International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9, Issue-7, May 2020

Improve OCR Accuracy with Advanced Image Preprocessing using Machine Learning with Python

Sanjeev Kumar, Mahika Sharma, Kritika Handa, Rishika Jaiswal

Abstract: Optical Character Recognition or Optical Character Reader (OCR) is a pattern-based method consciousness that transforms the concept of electronic conversion of images of handwritten text or printed text in a text compiled. Equipment or tools used for that purpose are cameras and apartment scanners. Handwritten text is scanned using a scanner. The image of the scrutinized document is processed using the program. Identification of manuscripts is difficult compared to other western language texts. In our proposed work we will accept the challenge of identifying letters and letters and working to achieve the same. Image Preprocessing techniques can effectively improve the accuracy of an OCR engine. The goal is to design and implement a machine with a learning machine and Python that is best to work with more accurate than OCR's pre-built machines with unique technologies such as MatLab, Artificial Intelligence, Neural networks, etc.

Keywords : Recognition text, OCR image analysis

I. INTRODUCTION

Optical Character Recognition (OCR) is a common and well-known technique and is the electronic conversion of typed or printed text to machine-readable text with Optical scanner and specially designed software. This software allows the computer to read different text images and convert them into dynamic and interesting data.

OCR usually involves three steps: .

1. To scan a document in OCR software

2. Identify a document in software, too **,**

3. Preservation of the product produced by OCR in the right format for you. **.**

An OCR engine is software that recognizes the text in any given image. Widely used as a way to import data from printed documents whether passport documents, bank statements, business emails, computer receipts,

Revised Manuscript Received on May 30, 2020.

***** Correspondence Author

Kritika Handa*, Computer Science And Engineering Department, ABES Institute Of Technology, Ghaziabad, India. Email: kritikahanda2308@gmail.com

Mahika Sharma, Computer Science And Engineering Department, ABES Institute Of Technology, Ghaziabad, India. Email: mahikas06@gmail.com

Rishika Jaiswal, Computer Science And Engineering Department, ABES Institute Of Technology, Ghaziabad, India. Email: rishikajaiswal32@gmail.com

Prof. Sanjeev Kumar, Computer Science And Engineering Department, ABES Institute Of Technology, Ghaziabad, India. Email: sanjeev.kumar@abesit.in

static data transcripts - or any documents - is a common way of making print documents digital to be edited, searchable, carefully stored .OCR may be a field of research in pattern recognition, AI and computer vision. Multiple methods have been proposed to improve the accuracy of text acquisition. Other methods have made an attempt to make mistakes after finding them.

Kukich[1] suggested using a n-gram dictionary or method based on the errors and returning the possible word to the dictionary using mathematical steps. These methods may reduce the total number of OCR errors in standard language names, but it is possible that the words may be correctly identified that are not in the dictionary of geographical names. The most popular position is to combine multiple OCR output to detect and fix errors automatically [4].Other methods are based on edge detection, binarization, integrated connectivity methods and texture optimization [2] [3]. How to improve installation quality document, thus using a pre- filtering techniques. This provides opportunities for development OCR accuracy is independent of any trained dictionary, specific language details or language model.

II. PROPOSED ALGORITHM

The proposed system uses Tesseract OCR engine while dropping various obstacles in order to increase performance and accuracy.

When it comes to OCR accuracy, there are two ways of measuring it:

A. Access to reading information: In particular, the accuracy of OCR technology is determined by the quality of the letters. How accurate the OCR software is to the characters depends on how often the character is specifically acknowledged compared to how often the character is found incorrectly. 99% accuracy means that 1 in100 characters is uncertain. While 99.9% accuracy means that 1 in 1000 characters is uncertain. OCR accuracy measurement is performed by taking the OCR run image result and comparing it to the original version of the same text.

B. Word level accuracy: To improve the accuracy of word levels, many OCR engines make use of secondary information in the language used in the text. If the language of the text is familiar (e.g. English), the recognized words can be compared to the dictionary of all the expanding words(e.g. all words are in the English language level).[2]Words containing unsupported characters can be “corrected” by finding the word within the dictionary for the highest possible similarity. For this project, we will do the subsequent -

- Focus on increasing the accuracy of character spacing.

- To identify the most reliable characters, a small "correction" at the word level is important.

Tesseract accuracy may increase surprisingly with the right Tesseract image preprocessing tools. If quality of the original source image is good, the result will be more accurate. But the word "Image quality" means we want -

As finer the quality of the original source image, the output will be more accurate. But the word “image quality” means that we want -

- sharp character borders
- marked contrasts
- well orient characters and
- little pixel sound as possible.

III. OCR APPLICATIONS

In recent years, OCR technology has been used throughout the industrial series, revolutionizing the document management process. Because of these people you no longer have to manually rewrite important documents when uploading them to electronic information. Instead, the OCR releases relevant information and enters the environment more precisely and with less time processing

A. Invoice Imaging: It is widely used in many business applications to keep track of financial records by scanning incoming invoices and withdrawing all obligatory information from them, reducing number errors, saving business time and money.

B. Legal Sector**:** In the legal industry, it has also been significant movement to make paper documents digital. To save space and eliminating the need to go through the boxes of paper files, documents are scanned and entered computer information.

C. Bankruptcy: OCR is used to process bank checks in person to intervene. Using OCR, a check can be inserted inside machine, the content has only a signature match-based validation against existing the database is checked and scanned immediately, as well a fair amount of money has been sent which is an effective period. The reduced downtime check permit is an economic benefit of all which is time efficient. A reduced processing time for cheque clearance is an economic gain for all.

C. Health: Health care professionals should maintain it dozens of reports and files and other related documents.

But using OCR, the data can be captured on a computer and stored in a unified place that can be easily accessed.

IV. TYPES OFCHARACTER RECOGNITION

A. Printed: OCR targets handwritten text, one character at a time. This is amazing to use for pattern matching and feature analysis. ICR (Intelligent Character Recognition), on the other hand can also monitor handwritten text and use machine learning techniques.

B. Handwritten**:** The Internet version is very advanced and uses handwriting checks. Instead of using a commonly used algorithm for learning, online mode allows us to capture interests, e.g. The sequence in which the fractions are drawn and the orientation of the strokes. The text is converted to grayscale format. Word recognition is shared outside of the OCR that occurs when the data is accessed by a specific medium, that is includes devices such as cameras, scanners. Due to the variety available in handwriting types, and taking into account the different styles across the globe, it's easy to conclude that the handwritten OCR is challenging to use compared to Printed OCR where the images to be processed have a custom style, similar to the font size written on them. The Tesseract is considered to be an excellent open source OCR engine. The accuracy of the Tesseract OCR is fairly high and can be greatly measured with a well-designed Tesseract pipe**..**Summarized below are the barriers of existing OCR programs:

Due to various issues such as shadows, biasing, blurred lines, dirty marks, overlaps, coffee residues, ink marks and any other missing marks will all reduce the chances of word recognition and accuracy.

This type of modification requires algorithms that are extremely pre-processing and can be featureless, before any type of character extraction is performed

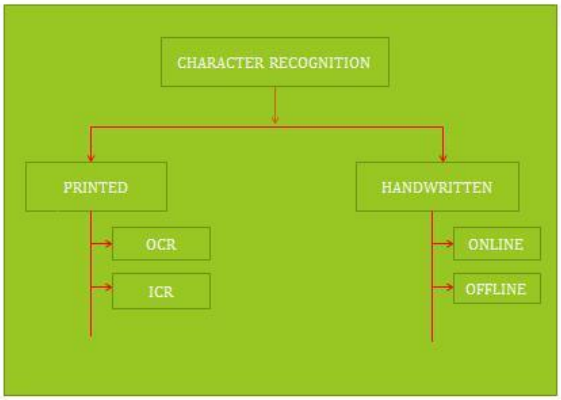

Figure 1. Types of OCR

V. STEPS OF PREPROCESSING

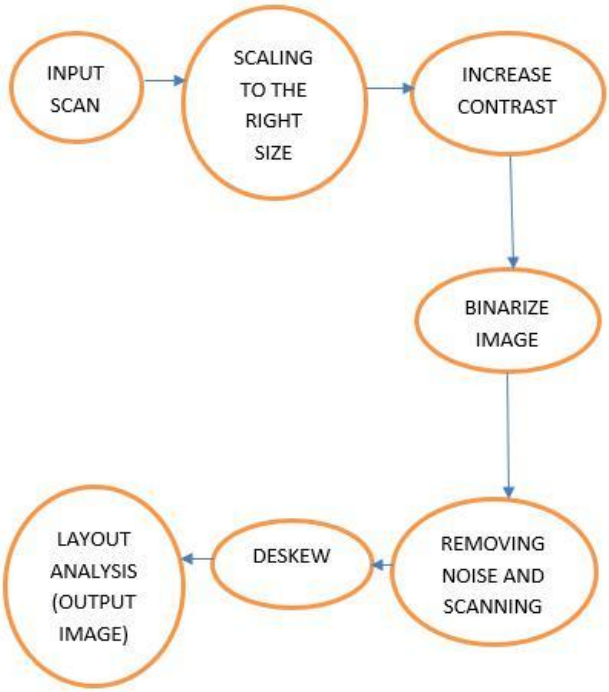

Fig 2: Steps of Preprocessing

A.Input Scan:

The first basic step to get the right OCR conversion is to ensure good quality source images (completely clean source and file source). Ensure that the original document is not damaged, corroded, blotted, printed in ink that does not match any type of marks

B. Scaling To The Right Size:

The next step is to scale the original source image to the correct size which is usually at least 300 DPI (Inch Dots).Setting a DPI below 200 will give spectacular and unimaginable results while a DPI above 600 will increase the output file size without improving the quality of the source file. Therefore, having a DPI of 300 is appropriate for this. Tesseract scans the image pixel by pixel, reads the image with the highest character dimensions above 20 pixels and will increase the computation time. This step also helps eliminate the single biggest obstacle of the Tesseract, where OCR accuracy drops characters below 20 pixels height. As shown in Fig.5 the first image has an average grain height of 12 pixels, which is why the image is limited to a value where the average grain height is 20 pixels. In the example given in Fig.5, the proximity of the direct OCR process is increased from 17.54% to 98.24%, effectively removing the image size. That way solving one of the engine's limitations, such as the Tesseract.

C. Increase Contrast:

If the difference is low, the end result is a poor visualization of the variables. Before using the OCR procedure it is necessary to increase the difference and size. Opt-out will be understandably increased the difference between text / image and its background.

D. Binarize Image:

Binarization is an OCR reconstruction technique where greyscale colors and images can be converted to black and white or monochrome- 1bit images effectively. The main difference between a domain and characters improves its accuracy. The binarization process imposes a color image limit.

Thresholding is the process of converting a pixel color value to a minimum when it is less than a limit or a size if it is greater than the limit value.

There are two types of limit:

- Global Thresholding- In a global mode a certain amount of limit is set for the whole image and not when the image contains different levels of light.
- Adaptive Thresholding- In this the opposite limit is taken into account for smaller image regions. This results in broken letters and noise around the characters, which is why binarization is followed by a blur of medium and twist, which wraps around the gaps. So, Adaptive thresholding one is most popular.

Benefits of Binarization-

Image size is reduced before sending it to an OCR engine is one of the a chance to be torn down.

E. Remove Noise And Scanning:

In this way we remove different types of sound at the same time During scanning or converting into a digital document the images can be contaminated by sound.

We need to be aware of the audio features and search for similar sound patterns in our text image to select the appropriate way to remove that noise.

F. Deskew:

De-skewing text (Skew text correction) is the process of opening a scanned or stray image - which is a blurry image in one place, or of uneven color. It can also be called rotation. This means de-skewing the image to present it in the right format and shape. Text should be horizontal and horizontal at any angle. When a photo is pointed in any direction, we need to leave it in a clockwise or anti-clockwise direction.

Given that the image has a rounded block of text at an unknown angle, we need to make adjustments to :

- Recognizing the block of text in the image.
- Evaluating the angle of the rotated text.
- Rotate the image to deskew.

G.Layout Analysis (Or Zone Analysis):

The process of document analysis searches and finds zones for recognition on the document. The document analysis process searches for and locates recognition sites in document images.

Here's how it works:

Al Analysis Document algorithms detect various basic elements in an image, e.g. names or parts of words, separators, linked parts, color beams, distorted text fields, etc. Based on this information, an icon for these blocks is created and evaluated: What type of block ?, What are the limits of a block? What kind of text can be (magazine, newspaper, book, page)?

VI. SCREENSHOTS

Figure 3. Input Image 1

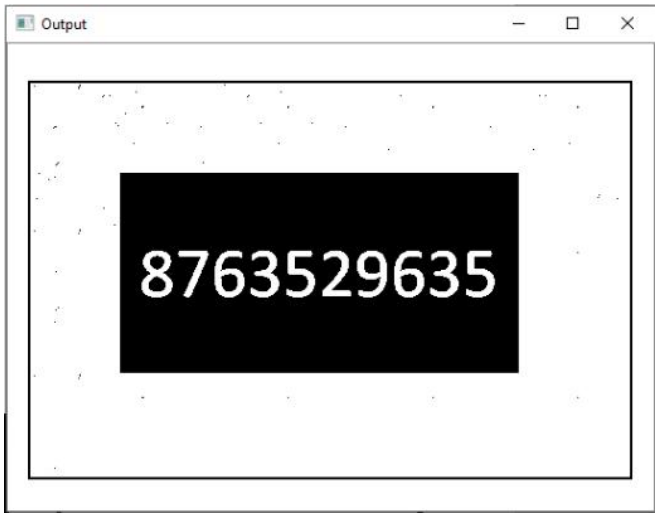

Figure 4. Output Image 1

Figure 5. Input Image 2

Figure 5. Output Image 2 OUTPUT-3

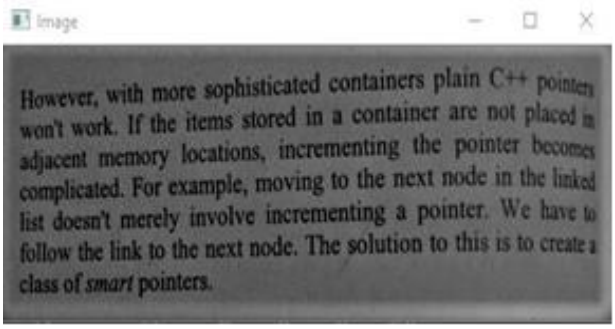

Figure 6. Input Image 3

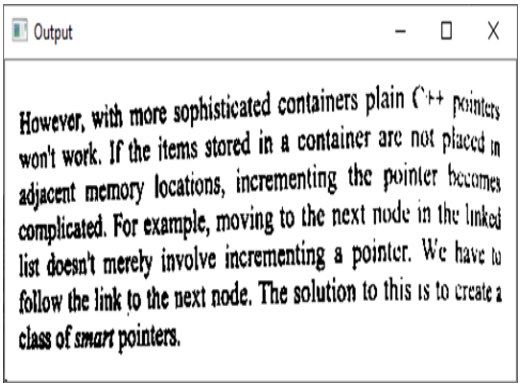

Figure 7. Output Image 3

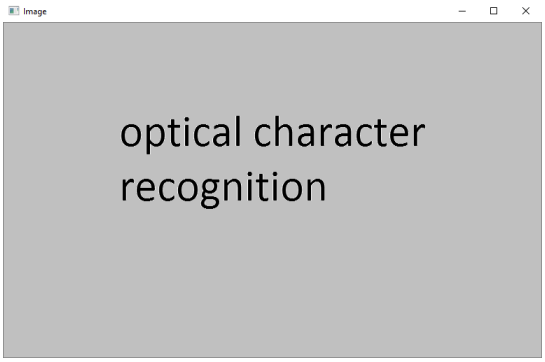

Figure 7. Input Image 4

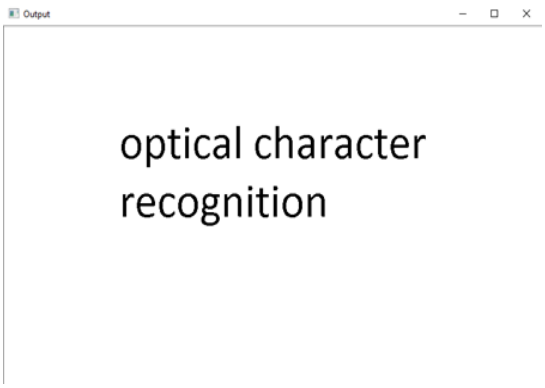

Figure 8. Output Image 4

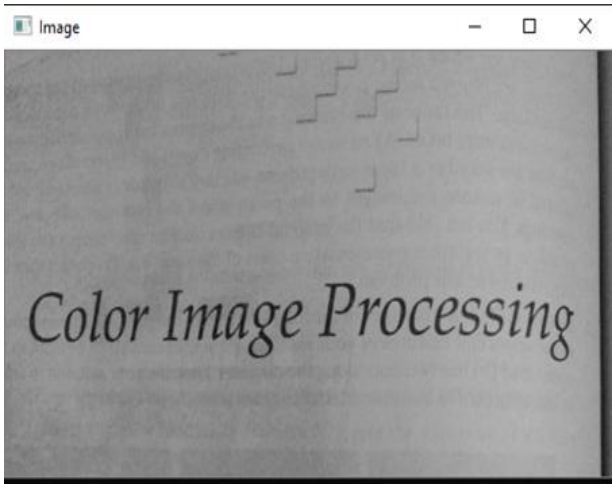

Figure 9. Input Image 5

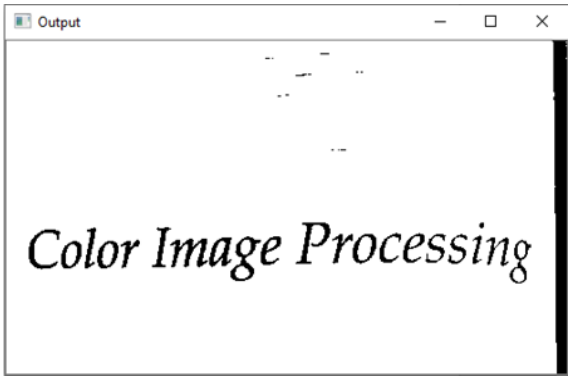

Figure 10. Output Image 5

VII. CONCLUSION

We took into consideration various sample images as the input. These images were either poor in quality, low in contrast or had large amount of noise in them. Our project has successfully enhanced those images and recognized the text between them as a result. The efficiency of our project reaches **99%** when taken into account all sample input images and their results. And thus, we have successfully increased the accuracy of the OCR engine.

REFERENCES

1. Karen Kukich, (1992) "System that automatically converts word into text".
2. N. Ezaki, M. Bulacu, L. Schomaker, (2004) "Text Detection for Visually Impaired Persons".
3. J. Park, G. Lee, E. Kim, J. Lim, S. Kim, H. Yang, M. Lee, S. Hwang, (2010) "Automatic detection and recognition of Korean".
4. Zeki Yalniz "System for automatic alignment for evaluation of books,"(2011).
5. Santosh Kumar Hengel and B. Rama2 "Analysis of Character Recognition Approach
6. SandeepTiwari, Shivangi Mishra, Priyank Bhatia, Praveen Km. Yadav Optical Character Recognition using MATLAB, International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE) Volume 2, Issue 5, May (2013).
7. H.F.Schantz,The History of OCR, Recognition Technology Users Association, Manchester Centre, VT, (2000)
8. Mohamed Fawzi1, Mohsen. A. Rashwan 1, Hany Ahmed 1, shaimaa Samir 1, SherifM. Abdou2, Hassanin M. AI-Barhamtoshy3, and Kamal M. Jambi3 process in Images for Camera-Based OCR Technology , (2015) 13th International Conference on Document Analysis and Recognition (ICDAR).
9. 9. N. VenkataRao, Dr. A.S.C.S.Sastry, A.S.N.Chakravarthy, Pkalyanchakravarthi optical character recognition technique
10. algorithms , journal of theoretical and applied information technology 20th January (2016). Vol.83. No.2 ,issn: 1992-8645 , e-issn: 1817-3195.
11. M.Hamanaka, K. Yamada, J. Tsukumo"On-line Japanese character recognition experiments by an off-line method based on normalization cooperated feature extraction",
12. Proceedings of the Second InternationalConference on Document Analysis and Recognition, (1993) 204-207.
13. Chunmei Liu, Chunheng Wang, Ruwei Dai " Low resolution character recognition by Image quality evaluation", Proceedings of the Eighteenth International Conference on Pattern Recognition ICPR 864-867, (2006).
14. LvXingiao, Dongshan Huang, ENming Song,Ping Li, CHunshan Wu "One Radical-Based on-line character recognition (OLCCR) system using support vector machine for recognition of Radicals", 1stInternational Conference on Bioinformatics and Biomedical Engineering, 558-561, (2007).
15. Hailong Liu, Xiaoging Ding "Recognition of handwritten character" (2005).
16. VeenaBensal, R.M.K. Sinha "Integrating Knowledge sources in Devanagari Text recognition system", IEEE Transact ions on Systems, Man. And Cybernetics-Part A: Systems and Humans, Volume 30, No.4, July(2000).
17. Ganis MD, Wilson CL, Blue JL. Neural network-based systems for handprint OCR applications. IEEE Transactions on Image Processing. 1998Aug;7(8):1097-112.
18. Patel C, Patel A, Patel D. Optical character recognition by OCR tool tesseract:A case study. International Journal of Computer Science. 2012 Jan 1;55(10)
19. Ye Q, Doermann D. Text detection and recognition in imagery: A survey.IEEE transactions on pattern analysis and machine intelligence. 2015 Jul1;37(7):1480-500.
20. Gatos, B., Pratikakis, I., Perantonis, S.J.,"Adaptive degraded document imagebinarization",39, 317–327, 2006.
21. Niblack,W., "An Introduction to Digital Image Process", pp.115–116.Englewood Cliffs, NJ,(1986)
22. Grundland M, Dodgson N, (2007) color to grayscale conversion. PatternRecognition 40:2891–2896
23. Cadik M, (2008) Estimation of color-to-grayscale image conversions.
24. Pratt W, (2007) Digital image process iley- Interscience.
25. Mahmoud, S.A., & Al-Badr, B., 1995,Survey and bibliography of Arabicoptical Text recognition, 41(1), 49-77. 24. Lund, W.B., Kennard, D.J., &Ringger, E.K. (2013). Thresholding Binarization Values to Upgrade OCR Output given in Document Recognition and Retrieval XX Conference 2013, California, USA, 2013. USA: SPIE

AUTHORS PROFILE

Prof. Sanjeev Kumar is an assistant professor in ABES Institute Of Technology, Ghaziabad. He has done M.Tech from USIT GGSIPU Delhi and pursuing Ph.D from IIT (ISM)Dhanbad in thesis Low cost effective and Intelligent authentication protocol for RFID. He has total experience of 13 years in experience research. He has International Certification of ISTE. He is also member of various discipline committee of the college.

Mahika Sharma is currently pursuing Bachelor of Technology in Computer Science and Engineering (CSE) branch from ABES Institute of Technology, Ghaziabad and going to become Computer Science Graduate in the year 2020. At ABES Institute of Technology, she secured second position in CS/IT Quiz(Department wise). She was the event manager of the Sports Committee(Girls) and earned several awards in sports related events. She has knowledge of C language, Java and Machine Learning Using Python.

Kritika Handa is currently pursuing Bachelor of Technology from ABES Institute of Technology, Ghaziabad in Computer Science and Engineering. She has knowledge of C language, Java, Web Technologies and Machine Learning.

Rishika Jaiswal is currently pursuing Bachelor of Technology from ABES Institute of Technology, Ghaziabad in Computer Science and Engineering. She has knowledge of C language, Web Technologies and Machine Learning.