International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 4 Issue 5, July-August 2020 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Machine Learning in the Field of Optical Character Recognition (OCR)

Mr. Rishabh Dubey

Student, Bharati Vidyapeeth's Institute of Management and Information Technology, Navi Mumbai, Maharashtra, India

ABSTRACT

Optical character recognition (OCR) deals with the process of identification of alphabets and various scripts. Optical character recognition (OCR) is one of the trending topics of all time. Optical character recognition (OCR) is used for pattern detection and artificial intelligence.

Machine learning is widely used in the field of OCR to provide good accuracy in the result. In Python, Pytesseract is an optical character recognition (OCR) tool for python. The paper starts with an introduction and brief background and history of Optical character recognition (OCR) systems. Then the various techniques of OCR systems such as optical scanning, location segmentation, pre-processing, feature extraction and post-processing. The different applications of OCR systems are highlighted next followed by the current uses of the OCR systems. The future of the Optical character recognition (OCR) systems with machine learning environment is presented.

KEYWORDS: *OCR, Machine learning, Detection, pytesseract, python*

How to cite this paper: Mr. Rishabh Dubey "Machine Learning in the Field of Optical Character Recognition (OCR)" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-4 | Issue-5, August 2020, pp.1664-1668, URL: www.ijtsrd.com/papers/ijtsrd33233.pdf

IJTSRD33233

INTRODUCTION

Incorporations, institutes, and offices an overwhelming volume of paper-based data challenges their ability to manage documents and records. Most of the fields become computerized because Computers can work faster and more efficiently than human beings. Paperwork is reduced day by day and the computer system took place of it. The Digitalization is taking place where the documents are more secured.

Generally identifying handwriting alphabets, image recognition is the main feature of the OCR. Optical character recognition (OCR) is the technique of performing automatic identification. Optical character recognition (OCR) is the process of classification of optical patterns contained in a digital image. The idea starts with the process of character recognition which is achieved through segmentation, feature extraction, and classification. This paper presents the basic ideas of OCR needed for a better understanding of the book or other handwritten and images.

Optical character recognition (OCR) is a process of converting a scanned document into a text document the scanned document can be in any form of documents like images so it can be easily edited if needed and becomes searchable. OCR is the mechanical or electronic translation of images of handwritten or printed text into machine-editable text that is the digitalization of document. The OCR system and its recognition engine interpret the scanned images and turn images of hand written or printed characters into ASCII data (Machine-readable characters). It occupies very less space than the image of the document. The document which is converted to a text document using OCR occupies 2.35 KB while the same document converted to an image occupies 5.38 MB. So, whenever document size is very large instead of scanning, OCR is preferable.

LITERATURE REVIEW:

Optical character recognition (OCR) system is most suitable for the applications like automatic number plate recognition, data entry for business documents (e.g. check, passport etc.), multi choice examination; almost all kind of form processing system.

Initially OCR systems were mechanical devices not computers, that were able to recognize characters, but the speed was very slow and less accuracy in result. Optical Character recognition is not a new problem but its roots can be traced back to systems before the inventions of computers.

In [2], M. Sheppard invented a reading and robot GISMO in 1951 that can be considered as the earliest work on modern OCR. GISMO was able to read musical notations and words on a printed page one by one. However, it can only recognize 23 characters. The machine also has the ability to copy a typewritten page.

In [3], J. Rainbow, developed a machine that can read uppercase typewritten English characters, one per minute. The early OCR systems were poor in performance due to errors and low recognition speed. That's why not much research efforts were put on the topic during 60's and 70's. The Developments were done on government organizations and big corporations like banks, newspapers and airlines etc. And because of the complexities associated with recognition, it was felt that three should be standardized OCR fonts for easing the task of recognition for OCR. That's why OCRA and OCRB were developed by ANSI and EMCA in 1970, that provided comparatively acceptable recognition rates.

Past thirty years, numbers of research have been done on OCR and emergence of document image analysis (DIA), multi-lingual, handwritten and omni-font OCRs taken place [2].

In [4], The author presented alphabeticity extraction technique in usual scene images. Along with expand alphabet recognition scheme using alphabeticity extraction technique in natural scene images. A graph matching method utilized structural alphabeticity of alphabet. It is recognition technique to simplify it consider relative of location and structural relation. It is robust technique to change of rotation or font. And to proof it; they experienced two cases that training font and test font are similar case or distinction case.

In [5], they evaluate a fresh alphabet recognition technique of certify plate number based on similar BP neural networks. And the enhanced the correctness of the identification scheme that aims to understand writing repeatedly the Chinese license plate. In the planned method, the quality is binarized with the sound is eliminate in the pre-processing stage, then the alphabet alphabeticity is extract with frame by means of the alphabet is normalize to size 8*16 pixels. The alphabet attribute is place addicted to the similar neural network as a final point, and the alphabet is documented. The anticipated technique in alphabet recognition is effectual, and hopeful grades have been obtained in experiment on Chinese certify plates.

Using the parallel neural networks with other methods they compare the alphabet gratitude presentation. They gave the compare of the alphabet recognition rate and recognition time among three methods in which 1st method represents their proposed method, 2nd method is the method of using simple BP neural networks, and 3rd method is the method of using template matching. They utilized 400 number-plates to test their algorithms. The experimental results shown that this process could improve more than 8-10% correct rate of alphabet recognition comparing with method 2nd, while 6-8% compare with 3rd method. The timing of recognition using their method is close to 2nd method, and is less than 3rd method. In addition, using method 1, the probability which two alphabets and more than two alphabets in the same number-plate are all mistake recognition is less than 4% that performance is better than that using other two methods. In maximum cases, the time of all seven alphabets recognition is close to 0.01 second, and the recognition rate of these alphabets is about 98.88%.

After many research efforts, the machine's ability to reliably read text is still very low as compare to the human. That's why now OCR research is being done on improving accuracy and performance speed of OCR for various style documents printed/ written in unconstrained environments. There is still not any software available for languages like Urdu or Sindhi etc which are complex.

Machine Learning OCR with Tesseract

Tesseract was originally developed at Hewlett-Packard Laboratories Bristol and at Hewlett-Packard Co, Greeley Colorado between 1985 and 1994, with some more changes made in 1996 to port to Windows. In 2005 Tesseract was open sourced by HP.

The capability of the Tesseract was limited to structured text data. The performance is quite poor with the unstructured text with significant noise. Further it is developed by Google. In 2006, Tesseract was considered one of the most accurate open-source OCR engines then available. The initial versions of Tesseract could recognize English-language text only. Version 2 Tesseract added six additional Western languages (French, Italian, German, Spanish, Brazilian Portuguese, Dutch). Version 3 extended language support significantly to include more languages.[9]

Basically, tesseract is an optical character recognition (OCR) tool for python. Tesseract will recognize and "read" the text present in images. Python-tesseract is a wrapper for Google's Tesseract-OCR Engine. That is useful as a stand-alone invocation script to tesseract, as it can read all image types supported by the libraries like the Pillow and Leptonica imaging, including jpeg, png, gif, bmp, tiff, and others. Additionally, if it is used as a script, Python-tesseract will print the recognized text instead of writing it to a file.

The subset of machine learning, Deep-learning based method performs better for the unstructured data. Tesseract version 4, which consist of deep-learning-based capability with the LSTM network (a kind of Recurrent Neural Network) based OCR engine, which is focused on the line recognition but also supports the legacy Tesseract OCR engine of Tesseract version 3 which works by recognizing character patterns. The latest stable version 4.1.0 is released on July 7, 2019. This version is significantly more accurate on the unstructured text as well. [10].

Methodology

A typical OCR system consists of several steps of processing. The first step in the process is to digitize the analog document using an optical scanner software or device. When the area containing text are located, each symbol is extracted through a segmentation process. The symbols which are extracted may then be pre-processed, eliminating noise, to facilitate the extraction of features in the next step.

These methods can be divided into five distinct steps

1. OPTICAL SCANNING
2. LOCATION AND SEGMENTAION
3. PRE-PROCESSING
4. FEATURE EXTRACTION
5. POST PROSESSING

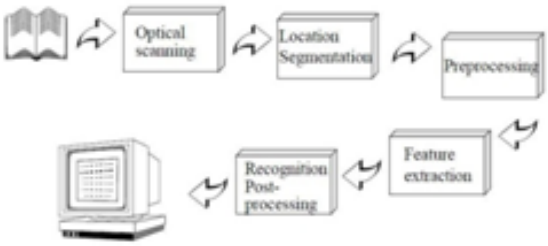

Figure: (OCR processing)

OPTICAL SCANNING

The scanning process is a process of capturing the digital images from the document. Optical scanners are used in the OCR, which generally consists of a transport mechanism plus a Sensor or device that converts light intensity into grey-levels. Printed documents generally consist of black print on a white background. Hence, when performing OCR, it is very common practice to convert the multilevel image into a bilevel image of black and white. This process is known as thresholding which is performed on the scanner to save memory space and computational effort. The thresholding process is important as the results of the following recognition are totally dependent on the quality of the bi-level image. Still, the thresholding performed on the scanner is usually very simple. A constant threshold is used, where grey-levels below this threshold is said to be black and levels above are said to be white. a pre-chosen fixed threshold can be sufficient for a high-contrast document with uniform background. However, a lot of documents used in practice have a rather large range in contrast. In these cases, to obtain good result more sophisticated methods for thresholding are required.[1]

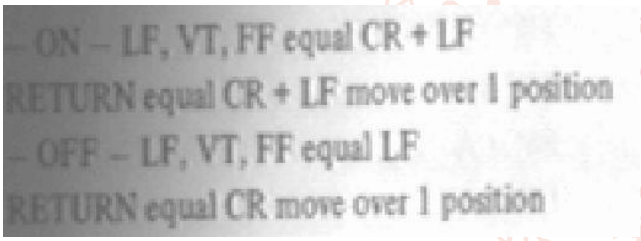

Figure:- (Original grey level image)

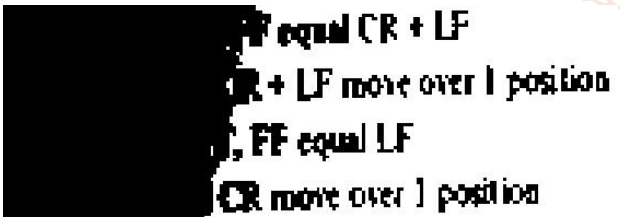

Figure:- (Image threshold with the global method)

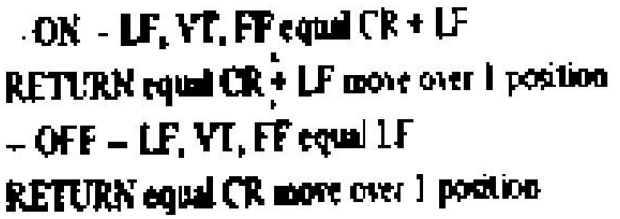

Figure:- (Image threshold with an adaptive method)

The best methods for thresholding are those which are able to vary the threshold over the document adapting to the local properties as brightness and contrast. However, such methods usually depend upon a multilevel scanning of the document which requires computational capacity and more memory. Therefore, such techniques are rarely used in connection with OCR systems, although they result in better images.

LOCATION AND SEGMENTAION

The next step is Segmentation is a process that determines the constituents of an image. It is necessary to identify the regions of the document where data have been printed and distinguish them from figures and graphics. For example, when performing automatic mail sorting through envelopes address must be located and separated from other prints like stamps and logos of organization, earlier to recognition. Segmentation is the isolation of characters or words when applied to the text. Usually, segmentation is performed by isolating each connected component. This technique is easy to implement but problems arise if characters touch or they are fragmented and consist of several parts. The main problems in segmentation are (a) extraction of touching and fragmented characters (b) distinguishing noise from the text (c) misinterpreting graphics and geometry with text and vice versa.[1]

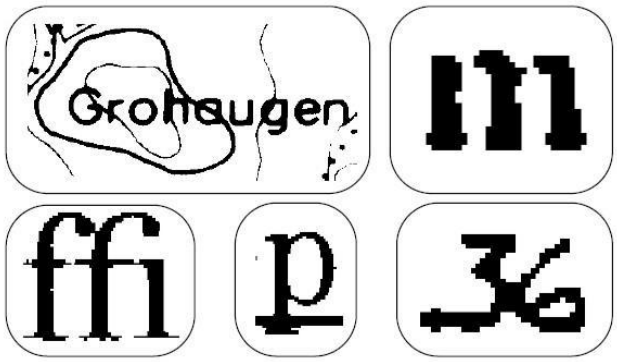

Fig:- (Fragmented symbols)

PRE-PROCESSING

The scanning process of the image may contain a certain amount of noise which depends on the resolution on the scanner and the success of the applied technique for thresholding, the characters may be blur or broken. Some of these defects, which may cause low-quality recognition of the characters from the image, can be eliminated by using a pre-processor to smooth the digitized characters. The smoothing approach consists of both filling and thinning. Where filling eliminates small break downs in characters, gaps between characters and holes in the digitized characters while thinning is used to decrease the width of the line. The most common techniques for smoothing, move a window across the binary image of the character, applying certain rules to the contents of the window. In addition to smoothing, pre-processing usually includes the normalization. The normalization is used to obtain characters of uniform size, slant, and rotation of the image at a 90-degree angle. The angle of rotation is identified for rotation. For rotating the pages and lines of text, Hough transform are commonly used for detecting skew. The Hough transform is a tool which is used to identify lines and other shapes as well. But to find the rotation angle of a single symbol is not possible until after the symbol has been recognized.[1]

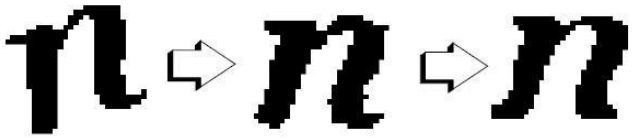

Figure:- Normalization and smoothing of a symbol

FEATURE EXTRACTION

The main function of the character recognition system is the formation of the feature vector to be used in the recognition stage. Feature extraction can be considered as finding a set of features that define the shape of the underlying character as unique as possible. The term features selection refers to algorithms that select the best subset of the input feature set. In this process methods that create new features based on transformations, or a combination of original features is called feature extraction algorithms.[1]

Capturing the essential characteristics of the symbols is the main objective of feature extraction and it is generally accepted that this is one of the most difficult problems of pattern recognition. The simple way of describing a character is by the actual raster image. Another case is to extract certain features that still characterize the symbols, but leaves out the unimportant attributes. The techniques for extraction of such features are often divided into three main groups, where the features are identified from:

- The distribution of points.
- Transformations and series expansions.
- Structural analysis.

POST PROSESSING

After the classification of characters, there are various approaches that can be used to improve the accuracy of OCR results. One of the approaches is to use more than one classifier for the classification of the image. The classifier can be used in cascading, parallel or hierarchical fashion. The results from the classifiers can be combined using various approaches. Where Contextual analysis can be performed to improve OCR results. The geometrical, mathematical and document context of the image can help in decreasing the chances of errors[1]. Other methods like Lexical processing which is based on Markov models and dictionary can also help in improving the results of OCR.

OCR AND MACHINE LEARNING

A large number of uses of machine learning is in the field of optical character recognition (OCR). Optical character recognition has been around for many years; however, the algorithms have become increasingly more accurate and more able to recognize handwritten text. Moreover, as these algorithms have become more sophisticated, there have been a wider variety of applications of optical character recognition. For instance, one common application of optical character recognition is in Google Translate. While most people are familiar with the basic text translate functions, Google Translate also offers real-time visual translation on their mobile app. They are able to do this through their image recognition machine learning algorithms, which utilize deep neural networks.

WORKING OF OCR WITH MACHINE LEARNING

To give a brief description of how this works, Google Translate uses a similar process as other image recognition algorithms. The app first has to detect the letters visible in the camera and isolate them from any other visual stimuli in the background. It does this by using a *neural network* to train the algorithm to identify and recognize patterns of pixel clusters and colours, so that it is eventually able to separate letters from non-letters, and recognize the individual letters. Once the letters have been recognized, they can then be looked up in the dictionary and be translated. The Google Translate app then finally outputs the translated letters on top of the original letters, so that the translate text appears in the camera, where the non-translated text would appear.

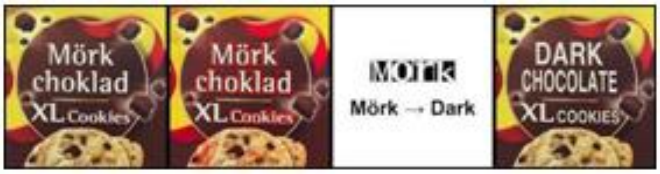

Figure:- (Representation of Google Translate OCR [6])

The Google Translate app demonstrates how optical character recognition can be applied towards very useful situations, but as of right now it is mainly compatible with images of printed text and does not function very well with handwritten text. However, a start-up called omni: us used machine learning to develop an AI which can read handwritten documents. Their services are aimed towards the insurance and banking industry, and they offer products that can process digital documents by extracting handwritten data from paper documents. Omni: us has accumulated a significantly large database of millions of pages, thus improving the accuracy of their algorithm significantly. By utilizing this handwriting recognition service, companies can save a lot of time and resources that they would have put towards data entry. However, one big concern is that due to the nature of the types of paperwork being digitized, such as credit card applications, there would be a lot of personal data being stored in this database, such as social security numbers or even signatures. Furthermore, it is not clear whether customers were informed or gave consent for their handwriting to be fed into the AI. Thus, there are a lot of concerns relating to the privacy and security of the AI used by omni.us.[8]

Both examples demonstrate how optical character recognition can be applied. However, one example of machine learning which has taken handwriting recognition to the next level, is a computer program called "My Text in Your Handwriting", developed by researchers at University College London. The researchers applied machine learning in developing the program so that it is able to take in handwritten script, analyse the input, and then output text in the style of the original person's handwriting. It is able to do this with only a paragraph of handwriting input to analyse. While the computer program is certainly not perfect and requires constant fine-tuning and training, however, in a test study conducted, participants were unable to identify the artificially generated handwriting 40% of the time.

Figure:- (Handwriting output example.[7])

There are many beneficial uses for this technology- for instance, victims who have suffered from strokes can use this technology to be able to "write" letters or "sign" legal documents. This technology thus can allow disabled persons to perform tasks which may have been previously impossible. Additionally, in a more general case, this technology can also be used for everyday purposes, such as producing a "handwritten" note to personalize a gift or a product. However, while there certainly are obvious benefits of having this program, there are some serious concerns too. Like other applications of handwriting recognition algorithms, there are concerns over where the database of handwriting input is being stored and who has access to it. Moreover, a much more pressing issue is related to this specific program. Specifically, if the algorithm or program fell into the wrong hands, it could lead to some unethical uses, such as generating handwriting to forge legal documents, or even generating fake historical documents. Nevertheless, the research team behind the computer program argues that this software could actually help to identify forgery, by using the model and database of handwriting to perform analysis which could help estimate the likelihood of forgery.[7]

CONCLUSION

After all this we can conclude that, although optical character recognition may appear to be simpler than other machine learning algorithms, there are a variety of applications where optical character recognition machine learning algorithms can be used. Where Handwriting is something that is considered fairly personal to oneself, and that can even carry very great importance such as through one's signature. That's why, it is very important that machine learning algorithms which utilize handwriting databases ensure that these databases remain private and confidential, and that people whose handwriting is being fed into these machine learning algorithms are aware that their handwriting is being used. If these security measures are ensured, then there certainly are very promising benefits that can arise from optical character recognition machine learning algorithms, further than what we have seen so far. Considering all this in mind we can develop the more accurate and higher performing OCR model for recognition of characters and text.

REFRENCESES

[1] Arindam Chaudhuri · Krupa Mandaviya Pratixa Badelia. Soumya K. Ghosh 1 3 Optical Character Recognition Systems for Different Languages with Soft Computing

[2] Satti, D. A., 2013, Offline Urdu Nastaliq OCR for Printed Text using Analytical Approach. MS thesis report Quaid-i.

[3] Mahmoud, S. A., & Al-Badr, B., 1995, Survey and bibliography of Arabic optical text recognition. Signal processing.

[4] Sushruth Shastry, Gunasheela G, Thejus Dutt, Vinay D S and Sudhir Rao Rupanagudi, "i"- A novel algorithm for Optical Alphabet Recognition (OCR).

[5] Feng Yanga, and Fan Yangb, "Alphabet Recognition Using Parallel BP Neural Network".

[6] Representation of Google Translate - https://ai.googleblog.com/2015/07/how-google-translate-squeezes-deep.html

[7] Handwriting output- https://www.ucl.ac.uk/news/2016/aug/new-computer-programme-replicates-handwriting

[8] Google Translate article- https://www.dailydot.com/debug/google-translate-real-time-image-recognition-new-languages/

[9] About OCR with tesseract- https://towardsdatascience.com/deep-learning-based-ocr-for-text-in-the-wild-efe4e67b78e4

[10] About the history of Tesseract- https://en.wikipedia.org/wiki/Tesseract