

Text Extraction from Images using Tesseract

Anand Shinde¹, Parvinder Singh², Jay Patil³, Jaideep Singh⁴, Trupti Baraskar⁵

¹⁻⁴UG Student, School of Computer Science and Engineering, MIT World Peace University, Pune, India

⁵Assistant Professor, School of Computer Science and Engineering, MIT World Peace University, Pune, India

Abstract - Important information can be found in captured images, scanned documents, magazines, newspapers, posters etc. The information in these images is highly available nowadays and they are very important in describing, representing and moving information which help people in communication, productivity, cost, analysis etc. The information from these image documents would provide a much higher ease of access if it is converted to text. The process by which text is extracted to plain text is known as Text Extraction. Text Extraction is useful in information editing, documenting, archiving, searching, or analysis of image text. However, a lot of differences in these texts because of size, orientation, and alignment, low resolution/pixelated image, and complicated and noisy background make the issue of text extraction an extremely difficult one. In this project we attempt to minimize these problems using Tesseract OCR Engine.

Key Words: Text Extraction, Image text, Tesseract, OCR, LSTM

1. INTRODUCTION

Textual Information contained in media have valuable information. Text extraction from image has stages of detection the text from given image, finding the location, extracting it, improving and recognizing the text from the image. differences in these texts because of size, orientation, and alignment, low resolution/pixelated image, and complicated and noisy background make the issue of text extraction an extremely difficult one. In this paper we attempt to minimize these problems using Tesseract OCR and display it using React JS and flask.

1.1 Area

- Text Extraction
- Python
- OCR
- Reacts JS
- Flask

1.2 Why Text Extraction?

Text extraction technology can be applied throughout multiple of industries, changing the document management process. This helps us scan physical documents & turning into fully searchable documents with text which is readable by computers. With text extraction, people will no longer need to manually retype important documents when storing them into databases. Instead, the system will extract the information and enter it automatically in the database. The result is efficient & accurate information analysis in less time.

Some of its applications are:

1. Legal:

In the legal industry, there is a process going on to digitize paper documents. In order to save space and discourage the need to search in boxes of paper files, documents are being scanned and entered into databases.

2. Banking:

Process cheques without human involvement. Overall, this will reduce wait times in many banks.

3. Healthcare:

Healthcare can also make use of image text extraction technology to process paperwork. Healthcare professionals have to deal with large number of documents for each patient, insurance forms, health forms, etc. To keep up with all of this information, it's necessary to input relevant data into a database to access it when required.

2. SYSTEM PROPOSED

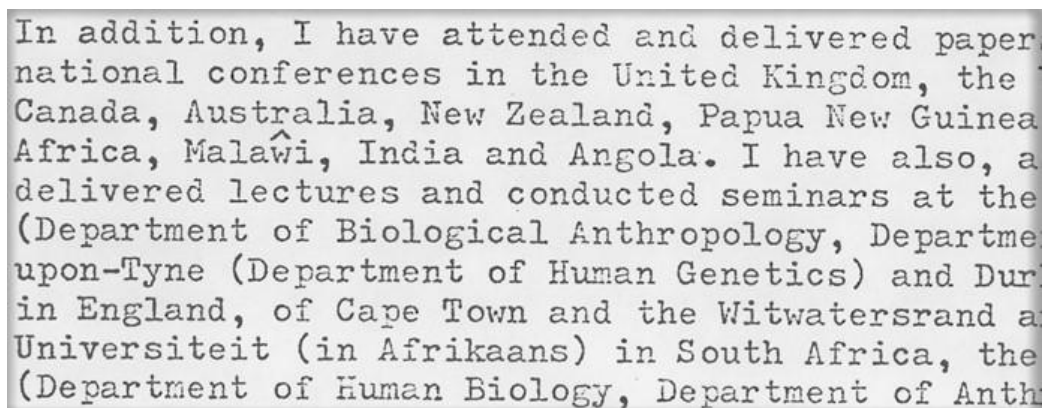
The System proposed consists of a Web application made using React JS, with flask acting as a backend. The text extractor will use the Tesseract OCR Engine to extract the Image text. The Web application will contain a section to submit/upload an image

which will then go through our text extraction program, after the program outputs the result, Flask will pull an API request to get the output text and display it on the Web page.

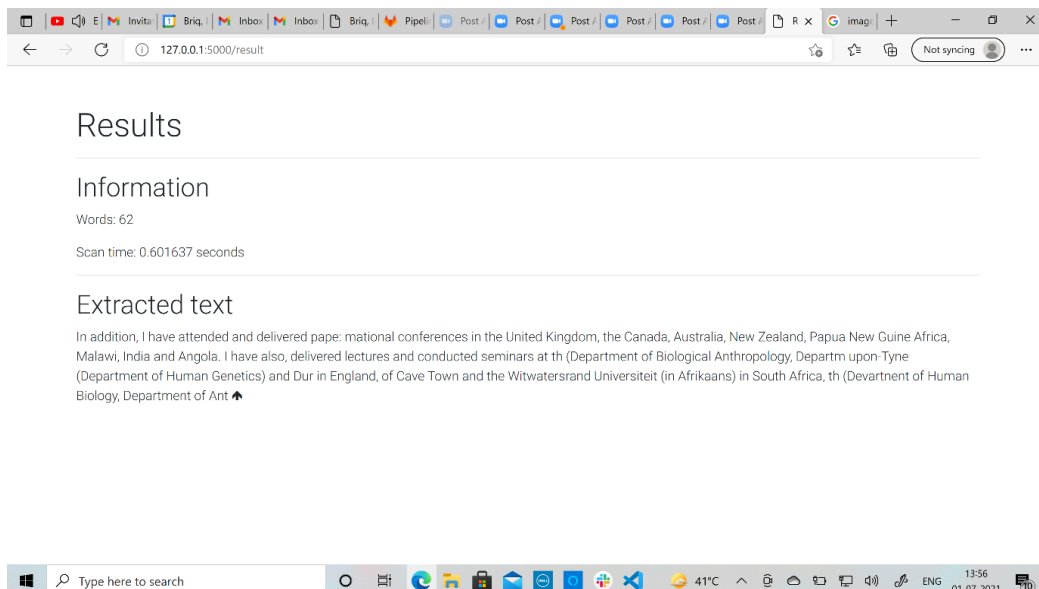
Table -1: Precision

Type	Accuracy
Machine Printed	90%
Handwritten (consistent)	76%
Handwritten (Inconsistent)	61%

Example 1:



Result:



Results

Information

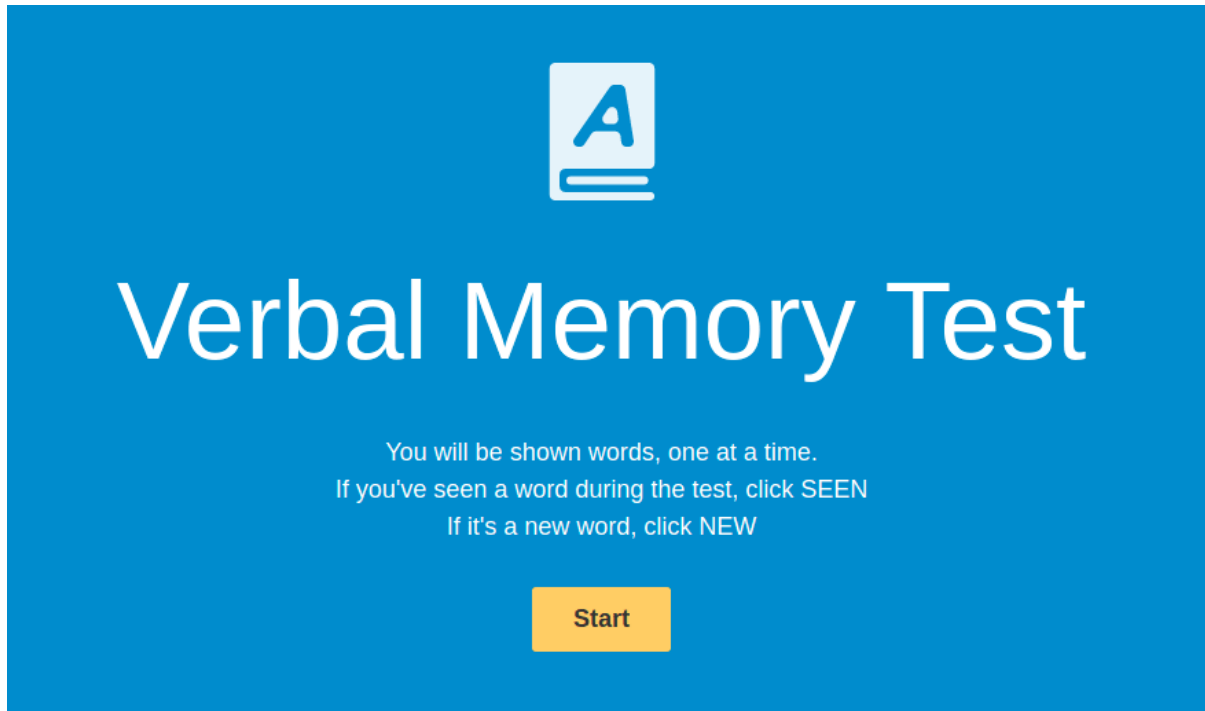
Words: 62

Scan time: 0.601637 seconds

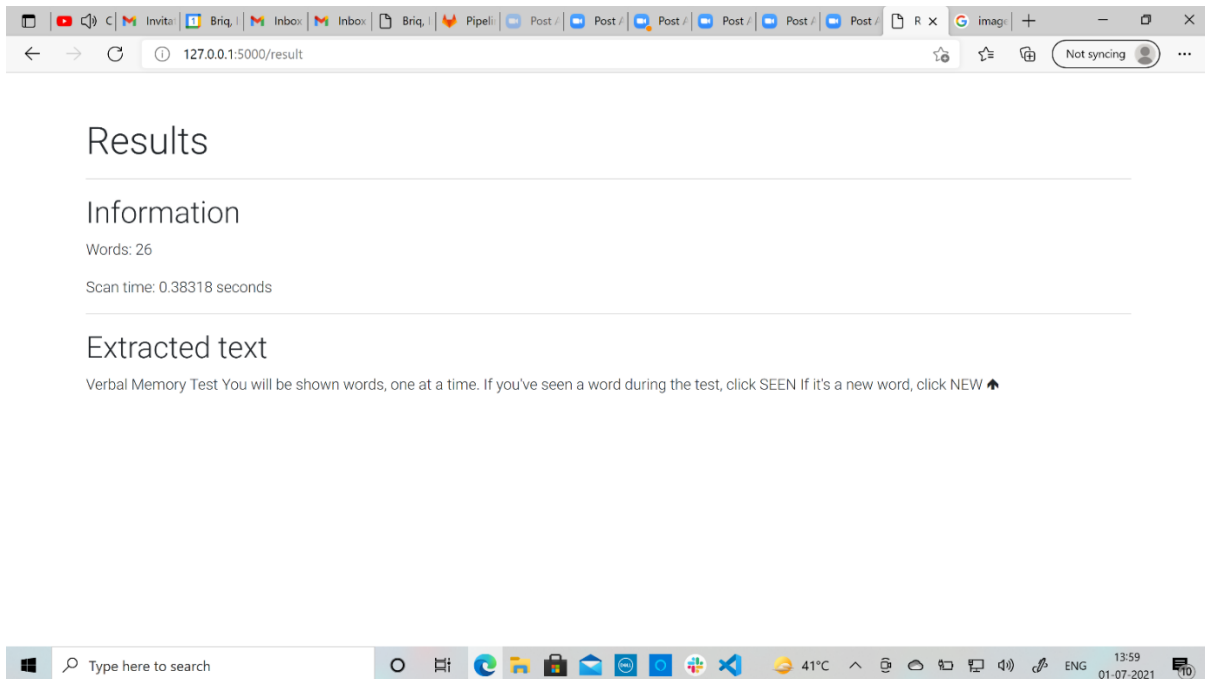
Extracted text

In addition, I have attended and delivered paper national conferences in the United Kingdom, the Canada, Australia, New Zealand, Papua New Guinea Africa, Malaŵi, India and Angola. I have also, a delivered lectures and conducted seminars at the (Department of Biological Anthropology, Departme upon-Tyne (Department of Human Genetics) and Dur in England, of Cape Town and the Witwatersrand a Universiteit (in Afrikaans) in South Africa, the (Department of Human Biology, Department of Anth

Example 2:



Result:



Example 3:

Am 4. Mai 1771

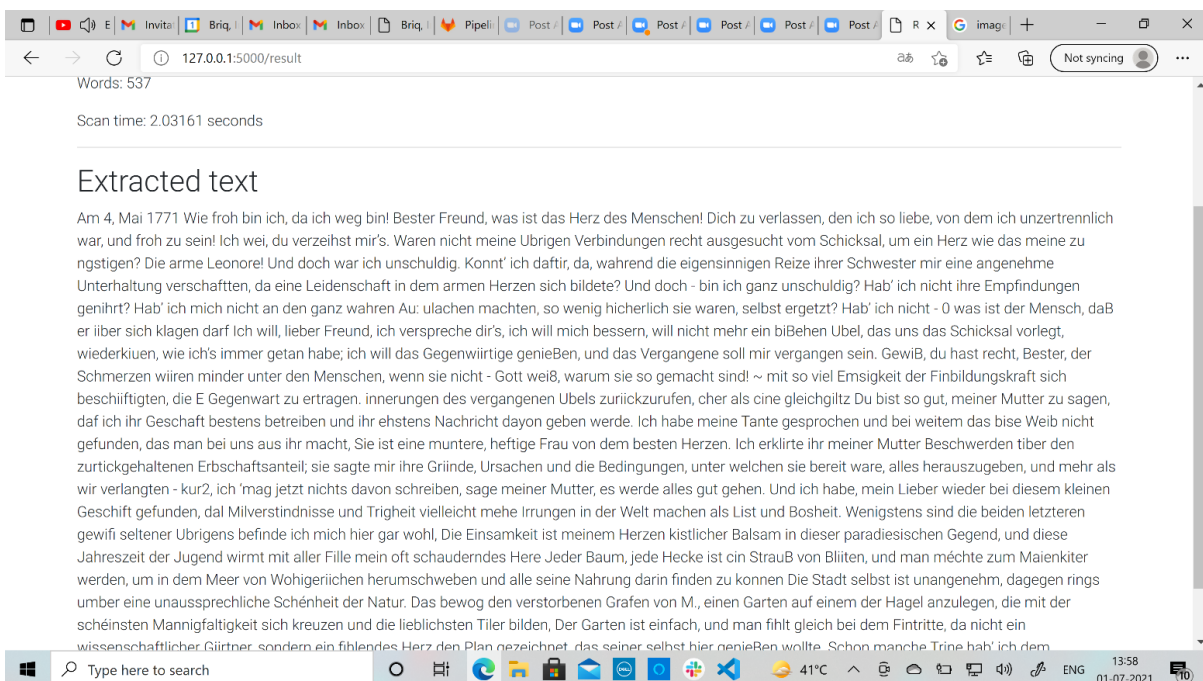
Wie froh bin ich, daß ich weg bin! Bester Freund, was ist das Herz des Menschen! Dich zu verlassen, den ich so liebe, von dem ich unzertrennlich war, und froh zu sein! Ich weiß, du verzeihst mir's. Waren nicht meine übrigen Verbindungen recht ausgesucht vom Schicksal, um ein Herz wie das meine zu ängstigen? Die arme Leonore! Und doch war ich unschuldig. Konnt' ich dafür, daß, während die eigensinnigen Reize ihrer Schwester mir eine angenehme Unterhaltung verschafften, daß eine Leidenschaft in dem armen Herzen sich bildete? Und doch - bin ich ganz unschuldig? Hab' ich nicht ihre Empfindungen genährt? Hab' ich mich nicht an den ganz wahren Ausdrücken der Natur, die uns so oft zu lachen machten, so wenig lächerlich sie waren, selbst ergetzt? Hab' ich nicht - o was ist der Mensch, daß er über sich klagen darf! Ich will, lieber Freund, ich verspreche dir's, ich will mich bessern, will nicht mehr ein bißchen Übel, das uns das Schicksal vorlegt, wiederkauen, wie ich's immer getan habe; ich will das Gegenwärtige genießen, und das Vergangene soll mir vergangen sein. Gewiß, du hast recht, Bester, der Schmerzen wären minder unter den Menschen, wenn sie nicht - Gott weiß, warum sie so gemacht sind! - mit so viel Emsigkeit der Einbildungskraft sich beschäftigten, die Erinnerungen des vergangenen Übels zurückzurufen, eher als eine gleichgültige Gegenwart zu ertragen.

Du bist so gut, meiner Mutter zu sagen, daß ich ihr Geschäft bestens betreiben und ihr ehstens Nachricht davon geben werde. Ich habe meine Tante gesprochen und bei weitem das böse Weib nicht gefunden, das man bei uns aus ihr macht. Sie ist eine muntere, heftige Frau von dem besten Herzen. Ich erklärte ihr meiner Mutter Beschwerden über den zurückgehaltenen Erbschaftsanteil; sie sagte mir ihre Gründe, Ursachen und die Bedingungen, unter welchen sie bereit wäre, alles herauszugeben, und mehr als wir verlangten - kurz, ich mag jetzt nichts davon schreiben, sage meiner Mutter, es werde alles gut gehen. Und ich habe, mein Lieber, wieder bei diesem kleinen Geschäft gefunden, daß Mißverständnisse und Trägheit vielleicht mehr Irrungen in der Welt machen als List und Bosheit. Wenigstens sind die beiden letzteren gewiß seltener.

Übrigens befinde ich mich hier gar wohl. Die Einsamkeit ist meinem Herzen köstlicher Balsam in dieser paradiesischen Gegend, und diese Jahreszeit der Jugend wärmt mit aller Fülle mein oft schauerndes Herz. Jeder Baum, jede Hecke ist ein Strauß von Blüten, und man möchte zum Maienkäfer werden, um in dem Meer von Wohlgerüchen herumschweben und alle seine Nahrung darin finden zu können.

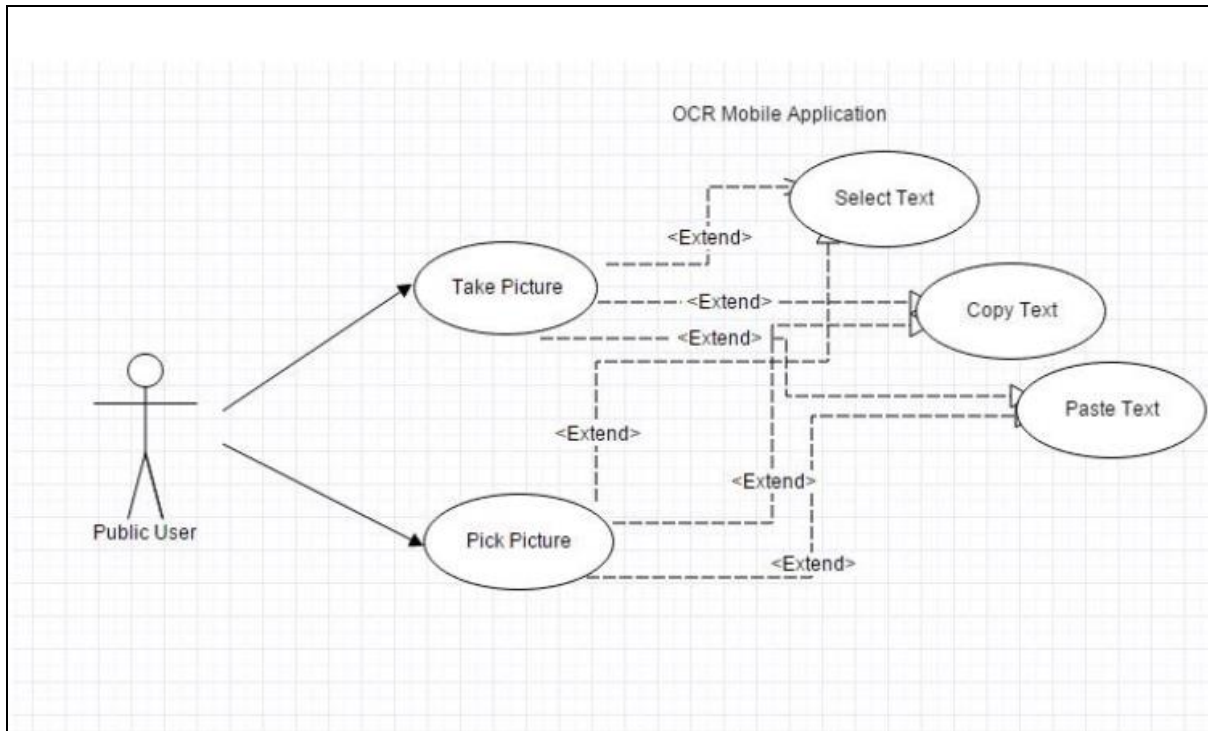
Die Stadt selbst ist unangenehm, dagegen rings umher eine unaussprechliche Schönheit der Natur. Das bewog den verstorbenen Grafen von M., einen Garten auf einem der Hügel anzulegen, die mit der schönsten Mannigfaltigkeit sich kreuzen und die lieblichsten Täler bilden. Der Garten ist einfach, und man fühlt gleich bei dem Eintritte, daß nicht ein wissenschaftlicher Gärtner, sondern ein fühlendes Herz den Plan gezeichnet, das seiner selbst hier genießen wollte. Schon manche Träne hab' ich dem Abgeschiedenen in dem verfallenen Kabinettchen geweint, das sein Lieblingsplätzchen war und auch meines ist. Bald werde ich Herr vom Garten sein; der Gärtner ist mir zugetan, nur seit den paar Tagen, und er wird sich nicht übel dabei befinden.

Result:

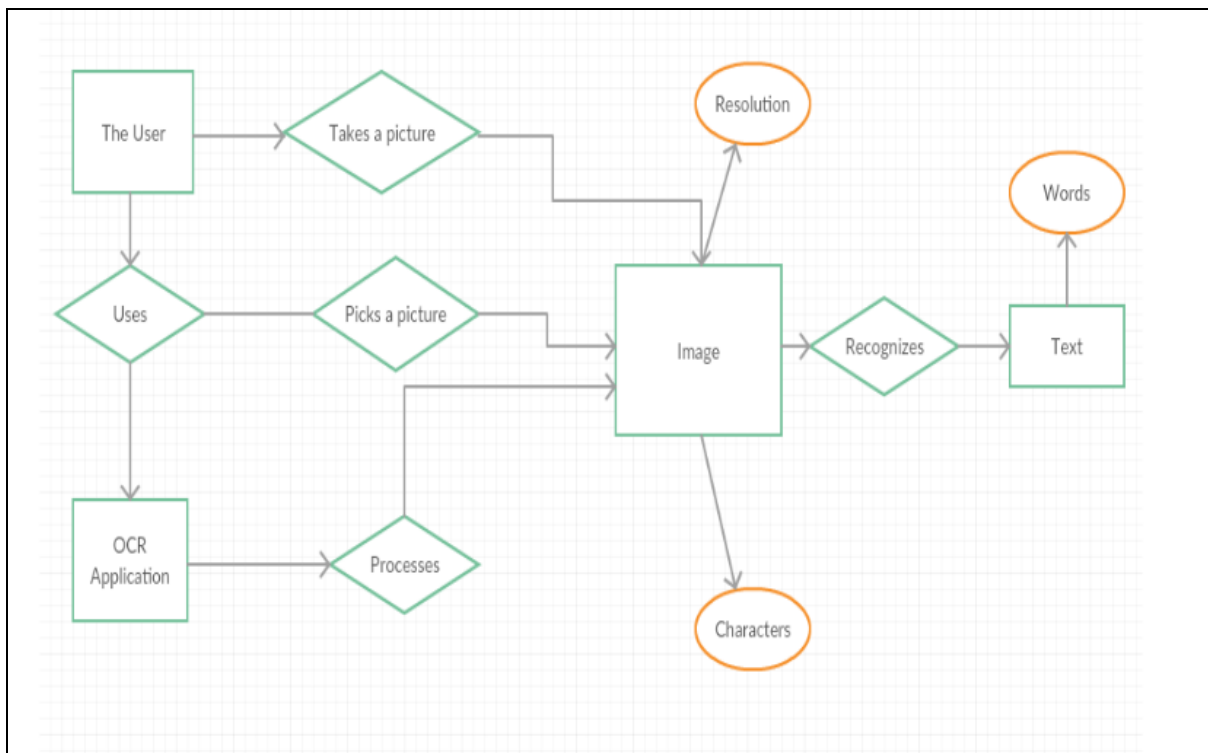


3. UML DIAGRAM

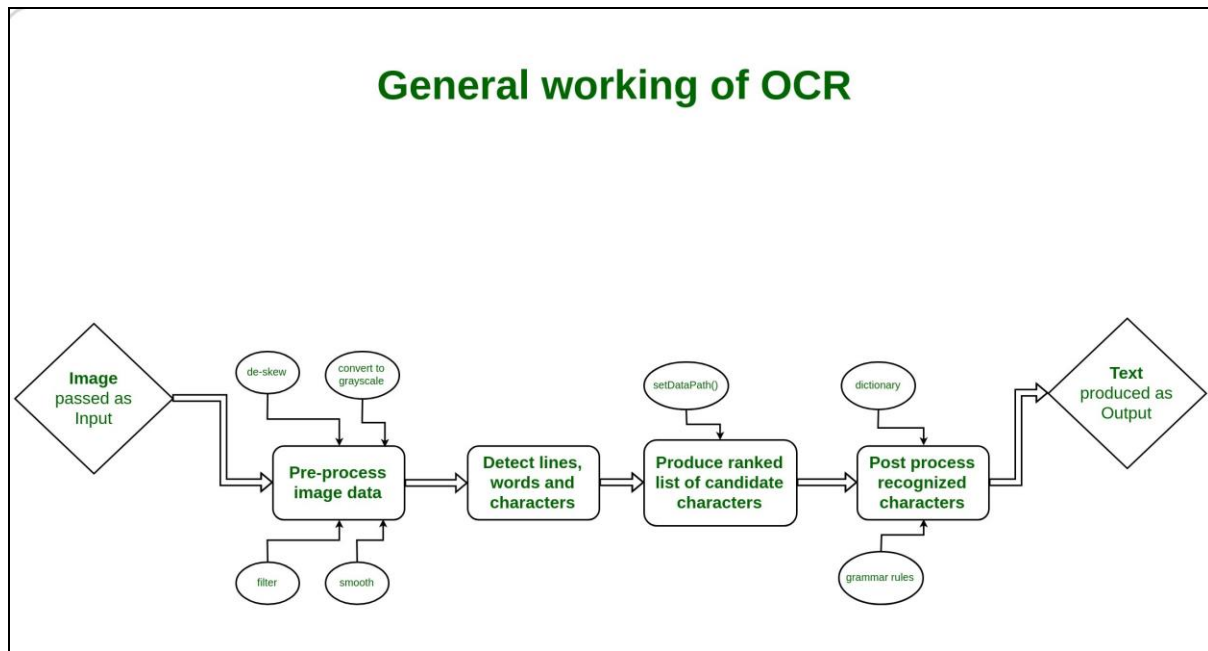
3.1 Use Case diagram:



3.2 Model Architecture:



3.3 Tesseract Architecture:



4. SYSTEM REQUIREMENTS

4.1 Software Requirements:

- React JS
- Flask
- Text Editor/IDE
- FTP Client
- Web Browser
- Graphics Editor (Nvidia)

4.2 Hardware Requirements:

- Processor: Intel Pentium IV 2.0 GHz and above
- RAM: 512 MB and above
- Hard disk: 80GB and above
- Monitor: CRT or LCD monitor
- Keyboard: Normal or Multimedia
- Mouse: Compatible mouse

5. CONCLUSION

We have created a working Text extraction Web application using Tesseract OCR engine, React JS and flask. The application successfully extracts text to use in various domains such as Banking, Academia, Finance, Legal, Healthcare, etc. for the purpose of storage, editing and analysis. To conclude, the Tesseract Engine used provided multilanguage support and various features such as alignment recognition, specific extraction and the extracted text is extremely accurate.

REFERENCES

- [1] A Novel Text Detection System Based on Character and Link Energies|| IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 23, NO. 9, SEPTEMBER 2014.

- [2] A Unified Framework for Multi-oriented Text Detection and Recognition|| IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 23, NO. 11, NOVEMBER 2014 4737.
- [3] Gradient Vector Flow and Grouping-based Method for Arbitrarily Oriented Scene Text Detection in Video Images|| IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, VOL. 23, NO. 10, OCTOBER 2013 1729.
- [4] Characterness: An Indicator of Text in the Wild Yao Li, Wenjing Jia, Chunhua Shen, and Anton van den Hengel IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 23, NO. 4, APRIL 2014.
- [5] Robust Text Detection in Natural Scene Images|| IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 36, NO. 5, MAY 2014
- [6] Scene Text Recognition in Mobile Applications by Character Descriptor and Structure Configuration|| IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 23, NO. 7, JULY 2014.

BIOGRAPHIES



Anand Shinde - School of Computer Science and Technology, MIT World Peace University.



Parvinder Singh - School of Computer Science and Technology, MIT World Peace University.



Jaideep Singh - School of Computer Science and Technology, MIT World Peace University.



Jay Patil - School of Computer Science and Technology, MIT World Peace University.



Dr. Trupti Baraskar – Assistant Professor, School of Computer Science and Technology, MIT World Peace University.