International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT12173177

Structured Detail Extraction from Government Documents

Shishir Kallapur, Shourya Sinha, Vinay Kumar, Shashank Singh
Department of Computer Science and Engineering, National Institute of Engineering, Mysore, India

Article Info
Volume 7, Issue 3
Page Number: 610-613

Publication Issue :
May-June-2021

Article History
Accepted : 20 June 2021
Published : 30 June 2021

ABSTRACT

Text recognition has been one of the most active and challenging research areas in the field of image processing and pattern recognition. This paper presents a method to extract various important data from government documents and store it in a JSON file and link different documents of a person so that different data can be used when required. The approach we have taken here is that first the user will have to upload the scanned copy of the document into the GUI and the data is extracted from the photo. We made sure to capture all the necessary data from the image so that even if some of the data in the image is in circular shape, the software captures all the data and from the captured data, we select the important data and store it in a JSON file.

Keywords: GUI, JSON

I. INTRODUCTION

We propose a simple approach to efficiently extract the text details contained in different Government Document Images, link them together and store it in a JSON file so that it can be used when required. When the user opens the software, he/she will have an option to upload the document image and an option to choose the type of document uploaded. Once the image has been uploaded the pre-processing of the image takes place. There are various processes in the pre-processing stage. After the pre-processing stage the cropped image, containing all the important data of the document is passed through Tesseract OCR Engine. The main function of Tesseract OCR is to convert the input image to string. This is made possible by training data against various feature extraction algorithms. Once the image has been converted to string, we apply various different methods to clean the text so that the text contains the accurate data which has been extracted from the image. Now that all the important data has been extracted, we structure the text and sort it according to our need and convert it to a JSON file so that it can be used when required. Further the data is stored in a .csv file for structured viewing and modification of the extracted data

II. LITERATURE REVIEW

Character recognition is not a new problem, but its roots can be traced back to systems before the inventions of computers. The earliest OCR systems were not computers but mechanical devices that were able to recognize characters and had very slow speed and low accuracy. In 1951, M. Sheppard

invented a reading and robot GISMO that can be considered as the earliest work on modern OCR. GISMO can read musical notations as well as words on a printed page one by one. However, it can only recognize 23 characters. The machine also has the capability to could copy a typewritten page. J. Rainbow, in 1954, devised a machine that can read uppercase typewritten English characters, one per minute. The early OCR systems were criticized due to errors and slow recognition speed. Hence, not much research efforts were put on the topic during 60's and 70's. The only developments were done on government agencies and large corporations like banks, newspapers and airlines etc.During the past thirty years, substantial research has been done on OCR. This has led to the emergence of document image analysis (DIA), multi-lingual, handwritten and omni-font OCRs and other such applications. Despite these extensive research efforts, the machine's ability to reliably read text is still far below the human. Hence, current OCR research is being done on improving accuracy and speed of OCR for diverse style documents printed/ written in unconstrained environments. There has not been availability of any open source or commercial software available for complex languages like Urdu or Sindhi etc.

III. PROBLEM STATEMENT

Humans are bound to make errors, some time or the other, especially while performing mundane and boring tasks like digitization or data entry for various important documents (govt. documents). Many times, we are unable to perceive certain digits and characters correctly due to various reasons – lack printed clarity, motion, illumination and so on. These are the primary problems that we intend to address.

IV. SOLUTION

OCR(Optical Character Recognition), involves a computer system designed to translate images of typewritten text (usually captured by a camera/scanning device) into a machine editable text or to translate pictures of characters into a standard encoding scheme representing them.Extraction of important text details from government documents could reduce the tedious office work for government officials and various organisations.For example if a person has to apply for a new government issued scheme then the official can just upload a photo into the software and the software will do the rest of filling the information. The Human error will be resolved.

V. EXISTING SYSTEM

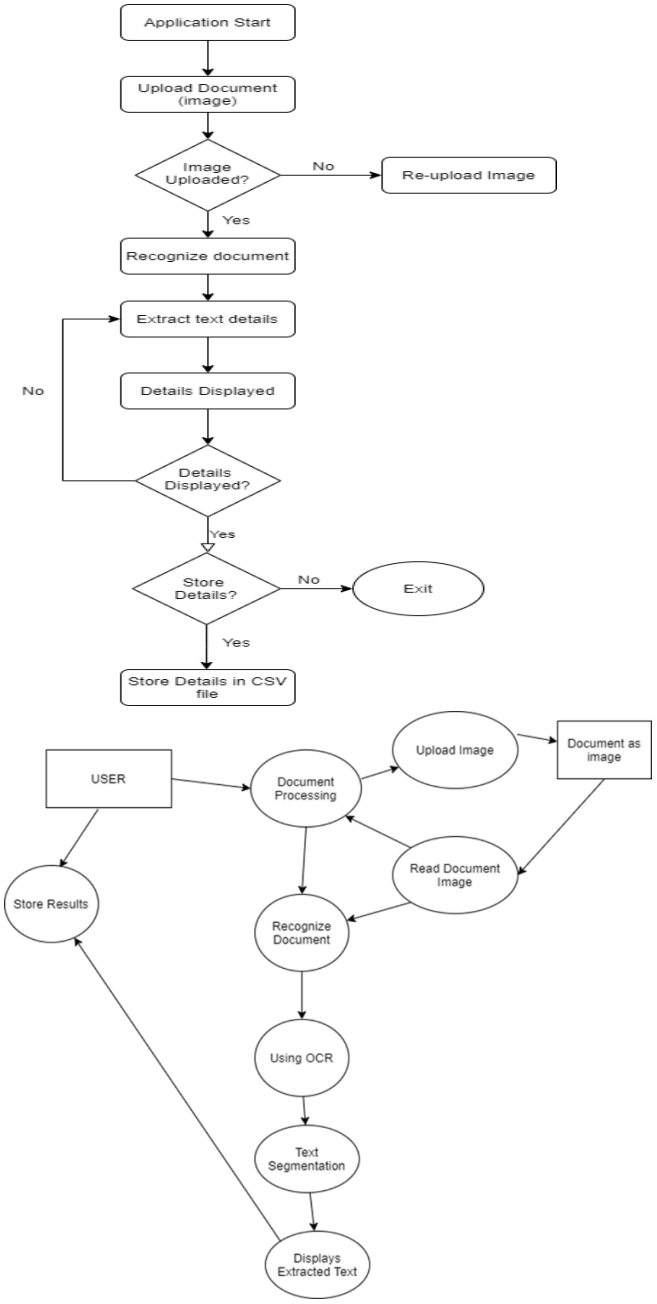
There is a growing demand for users to convert printed documents to electronic documents to maintain the security of their data. In the beginning there was no technology to convert image to text so all conversion would be done through humans. Then OCR was developed to convert characters to text. The efficiency of the existing OCR systems needs to be enhanced which improve the image readability for the system. Also the current systems lack in accurately extracting details positioned at a skewed angle.The Existing system deals with only homogeneous characters or characters of a single language.

VI. PROPOSED SYSTEM

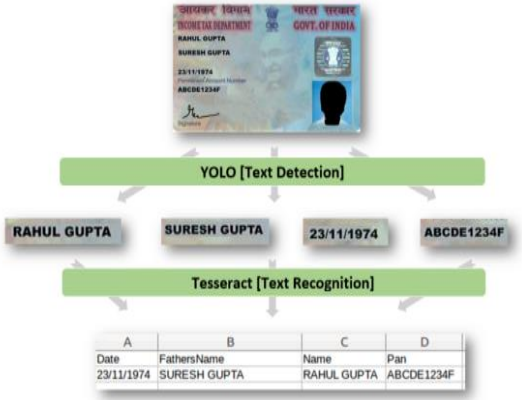
Our proposed system will cater to all the drawbacks in the existing system.The proposed system would be multilingual so the system would detect text in languages other than English. The efficiency of character recognition will be increased so that more characters can be recognized accurately. The software will also have the algorithm to recognize

characters which are skewed or at an angle so that it corrects the angle and recognizes the character.

VII. SYSTEM DESIGN AND IMPLEMENTATION



When the end user tries to extract text details from the document, the first phase through which the system runs is the document processing phase in which the document is pre- processed and made ready for the extraction. The document is uploaded as an image which is then read by the system and the necessary pre-processing functions are executed internally (pre-processing enables the system to produce more accurate results).The image-data is recognized and made to go through the Tesseract engine. In the Tesseract engine the data is trained against various feature extraction algorithms which extracts the text. After that the extracted data goes through a segmentation method in which it is tested against many constraints and regular expression to rule out non useful or gibberish text. After extracting and constructing it into meaningful data, the details are displayed. The end user can choose to store the details in a database file (JSON, CSV, SQL) or just discard it and extract another document detail.



VIII. ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible whose constant guidance and encouragement crown all efforts. First and foremost, we would like to thank our beloved principal Dr Rohini Nagapadma for being the patron and the beacon light for this project. We would like to express our sincere gratitude to Dr.V.K. Annapurna, HOD, Department of CSE, NIE for his relentless support and encouragement. It gives us immense pleasure to thank our guide Mr. Abhinandan S.P., Assistant Professor, Department of CSE, NIE and Mrs. Rashmi M.R., Assistant professor, Department of CSE, NIE for their valuable suggestions and guidance

during the process of the seminar and for having permitted us to pursue work on the subject.

IX. REFERENCES

[1]. S. V. Rice, F.R. Jenkins, T.A. Nartker, The Fourth Annual Test of OCR Accuracy, Technical Report 95-03, Information Science Research Institute, University of Nevada, Las Vegas, July 1995

[2]. R.W. Smith, The Extraction and Recognition of Text from Multimedia Document Images, PhD Thesis, University of Bristol, November 1987

[3]. R. Smith, "A Simple and Efficient Skew Detection Algorithm via Text Row Accumulation", Proc. of the 3rd Int. Conf. on Document Analysis and Recognition (Vol. 2), IEEE 1995, pp. 1145-1148

[4]. P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, Wiley-IEEE, 2003.

[5]. S.V. Rice, G. Nagy, T.A. Nartker, Optical Character Recognition: An Illustrated Guide to the Frontier, Kluwer Academic Publishers, USA 1999, pp. 57-60.

[6]. P.J. Schneider, "An Algorithm for Automatically Fitting Digitized Curves", in A.S. Glassner, Graphics Gems I, Morgan Kaufmann, 1990, pp. 612-626.

[7]. R.J. Shillman, Character Recognition Based on Phenomenological Attributes: Theory and Methods, PhD. Thesis, Massachusetts Institute of Technology. 1974.

[8]. B.A. Blesser, T.T. Kuklinski, R.J. Shillman, "Empirical Tests for Feature Selection Based on a Pscychological Theory of Character Recognition", Pattern Recognition 8(2), Elsevier, New York, 1976.

[9]. G. Nagy, "At the frontiers of OCR", Proc. IEEE 80(7), IEEE, USA, Jul 1992, pp 1093-1100.

[10]. H.S. Baird, R. Fossey, "A 100-Font Classifier", Proc. of the 1st Int. Conf. on Document Analysis and Recognition, IEEE, 1991, pp 332-340.

Cite this article as :

Shishir Kallapur, Shourya Sinha, Vinay Kumar, Shashank Singh, "Structured Detail Extraction from Government Documents", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 3, pp. 610-613, May-June 2021. Available at

doi : https://doi.org/10.32628/CSEIT12173177

Journal URL : https://ijsrcseit.com/CSEIT12173177