text or MS-Word) for later use, such as editing or searching in that data.

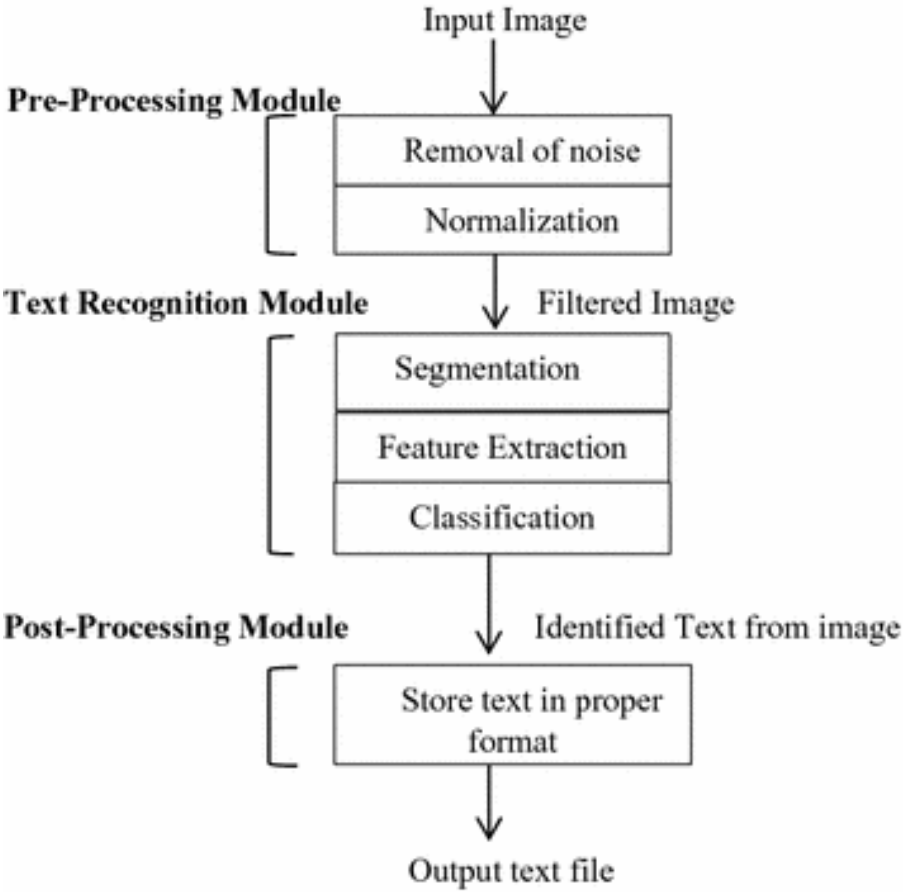

Fig 4.1 : Architecture of text recognition system

Chapter 5

Development and Testing

We currently have data in the form of an image, which may be further examined in order to extract the key information. The output results of our text recognition approach, which was not employed in this paper, are shown in the following picture

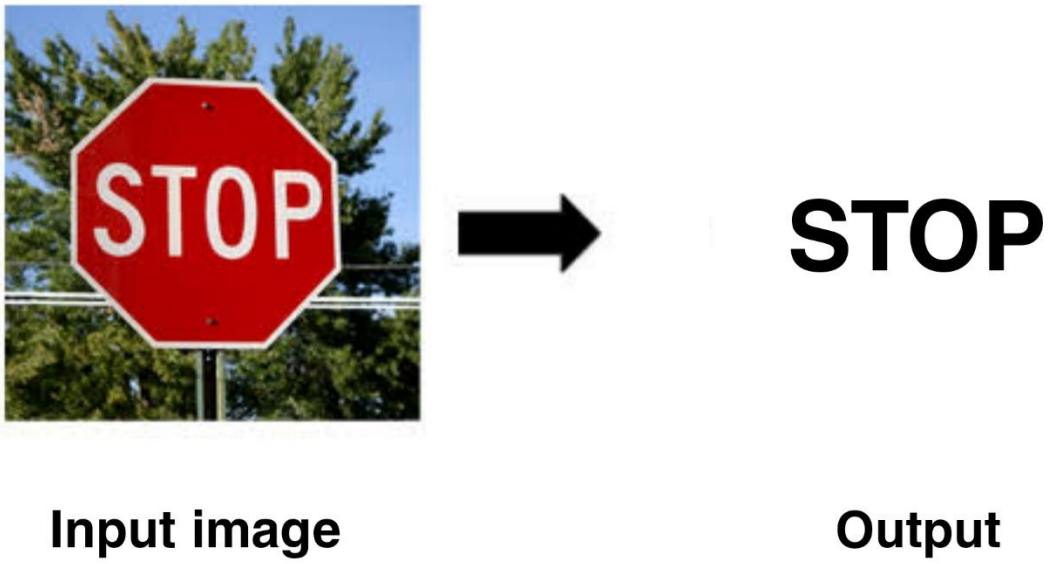

Fig 5.1 : Output of Image processing

Fig 5.2: Text recognition system

Testing:

Testing is a crucial phase that aids in error detection. Finding errors that could happen during the implementation phase is the process of testing. Additionally, it is a means of determining whether the product satisfies the requirements and examining the component functionalities.

Chapter 6

Conclusion

Optical Character Recognition can be used to recognize symbols, and it has proven to be reliable and accurate under a variety of conditions with variable resolutions, noise levels, and filters. The method is unique in a way since it can handle both character recognition and symbol recognition simultaneously.
The Neural Network(NN) classifier is a potential remedy for OCR and symbol identification and could be the ominous element that is currently lacking. Finding the content of each symbol can be greatly aided by combining the old static and adaptive classifier with the NN classifier.
Symbol recognition using optical character recognition is definitely feasible, and both character and symbol recognition will benefit from the ongoing improvement of classifier accuracy.

Future Work

More test photos are needed to calculate the dynamic morphological parameters in order to create a more reliable algorithm. The most crucial component of locating the blobs into which the information is separated, even while it provides an overall good outcome, should be added as an optional input for the testers.
To make sure the program is effective and used for future purposes, I will also focus on the text with skew.

REFERENCES

[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