INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

Handwritten Optical Character Recognition (OCR)

JREVANTH
B-Tech Student

G HEMANTH
B-Tech Student

G RANJITH
B-Tech Student

G MEANAKA
(Assistant Professor)

Department of Information Technology
CMR Technical Campus
Kadlakoya (V), Medchal, Hyderabad-501401.

Abstract: The objective of the current project is to develop an Optical Character Recognition (OCR) engine for information Just in Time system that can be used to convert the handwritten textual image to text. Tesseract is open-source OCR engine used to develop user-specific handwriting recognition. The offline handwritten characters recognition has been one of the most challenging problems. In this project, an offline handwritten character recognition tool has been developed based on the Tesseract open-source OCR engine.

I. INTRODUCTION

In online character recognition, the trajectories of pen tip movements are recorded and analysed to identify the linguistic information expressed. With the latest technological advancements in pen input devices, new interfaces are designed to capture the precise pen-trajectory information and subsequent analysis of online handwritten data, with user comforts in writing. It is now possible to write on an ordinary paper and immediate wireless transmission of handwritten annotations to a remote server . With these technological advances, handwritten annotations in digital notebooks may be digitized in no time. This may, in due course, generate huge information archive with basic requirements of in-time retrieval of relevant handwritten annotations through typed/handwritten query. Traditionally, documents containing handwritten information are difficult to archive in digital form. Even with the help of latest optical scanners, content based indexing techniques and research tools; it is difficult to find digitized versions of document pages based on user queries. Robust techniques for indexing and retrieval of handwritten information are needed for various applications involving historical manuscripts, scientific notes, personal records, criminal records etc. In most of these applications there is a need for indexing and retrieval based on textual content. To address this challenge, we present a recognition-based approach for indexing and retrieval of handwritten annotations in document pages. Some work has recently been done on content-based retrieval of handwritten documents. Bertrand et.al. have developed a technique for structural document recognition and recognition of handwritten names. In another work, Matthew et.al.developed a stroke feature based technique for retrieval of handwritten Chinese annotations based on typed/handwritten query. Srihari et.al. had used stroke/shape features for indexing and retrieval of handwritten documents based on writer characteristics, textual content and writer profile. However, recognition-based indexing of handwritten annotations is still an open problem for the researchers. Efficient techniques for approximate string matching need to be developed for real-time retrieval of handwritten annotations based on typed/handwritten query. David Doermann, in his survey [has highlighted key issues involved in indexing and retrieval of document images. One of the possible approaches for the said problem was addressed by Han Shu using Hidden Markov Models. In this paper we have discussed a potential solution for indexing/retrieval of

ISSN: 2057-5688

handwritten annotations using q-gram based approximate string-matching technique and subsequent optimization through Hidden Markov Model. In the following section of this paper, we have discussed an overview of the existing iJIT system, the digital pen and paper technology used in this work and the key motivations behind the current experimentation. In the subsequent sections, we have discussed the principal design issues involved in the development of an effective handwritten annotations retrieval system based on typed/handwritten query.

II. LITERATURE & SURVEY

Popular OCR Applications in the Real World OCR has widespread applications across industries (primarily with the aim of reducing manual human effort). It has been incorporated in our everyday life to an extent that we hardly ever notice it! But they surely strive to bring a better user experience. OCR is used for handwriting recognition tasks to extract information. A lot of work is going on in this field and we have made some really significant advancements. Microsoft has come up with an awesome mathematical application that takes as input a handwritten mathematical equation and generates the solution along with a step-by-step explanation of the working. OCR is increasingly being used for digitization by various industries to cut down manual workload. This makes it very easy and efficient to extract and store information from business documents, receipts, invoices, passports, etc. Also, when you upload your documents for KYC (Know Your Customer), OCR is used to extract information from these documents and store them for future reference. OCR is also used for book scanning where it turns raw images into a digital text format. Many large-scale projects like the Gutenberg project, Million Book Project, and Google Books use OCR to scan and digitize books and store the works as an archive. The banking industry is also increasingly using OCR to archive client-related paperwork, like onboarding material, to easily create a client repository. This significantly reduces the onboarding time and thereby improves the user experience. Also, banks use OCR to extract information like account number, amount, cheque number from cheques for faster processing. The applications of OCR are incomplete without mentioning their use in self-driving cars. Autonomous cars rely extensively on OCR to read signposts and traffic signs. An effective understanding of these signs makes autonomous cars safe for pedestrians and other vehicles that ply on the roads. There are definitely many more applications of OCR like vehicle number plate recognition, converting scanned documents into editable word documents, and many more. I would love to hear your experience of using OCR – let me know in the comments section below. The digitization using OCR obviously has widespread advantages like easy storage and manipulation of the text, not to mention the unfathomable amount of analytics that you can apply to this data! OCR is definitely one of the most important fields of Computer Vision.

III. IMPLEMENTATION

ARCHITECTURE

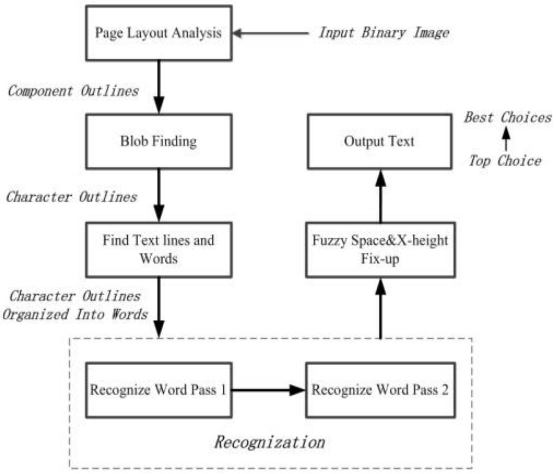


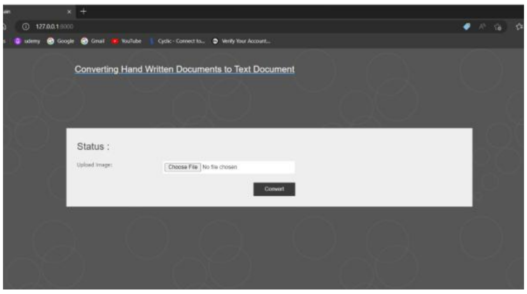
Figure 6. Process flow of the Tesseract engine.

Modules Description

User: In this module user can upload handwritten image to system.

System: Recognizing the Characters In this step, the black areas are further processed to identify letters or digits. Usually, an OCR focuses on one character or block of text at a time. The recognition of characters is carried out by using one of the following two types of algorithms: Pattern recognition. The pattern recognition algorithm involves inserting text in different fonts and formats into the OCR software. The modified software is then used for comparing and recognizing the characters in the scanned document. Feature detection. Through the feature detection algorithm, OCR software applies rules considering the features of a certain letter or number to identify characters in the handwriting image.

IV. RESULT

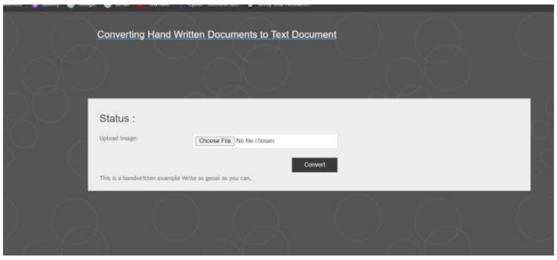

127.0.0.1:8000

Converting Hand Written Documents to Text Document

Status :

Upload Image: No file chosen

Fig-3.2: Home Page


Converting Hand Written Documents to Text Document

Status :

Upload Image: No file chosen

This is a handwritten example Write as geoal as you can.

Fig-7.3: output screen

V. CONCLUSION

The performance of the proposed system depends on the character recognition module (tesseract-ocr) engine. In the present work, we have developed Tesseract based system to meet the requirements for the current work recognition of handwritten annotations. The performance of the system was satisfactory.

REFERENCES

1. Wei Lu, ZhijianLi,Bingxue Shi . ”Handwritten Digits Recognition with Neural Networks and Fuzzy Logic” in IEEE International Conference on Neural Networks, 1995. Proceedings.,
2. P. Banumathi , Dr. G. M. Nasira .”Handwritten Tamil Character Recognition using Artificial Neural Networks” in 2011 International Conference on Process Automation, Control and Computing
3. B. V. S. Murthy. ”Handwriting Recognition Using Supervised Neural Networks” in International Joint Conference on Neural Networks, 1999. IJCNN ’99.
4. J.Pradeep, E.Srinivasan, S.Himavathi. ”Neural Network based Handwritten Character Recognition system without feature extraction” in International Conference on Computer, Communication and Electrical Technology ICCCET 2011

5. Shun Nishide, Hiroshi G. Okuno, Tetsuya Ogata, Jun Tani. ”Handwriting Prediction Based Character Recognition using Recurrent Neural Network” in 2011 IEEE International Conference on Systems, Man, and Cybernetics
6. T. Kohonen, Self-Organization and Associative Memory, 2nd Ed., New York, Springer, 1988.
7. Y. Yamashita and J. Tani, Emergence of Functional Hierarchy in a Multiple Timescales Recurrent Neural Network Model: A Humanoid Robot Experiment, PLoS Computational Biology, Vol. 4, e1000220, 2008.
8. Lei Wang, Lei Zhang , Yanqing Ma . ”Gstreamer Accomplish Video capture and coding with PyGI in Python language” in 2017 24th AsiaPacific Software Engineering Conference (APSEC)
9. Rahul R. Palekar , Sushant U. Parab , Dhrumil P. Parikh , Prof. Vijaya N. Kamble. ”Real Time License Plate Detection Using OpenCV and Tesseract” in International Conference on Communication and Signal Processing
10. ”OpenCV” https://en.wikipedia.org,[Online] Available: https://en.wikipedia.org/wiki/OpenCV/. [Accessed 05 March 2018]
11. [1 FatihErtam, GalipAydOn. ”Data Classification with Deep Learning using Tensorflow” in 2017 International Conference on Computer Science and Engineering (UBMK)
12. ”An open-source machine learning framework for everyone” https://www.tensorflow.org/,[Online] Available: https://www.tensorflow.org/. [Accessed 05 March 2018]
13. Rohan V. Vaidya , Darshan K. Trivedi. ”M-health : A complete healthcare solution” in 2017 International Conference on Computing Methodologies and Communication (ICCMC)
14. ”NIST Special Database 19” https://www.nist.gov, [Online]. Available: https://www.nist.gov/srd/nist-special-database-19. [Accessed 05 March 2018].
15. SihamTabik, Daniel Peralta, Andrs Herrera-Poyatos, Francisco Herrera. ”A snapshot of image Pre-Processing for convolutional neural networks: Case study of MNIST” in International Journal of Computational Intelligence Systems 10(1):555 January 2017
16. Vikas J Dongre , Vijay H Mankar. ”DEVNAGARI DOCUMENT SEGMENTATION USING HISTOGRAM APPROACH” in International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.1, No.3, August 2011.