IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

OPTICAL CHARACTER RECOGNITION ENGINE

Shashank Mishra | Shivang Sinha | Anupam Alok | Aditya Saxena| Mrs. Ritu Dewan
Department of Computer Science Sharda University, Greater Noida, India

Abstract- **An OCR engine that scans any document written or printed text and recognizes it and prints it on the screen.**

I. Introduction

Optical character recognition engines scan any document whether it is handwritten, printed, or an image of a text, and gives the output of whatever text is written on that document. Through this, it is much easier to store, read, edit, and format any document and can be used in processing tasks such as language translation and text mining. OCR is the main key towards automation which is the future. This paper gives you an overview on what work has been done in the area of OCR development and what you should expect in the upcoming future. This paper covers many languages to give you a wide variety of work related to this enine.g

II. Problem Identification

OCR engine makes it very easy to store, edit any document online or offline in a compatible way so that it can be easily accessed from anywhere for reading or editing. OCR takes any format of the document even if it is a photo of any text it will recognize it and print it on your screen so that you can read, edit or transfer it easily. To retrieve anything from a scanned image makes documentation very easy and efficient.

III. Literature Survey

1. **2009 By Sandip Rakshit Amitava Kundu Mrinmoy Maity** They worked and developed a Tesseract OCR engine to recognize handwritten Roman numerals. It was extensively tested with the handwritings of many different users. They collected different samples to create a single working model and then it was tested on fresh samples. These samples were trained using a tesseract inbuilt tool called Tesseract. The samples are collected in two different test dataset named as TD1 and TD2, which gave an accuracy of 92.1% for the first and 86.5% for the second one. Rejection rates were high because Tesseract was not able to properly segment the roman digits. It was a major drawback of this system which can be improved by more training of the system.

2. 2015 by Hewlett Packard and the University of Nevada proposed the original design goal was that it should recognize white-on-black text. It follows a step-by-step pipeline procedure. The Tesseract OCR engine was designed from the beginning to be language-independent, but the rest of the engine was developed for EnglishThe biggest advantage of Tesseract OCR is its availability as open code.

3. 2009 By Sandip Rakshit, Debkumar Ghosal, Tanmoy Das, Subharjit Dutta proposed a Developed an OCR Engine to read Bangla language characters. As Bangla has a much larger number of characters as compared to roman, English, or any other scripts, it was difficult to train a system. They used an open-source OCR engine which is Tesseract 2.01 which is licensed under 2.0. They used the tool bbTesseract to train their samples. Different sets were taken into consideration so that this system has a better accuracy rate. when put to test their system gave an accuracy rate of 92.15%, which is very high. It has a major drawback that the system failed many to segment cursive words or under-segment a word that resulted in rejection. which can obviously be made better by more training of the system.

4. 2019 by Mr.Radhe Shayam and his team of members who worked on OCRH(optical character recognition hypertext). OCRH is an upgraded version of OCR. OCRH uses advanced methods of intelligent character recognition. This OCR worked on an Artificial neural network, which was one of its kind. This OCR is working on the tesseract software. For better accuracy, they took several samples into consideration for a better result, These samples are divided into 45% for training and 55% for testing. This format gives an almost 89% accuracy. This method was very distinct as compared to others, so the results were also very favorable and this new approach has also opened many new possibilities.

5. 2013 by Anindita Dey, and Ankita Kumari Character proposed a Recognition System is a system that provides alphanumeric recognition of printed and handwritten characters easily scanned. This technology allows for automatic identification of characters with a viewing machine. This technology is almost complete with the required draft or check, and is clearly accurate to the handwritten claim, although it requires signature verification. In the legal field, there has been an important movement to digitize documents. A new way of extracting features from text images and real character recognition for producing a text document is presented here. The proposed approach promises a much simpler solution to the problem of visual cues.

6.Satti, D.A., 2013, Offline Urdu Nastaliq OCR for Printed Text using Analytical Approach. MS thesis report Quaid-i-Azam University: Islamabad, Pakistan **proposed this In January 2013 Danish Altaf Satti from department of computer science from Quaid-i-Azam University, Islamabad, Pakistan has proposed a novel idea and mechanism for developing Offline Urd Nastaliq OCR for Printed Text using Analytical Approach under the supervision of Dr. Khalid Saleem. Mr Danish Altaf Satti from Quaid-i-Azam University presented his novel igature based analytical approach for recognition or Urdu Nastaliq text. The proposed technique is effective in recognizing cursive scripts like Urdu successfully and it is font-size invariant. In the end it evaluated the proposed technique in a dataset of**

2347 primary Urdu Nastaliq ligatures and found that the system is able to recognize Nastaliq ligatures with an accuracy of 97.10%. This accuracy is really appreciable Danish Altaf Satti presented a novel ligature based analytical approach for recognition or Urdu Nastaliq text.

7. In 2013 Nick trained the Tesseract for Ancient Greek OCR. He used this OCR to develop a digital version of poetry and books. He Trained the tesseract by describing the shapes of character and giving a variety of information on how the language works. For character shape training he specified the coordinated and UTF-8 code-point for each character. For vocabulary training, He divided the wordlist into two lists namely comprehensive list and most frequent word list. There was a use of some new features like an A-weighted wordlist, Explicit support for diacritics, Configuration options. Tesseract works better and with accuracy if the data used for training is large and reliable. As a result of the high flexibility of this tool, it is much easier to train any language or script on this.

8. **2019 By K.Karthick, K.B.Ravindrakumar, R.Francis, S.Ilankannan** proposed In the last decade it was easy to use a mouse and keyboard between us and computers but human based communication and the ability to interact with a machine will surely make things easy to do. This paper is about continuous research between humans and machines. Some of its methods are speech recognition Ocr,omr. The computer vision
and digital image processing are fast growing fields that are essential in many aspects of other areas like multimedia, artificial intelligence, robotics and much more. Image analysis involves the study of segmentation, feature extraction, and classification techniques. Humans interact quite naturally with each other. We have also discussed the handwritten ocr systems classification and steps required for automatic identification technique.

9. 2017 ,Review on tesseract ocr engine and performance by Aseem thakkar and Prof Vatsal Shah proposed thatTesseract is an open source ocr engine which has outperformed other commercial ocr engines at 4th unlv annual test of ocr accuracy. The Tesseract engine takes input of binary image with optionally defined polygonal text regions. The processing follows a simple pipeline, but the techniques that are used at different stages set it apart from the OCR engines of its time It was developed by HP as an add-on to flatbed scanner where it failed miserably. It vanished for 10 years after that hp made it an open source . It was taken over by google when it was released as open source in 2005 but the code has changed alot since then which includes converting it to unicode.

10. 2020 By Shivani Sharma and Arun Mittal They proposed a system that was developed on Tesseract. In the past we have seen that many of the OCR developers have used Tesseract because it is one of the most reliable engines because it gives you a high accuracy rate. as the new version of tesseract is on the way tesseract 4.0 which is creating a lot more hype in society. Shivani and Arun developed their engine in several steps namely- Input of scanned image, Scaling of the right size,

Increase contrast, Binarize the image, Removing noise, Deskew, and then final output. The benefit of the proposed was that it not only takes the input image and gives you output text but also does searching and editing.

10. 2020 By Shivani Sharma and Arun Mittal They proposed a system that was developed on Tesseract. In the past we have seen that many of the OCR developers have used Tesseract because it is one of the most reliable engines because it gives you a high accuracy rate. as the new version of tesseract is on the way tesseract 4.0 which is creating a lot more hype in society. Shivani and Arun developed their engine in several steps namely- Input of scanned image, Scaling of the right size, Increase contrast, Binarize the image, Removing noise, Deskew, and then final output. The benefit of the proposed was that it not only takes the input image and gives you output text but also does searching and editing.

11. Kahan, S. Pavlidis, T. Baird, H. S., "On the Recognition of Printed Characters of Any Font and Size," IEEE Trans. Pattern

Analysis and Machine Intelligence IEEE (2012) they come to this conclusion because we are researching various journals published
in IEEE Transaction on Pattern Analysis and Machine Intelligence. Publisher is IEEE. This OCR tool describes the current state of the selected system that detects printed fonts and various sizes of Roman characters. The program incorporates many other strategies to improve overall visibility. Altho. the results they get are good. Weight loss will be followed by fatigue and constant tiredness. The emerging clobber and other shapes are mapped, using a stand-alone approach, to the banner features provided for inclusion in the Base mathematical category. Here the binary features are the ones that can be valued in one of two ways (plus or minus), which have been found to be useful in producing classification systems. Major experiments have shown better than 97 percent (which is good to achieve) in choosing the right combination of six different fonts, and over 99 percent in most single fonts, in the range of multiple points.

12. Och, F. J., "Minimum error rate training in statistical machine translation," Proc. 41st In 2003,Annual Meeting on Association for Computational Linguistics, 160-167 **they come to this issue published in the run-up to the 41st annual meeting of the Association for Computational Linguistics.The training process for modeling machine translation models is based on a large base or a related process. Kernel-based learning (e.g., Support Vector Machines) has been used successfully in many complex problems in Natural Language Processing (NLP). In NLP, although a combination of factors is important in improving performance, it is selected heuristically. Kernel's methods change this situation. The advantage of kernel methods is that the active component combination is fully extended without the usual losses and increases computational costs. Kernel-based text analysis shows excellent performance with accuracy; however, these methods are usually very slow to use in large-scale literature analysis. Word Segmentation and Japan Dependency Parsing show that our new classifiers are almost 30 to 300 times better than kernel-based classifiers.**

13. IN 2015 Muna Ahmed Awel proposed Character recognition, typically abbreviated to optical character recognition or shoFened OCR, is that the mechanical or electronic translation of pictures of written, typed or written text (usually captured by a scanner) intcygiachine-editable text although educational analysis within the field continues, the main focus on character recognition has to implement verified techniques. Optical character recognition technology was unreal within the early

1800 s, once it had been proprietary as reading aids for the blind. In 1870, C. R. Carey proprietary a picture transmission system victimisation photoceLLs,The 1980's saw the emergence of OCR systems supposed to be used with personal Computers. Nowadays, it's common to realize PC-based Morato develop optical character recognition system victimisation neural networks for Hindi characters and trained with one thousand dataset.In the analysis works revised in this paper, character recognition systems use completely different approaches and several of them get sensible accuracy. What we will perceive from this paper is feature extraction techniques ought to be select per the character you operating as a result of every scripts or alphabets has its own nature therefore want to realize techniques that work or appropriate for characters.The higher able to extract options from character additionally we will sight and acknowledge characters in highest accuracy. The higher ability to extract options from character Additionally we will sight and acknowledge characters in h@hest accuracy. Our future work can target optimizing this recognition results by new approaches for segmentation
and new forms of options.This methodology may be applied to either machine written.

14. IN 2014 --- RAJU KUMAR SHAH AND
PARUL MEHTHA proposed Character recognition technique has been completed through studies on completely different characters as an example, English, Arabic, Chinese, Devanagari, Bangla, Farsi and kannada and then on. Totally, the whole technique is applied in 3 section Preprocessing. It's common to realize PC-based OCR systems that are unitically accessible. However, most of tI}gse systems are unit developed to figure with Latin-based scripts. Optical character recognition systems for Latin characters are accessible for over a decade.There area unit analysis has conjointly What we will perceive from this paper is feature extraction techniques ought to be select in step with the character you operating as a result of every scripts or alphabets has its own nature therefore would like to realize
.techniques that match or appropriate for characters.What we will perceive from this paper is feature extraction techniques ought to be selected in step with the character you operate as a result of every script or alphabet has its own nature therefore would like to realip techniques that match or appropriate for characters. been directed at different non-Latin scripts like Arabic, Japanese, Chinese, Hindu, llbean. In order to develop AN OCR system it needs the event and integration of the many subsystems. The beginning is preprocessing like skew detection and correction, noise detection and removal, thinning, and normalisation.Then segmentation of document pictures into line word and characters. This is followed by feature extraction for representing character pictures and a categorification module that label characters to their correct class. Finally, This is one of the riskiest parts in AN OCR development. The aim is to extract vital patterns from characteristics. The chosen options area unit is expected to contain patterns that differentiate one character from different and relevant data from the input knowledge, once the input document area unit is terribly recent and harmful whether or not we tend to take it in camera or scanned can be terribly troublesome to look at the character.

15. In 2014-- N. Stamatopoulos, and S.J.Perantonis proposed the writing history measure is primarily focused on record recovery. Word-spotting procedures for watching out and characterization will probably gracefully reasonably acknowledge correctness that adjust movement recovery of composed pages from a client provided code question.transforming entire chronicled reports into text records could be a somewhat more irksome assignment. To the best of our information, there don't appear to be stores of works following this methodology. Besides, all associated works will in general be focused on the particular qualities of the
comparing history they technique, similar to substance and style In this paper, Associate in Nursing disconnected acknowledgment framework for either machine composed of composed authentic archives is offered. It comprises a pre-handling stage any place reports square measure recovery into

parallel pictures, a high — down division procedure that removes the characters, the production of an info.Our future work can have practical experience in improving this acknowledgment results by misusing new methodologies for division and new styles of choices.

16. 2014 ,IMPROVING THE EFFICIENCY OF TESSERACT OCR ENGINE by Sahil badla **proposed that this current venture's fundamental goal is to research standards of ocr on tesseract motor and to improve its effectiveness and runtime Ocr is utilized to change over printed text into editable content in different application over an assortment of gadgets, for example, scanners,pc,tablets As cell phones are assuming control over the PCs in all spaces however ocr still remaining part not all that refined field This paper centers fund improving tesseract ocr proficiency** for the Hindi language to run on cell phones. This paper presents a preprocessing method on tesseract to improve acknowledgment of characters keeping the runtime low Making it running easily and proficiently on cell phones.

17. 2020 A study on the Accuracy of OCR Engines proposed that This current venture's fundamental goal is to research standards of ocr on tesseract motor and to improve its effectiveness and runtime Ocr is utilized to change over printed text into editable content in different application over an assortment of gadgets, for example, scanners,pc,tablets As cell phones are assuming control over the PCs in all spaces however ocr still remaining paxs one not all that refined field This paper enters around improving tesseract ocr proficiency for the Hindi language to run on cell phones. This paper presents a preprocessing method on tesseract to improve acknowledgment of characters keeping the runtime low Making it running easily and proficiently on cell phones.

18. Bhansali, M., & Kumar, P, 2013 proposed an Alternative Method for Facility Check Clearance Using Smart Phones Application.Global Journal of Application or Innovation in Engineering andManagement - We came to this diary from International Journal of Application or Innovation in Engineering and Management (IJAIEM). In 2013 Mohit Bhansal from Department of Computer @ience and Engineering, Amity University and Praveen Kumar 2Assistant Professor, Department of Computer Science and Engineering, Amity University Noida had accompanied a thought for building up an application which turns into an elective
strategy to abbreviate the extensive cycle of storing actual duplicate of the checks in banks and its leeway making the whole check clearing measure a lot less complex and speedier. OCR (optical character acknowledgment) innovation is utilized to peruse the MICR data from the picture, and in the last advance is proportional to a 200 DPI goal , changing for any non-even goal identified. The back pictures are prepared in the equivalent with the exception that a goal scaling is resolved from the añereflects of handling the front picture, as there is no MICR data to measure on the back. The outcome is a picture trade consistent TIFF picture. We found that ACPS inside a general portable financial contribution furnishes money with numerous appealing advantages.

19. 2019 By K.Karthick, K.B.Ravindrakumar, R.Francis, S.llansannan proposed that the most recent decade it was anything but difficult to utilize a mouse and console among us and PCs however human based correspondence and the capacity to associate with a machine will without a doubt make things simple to do. This paper is about constant examination among people and machines. A portion of its techniques are discourse acknowledgment Ocr,omr. The PC vision and advanced picture handling are quickly developing fields that are basic in numerous parts of different territories like interactive media, man-made reasoning, mechanical technology and significantly more. Picture investigation includes the investigation of division, highlight extraction, and grouping strategies. People communicate normally with one another. We have additionally examined the transcribed ocr frameworks order and steps needed for programmed distinguishing proof procedure.

IV. Conclusion

When the Tesseract engine was taken over by Google, it made a transition. Nowadays it is of the most accurate open source OCR engines available and it has been growing ever since then. The Tesseract is compatible with platforms such as Linux, Windows and Mac OSX. Its source code can also be compiled for other platforms, including Android and iPhone .The Tesseract Supports 149 languages for optical character recognition. The advantages of using this engine are that it supports many different languages on many different While running it is capable of achieving high accuracy such as 95% but in the case of some complex images having multi layered backgrounds or fancy text. It provides better accuracy in results if the pictures are in the grayscale mode as instead of color. There is a tradeoff between processing and accuracy. The more time we invest on preprocessing then we get more accuracy but it increases the runtime too.