Review of Optical Character Recognition using Pytesseract

1st Abhishek kumar Jha
UIE(University Institute of Engineering)
Chandigarh University
Mohali, India
abhiiishek.jha@gmail.com

2nd Arun Prakash Singh
UIE(University Institute of Engineering)
Chandigrh University
Mohali, India
fjapsinghcse98@gmail.com

3rd Eknoor Singh
UIE(University Institute of Engineering)
Chandigrh University
Mohali, India
singh18eknoor@gmail.com

4th Priyanka Aggarwal
UIE(University Institute of Engineering)
Chandigarh University
Mohali, India
priyankaaggarwal2004@gmail.com

5th Kumari Kashish
UIE(University Institute of Engineering)
Chandigarh University
Mohali, India
kashishgupta9877@gmail.com

6th Er. Pushkar Sharma
UIE(University Institute of Engineering)
Chandigarh University
Mohali, India
pushkar.e14833@cumail.in

***Abstract*—** In recent years the research done in the field of ocr is advanced to the next level of understanding irrespective of machines and computers. The development of digital libraries throughout the world has led to the conversion of all photographic images into an edited format which is easy to understand and in the upcoming centuries irrespective of Handwriting Styles. The Ocr converts the document with the help of ANN and AI to increase the training of the system to improve its accuracy rate. The ocr faces a complex problem because of different writing pattern and different character for different type of languages various recognition techniques like Character Normalization, Correlation, Neural Network, Recognition, Hidden Markov Model, and Correlation method. In our research paper we explain the different types of ocr that are available in the market and the customer can use it on his basic requirement need. This research paper deals with the new technique that is applicable for creating an improved Ocr which results in improved accuracy.

Keywords— ANN(Artificial Neural Network), AI(Artificial Intelligence), OCR(Optical Character Recognition, ANSI(American National Standard institute, GISMO(Geophysical Institute Seismology Matlab Objects) etc.

1. Introduction -

In today's world we have a large amount of documents into our possession both the modern and historical documents, Hence the development of digital libraries has helped in a reliable and accurate manner. The Historical documents which are generally found in fossils and are developed by our great historians are of more importance to us because they represent our cultural heritage. Optical Character Recognition (OCR) is a software which helps in converting the text and images into the handwritten documents the results are not much efficient and satisfactory. The OCR is sometimes unable to execute the historical documents because of their low quality images, lack of font style and the presence unknown facts. Hence recognition of historical documents is the most challenging task in OCR. Even though the human brain has the capability to easily recognize the text from an image.

OCR is still a problem after the advancement in research because of the language barrier in modern and past times, because of their writing pattern that is stored in the form of images and their language techniques. The techniques from different subsystems of computer science that are required for developing and OCR are image processing, pattern classification and natural language processing. The word spotting technique is used to search the similar words in the document and to collect them into a cluster by using image matching. These clusters provide a better accurate result while executing different steps on the handwritten or image document. Moreover all the work tend to focused on defining the unique in each characteristics by their content and writing style. The primary step is to convert the images into binary sets, a top-down breakdown helps in extracting the characters.

The database is a combination of extracted characters converting the document into a text file. It can be applied to different types of documents to deal with characters that do not appear more often in the document. Hence the combination of several approaches binarization, segmentation and pattern recognition deals with such as image enhancement and clustering that leads to a recognition system for a document which is printed or handwritten.

2. Literature Review

The OCR has faced problems since it came into the introduction before computers were introduced. The first OCR was not a computer but a device which was able to recognize the characters but the efficiency rate of that machine was slow and it also had the slow processing speed.

The year 1929 was the first of its kind when Gustav Tauschek patented the OCR in Germany and second was acquired by Handel in 1933 from the USA. The machine made by Tauschek recognize the characters with the help of templates and photo detectors. RCA(Radio Corporation of America) started working on the computer based OCR in the year 1949 to help blind people. But lately their machine was way to expensive and was unable to achieve the desired goal because the machine was firstly converting the characters into machine language and then spoke the

letters for the blind people. The year 1951 plays a crucial role in development history of an OCR because M.sheppard laid the foundation of the modern OCR and named it GISMO. GISMO was able to perform different functions as it was able to read all the musical symbols on the printed pages one by one, but it was only able to recognize 23 alphabetical characters. The next year J.Rainbow which had the speed of reading the English uppercase characters one per minute. The complexity and problems that were faced during the development of OCR lead to a common decision that there should be standardized OCR fonts on which system will execute its process. This mutual decision lead to formation of ocra and ocrb with the collaboration of ANSI and EMCA which helps in providing the efficiency rate of an ocr. Kurzweil Computer started selling of OCR as a commercial product in the year 1978 and firstly it was used as program to upload the documents into the online database. The Kurzweil computer became a subsidiary of Xerox in the year 1967 and they introduced a new technique for sharpening of pixels which helps in restricting the histogram. Wantanable also proposed a histogram difference method which helps in selecting the maximum threshold at gray level. This helps in modifying the histogram so that the histogram is useful for thresholding. Another technique is to deal with parametric technique such as to take the sum of gaussian distributions with the help of least sense square and statistical decision. These methods are time consuming and requires high computational power. There is also one technique which finds the optimal solution with the help of nonlinear multimodal function optimization. There also techniques which uses the fuzzy logic and the artificial neural network to execute the ocr. Crowding is a method which helps in providing a better and reliable results of bimodal histograms in multimodal optimization problems. This lead the ocr to the formation of document image analysis(DIA), handwritten and multiple lingual ocr. Even after the advancement and the research done, the ocr ability to understand the contents is far below from human. The current research that is being executed is to improve the efficiency of ocr and more fonts and writing styles should be added into the current system of ocr. This research helps in explaining the extraction of images and the final proposed system.Threshold from the system can also be achieved with the help of genetic algorithms.

3. Types of Optical Character Recognition -

In recent years the research done in the field of ocr is advanced to the next level of understanding irrespective of machines and computers. This part of the research paper will explain the different types of ocr that are available in the market and the customer can use it on his basic requirement need. These ocr's were based on their different font-restrictions, character recognition and image extraction. The input of an ocr can be of two types either it can be handwritten or machine printed recognition system. The earlier ocr's were easy to predict the results because of the common character style and the position of the ocr on the document page. The ocr faces a complex problem because of different writing pattern and different character for different type of languages. These can broadly categorized into two parts - online and offline. The online system can execute its process while the users are writing the characters. They can also help in capturing the speed of writing the characters. Hence they are less complex. Whereas the offline system operates its execution on the constant data, it results in a complex pattern recognition. The demand of online system is much more than offline because it provide better accurate results, are easier to develop, and can be incorporated in tablets and Fig[1] gives a pictorial representation of types of OCR.

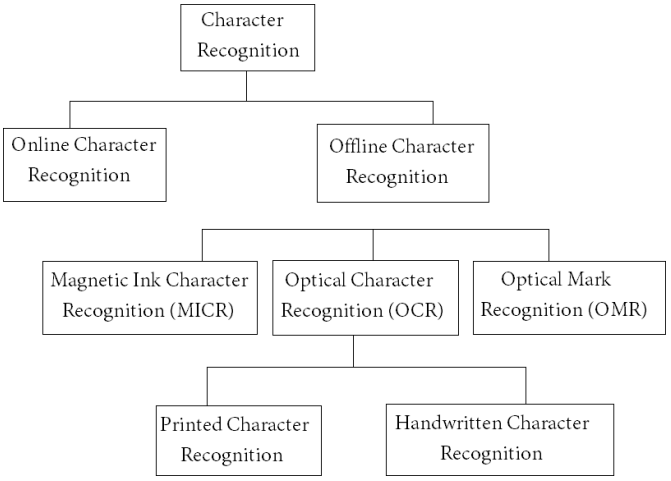


Figure 1 – Types of Character Recognition

4. Recognition Techniques -

Another technique of character recognition is offline based and the similarity of the pattern between the document and the database created with the extraction of the characters. There are different steps required to achieve character recognition are - image processing, neural network, character positioning and segmentation.

Different types of approach used are -

4.1 Neural Network -

Back Propagation neural network is used to achieve pattern recognition because it gives a direct result of success and failure. If there is a failure then the system will automatically revert back the last step and select a different step to achieve success. It is a combination of two phases-Back and Forward Propagation. Fig[2] explains the neural network architecture.

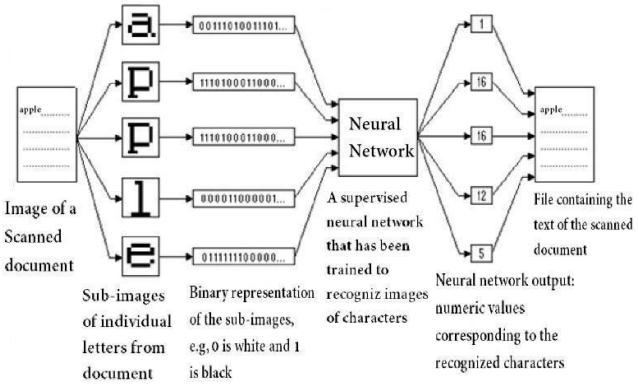


Figure 2 – Example of Neural Network of OCR

4.2 Character Normalization -

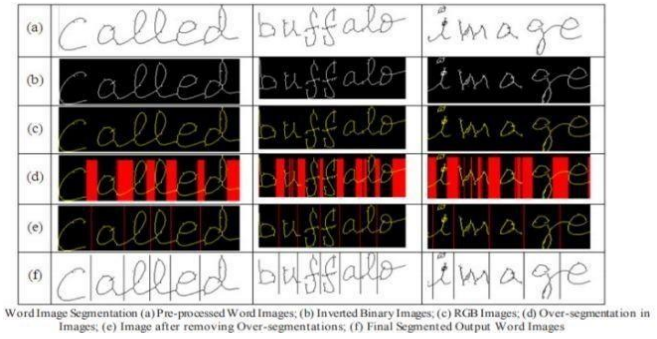
This step helps in normalizing all the characters, numbers to a standard size including all the headings, subscript and superscript. So that the input given to the system for recognition comprises all the characters of same size. This will help in faster evaluation.

4.3 Correlation Method -

This step helps in reducing the noise from an image because even a perfect image will carry some amount of noise in it and to remove the noise, the image is converted into a binary image and process is called as digitization and those pixels which are less than 30 are removed from the image because they provide a very less impact on the image. Hence the digitized image is then sent to the system for further processing.

4.4 Segmentation -

The size of an image is selected according to the template and the image is cropped so that the image perfectly fits into the template. Fig[3] The characters are segmented on the basis of preprocessing and RGB images.



Word Image Segmentation (a) Pre-processed Word Images; (b) Inverted Binary Images; (c) RGB Images; (d) Over-segmentation in Images; (e) Image after removing Over-segmentations; (f) Final Segmented Output Word Images

Figure 3 – Segmentation in OCR

4.5 Recognition -

The segmented image is loaded into the system and the character which has the maximum correlated value is considered as the character present in the image and converted into a text document.

4.6 Hidden Markov Model -

It comprises a hidden layer and is not visible directly but is observed through the input layer used for computations. The hidden layer consists of probabilities that are assigned to each character in the image and based on their values they are selected and discarded and the function is called a Probability Density function.

5. Steps of Optical Character Recognition

5.1 Image Acquisition - It is the collection of images for conversion into printed format from other sources.

5.2 Pre-Processing - This phase is referred to as improving the quality of the image and increasing the sharpness of an image, so that all the characters in an image are clearly visible and the system can perform character recognition with better accurate results.

5.3 Character Segmentation - This step helps in extracting single characters from an image and they are sent to the recognition system. As well as the characters which are broken and if any image has some noise present in it. In these types of situations advance character segmentation technique is used.

5.4 Feature Extraction – After the segmentation the characters are separated based on their features which are extracted from high quality images and are observed with the help of inter-class variations and those characters are selected which are efficiently computable.

5.5 Character Classification - This step helps the segmented characters to arrange them into different categories and classes. After evaluating their result they will divided into two categories-

1. Structural Pattern Classification - Feature extracted from the structure of image.

2.Statistical Pattern Classification - Based on their Probabilistic models and other methods to classify the characters.

5.6 Post Processing -

After performing the following steps it is not possible to achieve accurate result, hence we apply certain techniques to a better result. These techniques involves nlp, lc to remove errors from the system. The complexities of ocr should also be reduced so that it take less time to execute the overall process. It also involves spell checker and a dictionary to improve the accuracy of the image entered as an input.

5.6.1 Image Acquisition - It gives digital image as an input and converts into a machine editable format which are easily processed by the computer. It also involves compressing up of image and the binarization of the image involves two levels to characterize the image. It can be lossy and lossless.

5.6.2 Pre-Processing - It helps in enhancing the quality of an image. Each binary image contains threshold value and they can be set as local and global value. The techniques used to find out the skew in document - projection profiles, nn methods. Hence pre-processing of images also helps in projections and clustering of pixels.

5.6.3 Character Segmentation - This process can be processed as explicitly and implicitly both in the classification phase.

5.6.4 Feature Extraction – The unique features are extracted from the characters with the help of geometrical and statistical features and pca can be used to reduce the dimensionality of an image.

5.6.5 Classification – There are two types of approach: statistical and structural approach to classify the image. Statistical classifiers are bayesian classifiers, neural network classifier. Hence document is formed with the help of single characters which are processed.

5.6.6 Post-Processing - The final step is to provide a better quality image to the system and to provide accurate result. Contextual and lexical processing is done to reduce the chances of errors. Fig[4] differentiates the training phase and testing phase which are the essential requirements to develop an OCR.

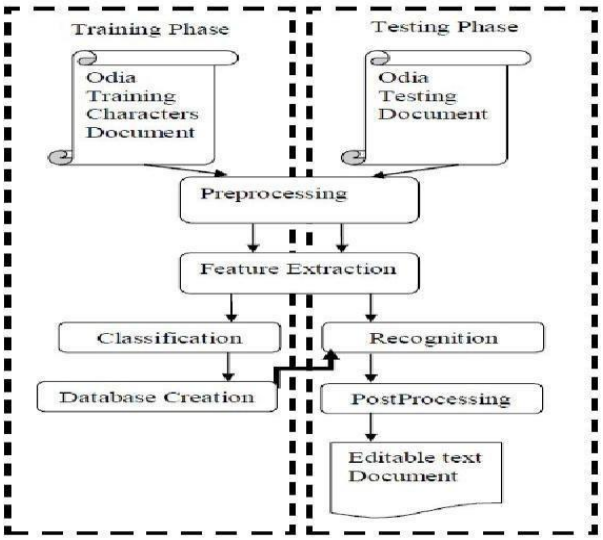


Figure 4 – OCR Process

7. Applications of Optical Character Recognition -

The ocr helps in evaluating the bank cheques and signature verifications. It helps in evaluating the documents in offices where they are present in printed form.

Ocr also helps the blind people and visually impared people to execute their work.

8. Conclusion -

This paper briefs the different techniques that are required to execute ocr. Hidden markov models and neural networks give best results among all the techniques that are used to develop optical character recognition. The results are efficient and the major concern that should be given is to the quality of an image so that characters are easily recognized and provide faster results to make an error free system. The subprocess that are necessary and equally important such as segmentation, feature extraction and classification. But the arabic and urdu is still a challenge for an ocr to detect and this leds to many fields that require research to be done.

In future, the new font will be introduced into the system which will provide a better clear image text document. Also, ocr will be applied on a website to work independently and perform its execution.

REFERENCES

[1] Honggang Wang, Ming C. Leu and Cemil Oz, "American Sign Language Recognition Using Multidimensional Hidden Markov Models", Journal of Information Science and Engineering, January 2006.

[2] Lionel Pignou. Sander Dieleman, Pieter-Jan Kindermans, Benjamin Schrauwen, "Sign Language Recognition Using Convolutional Neural Networks", March 2015.

[3] Rafiqul Zaman Khan and NoorAdnan Ibraheem, "Hand Gesture Recognition: A Literature Review", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3 , No.4 , July 2012.

[4] Karthick Arya, Jayesh Kudase, "Convolutional Neural Networks based Sign Language Recognition", International Journal of Innovative Research in Computer and Communication Engineering, Volume 5, Issue 10, October 2017.

[5] Vivek and N. Dianna Radpour, "Using Deep Convolutional Networks for Gesture Recognition in American Sign Language", Department of Computer Science, State University of New York at Buffalo, 2017.

[6] Optical character recognition (ocr) based vehicle's license plate recognition system using python and opencv Milan Samantaray, Anil Kumar Biswal, Debabrata Singh, Debabrata Samanta, Marimuthu Karuppiah, Niju P Joseph 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 849-853, 2021

[7] Optical character recognition for English handwritten text using recurrent neural network R Parthiban, R Ezhilarasi, D Saravanan 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), 1-5, 2020