PROJECT REPORT OF

SMART OCR FOR DOCUMENT DIGITIZATION

By

MADHURA MIRIKAR

INTRODUCTION



**OVERVIEW:**

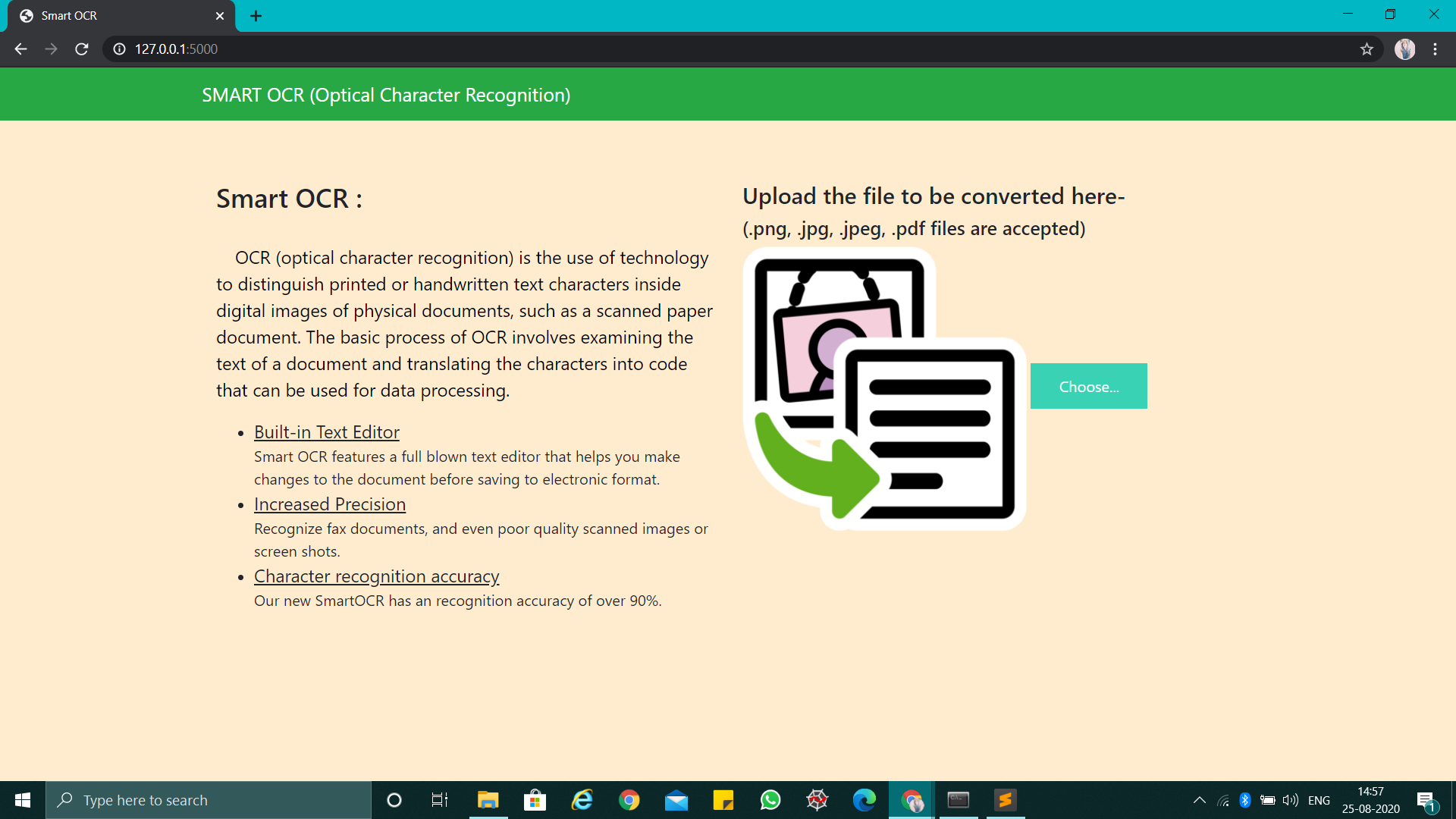
Optical character recognition or optical character reader (OCR) is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo (for example the text on signs and billboards in a landscape photo) or from subtitle text superimposed on an image (for example: from a television broadcast).

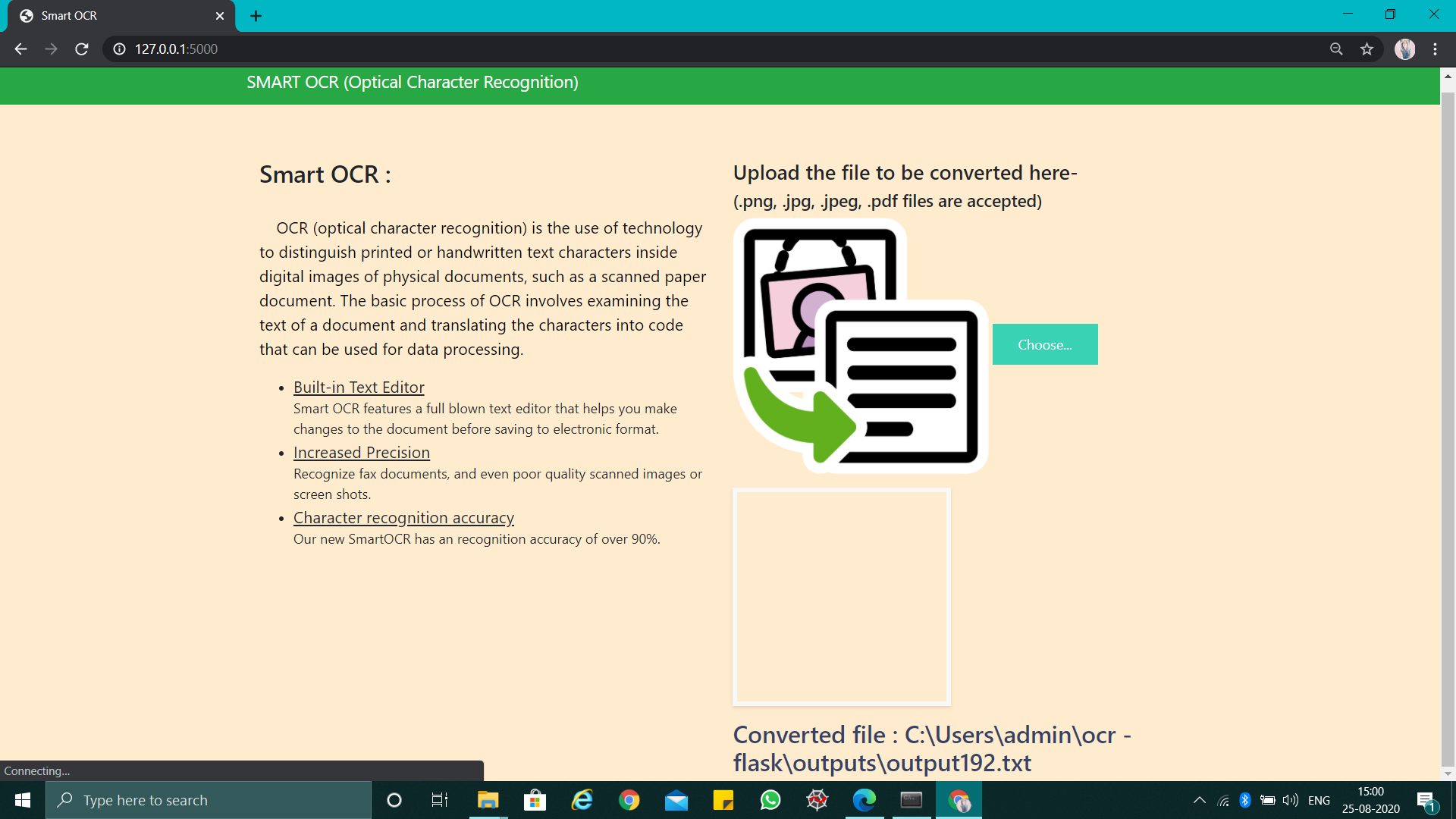
With the advent of OCR techniques, much time has been saved by automatically extracting the text out of a digital image of any invoice or a document. Currently, this is where most organizations that use OCR for any form of automation are Digital copies of invoices or documents are obtained by scanning or taking pictures. The text is extracted from these documents is entered into a template-based data entry software.

**PURPOSE:**

The project aims at creating an application from where the user can upload a pdf document, the document is analyzed by Optical character recognition package to extract text from it. The extracted text is again saved in a text document in the local drive.

RESULTS





APPLICATIONS

OCR engines have been developed into many kinds of domain-specific OCR applications, such as receipt OCR, invoice OCR, check OCR, legal billing document OCR.

They can be used for:

* Data entry for business documents, e.g. Cheque, passport, invoice, bank statement and receipt
* Automatic number plate recognition
* In airports, for passport recognition and information extraction
* Automatic insurance documents key information extraction
* Traffic sign recognition
* Extracting business card information into a contact list
* More quickly make textual versions of printed documents, e.g. book scanning for Project Gutenberg
* Make electronic images of printed documents searchable, e.g. Google Books
* Converting handwriting in real time to control a computer (pen computing)
* Defeating CAPTCHA anti-bot systems, though these are specifically designed to prevent OCR.The purpose can also be to test the robustness of CAPTCHA anti-bot systems.
* Assistive technology for blind and visually impaired users
* Writing the instructions for vehicles by identifying CAD images in a database that are appropriate to the vehicle design as it changes in real time.
* Making scanned documents searchable by converting them to searchable PDFs

CONCLUSION

Numerous algorithms, methods and techniques have been

proposed to optical character recognition in scene imagery, yet

there are not enough literature surveys in this field. In this project,

we have proposed an organization of these methods, algorithms

and techniques. It is hoped that this comprehensive survey will

provide insight into the concepts involved, and perhaps provoke

further advances in the area.

We highlight that for designing any

application related to the OCR, one must pay great attention to

each phase to obtain high accurate character recognition rate, but

still we cannot propose comprehensive algorithms for each phase

because it depends upon datasets, application specifics, and

parameter specifics. Finally major applications related to the

OCR are discussed.

FUTURE SCOPE

Although the state-of-the art OCR enables text recognition with

high accuracy, we think that there could be many more practical

applications of OCR. As a future work we are planning to use

OCR for such practical applications for daily personal use. We

are planning to incorporate mobile devices with OCR in one OCR

system. An automated book reader or a receipt tracker constitutes

some of our future OCR based applications.