Abstract

Cloud computing and Optical Character Recognition (OCR)
technology has become an extremely attractive area of research over the last few years. Sometimes it is difficult to retrieve text from the image because of different size, style, orientation and complex background of image. There is need to convert paper books and documents into text. OCR is still imperfect as it occasionally mis-recognizes letters and falsely identifies scanned text, leading to misspellings and linguistics errors in the OCR output text. A cloud based Optical Character Recognition Technology was used. This was powered on RunPod cloud in form of a Web Application Programming Interface that load images to an Optical Character Recognition server, process with necessary recognition, export parameters, and obtain the results of the processing. The key idea is to bring together the advantages of the Optical Character Recognition technology and cloud computing in one place in other to enable quicker access and faster turn out time, processing period, and increasing efficiency across the board for application users. The methodology adopted is the object oriented methodology. This was achieved using html, css and javascript programming language.

General Terms:

Cloud computing: Cloud computing is on-demand access, via the internet, to computing resources—applications, servers (physical servers and virtual servers), data storage, development tools, networking capabilities, and more—hosted at a remote data centre managed by a cloud services provider (or CSP). The CSP makes these resources available for a monthly subscription fee or bills them according to usage.

OCR:

Optical character recognition (OCR) is sometimes referred to as text recognition. An OCR program extracts and repurposes data from scanned documents, camera images and image-only pdfs. OCR software singles out letters on the image, puts them into words and then puts the words into sentences, thus enabling access to and editing of the original content. It also eliminates the need for manual data entry.

RunPod Cloud: RunPod provides two cloud computing services: Secure Cloud and Community Cloud. Secure Cloud runs in T4 data centres by our trusted partners. Our close partnership comes with high-reliability w/ redundancy, security, and fast response times to mitigate any downtimes.

Keywords: Cloud computing ,Optical character Reader, RunPod cloud, Image, text.

Introduction

There is a paradigm shift in the manner people learn, trade, communicate and share knowledge with the drastic introduction of modern computers into every aspect of life The dawn of digital computers has made it unavoidable that everything processed by the digital computer must be processed in digital form. For example, most famous libraries in the world like the Boston public library has over 6 million books, for public consumption, inescapably has to change all its paper-based books into digital documents in order that they could be stored on a storage drive. It can also be projected that every year over 200 million books are being published, many of which are disseminated and published on papers [6].Thus, for several document-input jobs, Optical Character Recognition is the utmost economical and swift process obtainable. Every year, the innovation liberates storage space once meant for file organisers and boxes loaded with paper records. Before OCR can be utilized, the source material must be scanned with the use of an optical scanner to read the page as bitmaps.(patterns of dots).. The Optical Character Recognition software then processes these scans to differentiate between images and text and determine what letters are represented in the light and dark areas. —Older OCR systems match these images against stored bitmaps based on specific fonts. The hit-or-miss results of such pattern-recognition systems helped establish OCR's reputation for inaccuracy.

Review of Optical Character

Recognition Technologies- Tesseract Tesseract is an optical character recognition engine for
various operating systems. It is free software, released under the Apache License, Version 2.0, and development has been sponsored by Google since 2006. The Tesseract engine was originally developed as proprietary software at Hewlett Packard labs in Bristol, England and Greeley, Colorado between 1985 and 1994, with some more changes made in 1996 to port to Windows, and some migration from C to C++ in 1998. A lot of the code was written in C, and then some more was written in C++. Since then all the code has been converted to at least compile with a C++ compiler.

GOCR: GOCR claims it can handle single-column sans-serif fonts of 20–60 pixels in height. It reports trouble with serif fonts, overlapping characters, handwritten text, heterogeneous fonts, noisy images, large angles of skew, and text in anything other than a Latin alphabet. GOCR can also translate barcodes.

OCRopus; OCRopus is a free document analysis and optical character recognition (OCR) system released under the Apache License, Version 2.0 with a very modular design through the use of plugins. These plugins allow OCRopus to swap out components easily. OCRopus is currently developed under the lead of Thomas Breuel from the German Research Centre for Artificial Intelligence in Kaiserslautern, Germany and is sponsored by Google. OCRopus is developed for Linux; however, users have reported success with OCRopus on Mac OS X and an application called TakOCR has been developed that installs OCRopus on Mac OS X and provides a simple droplet interface.

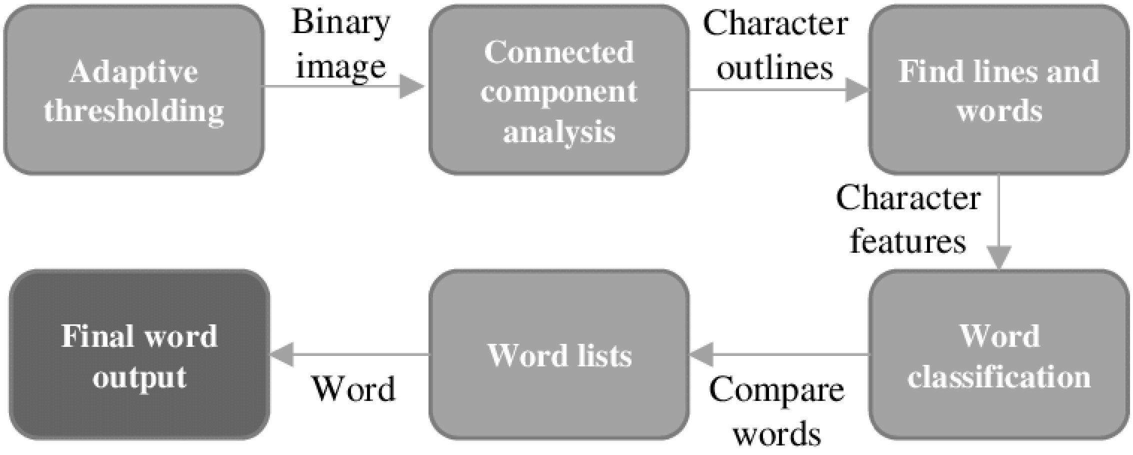
THE EXISTING SYSTEM

Over the years Optical Character Recognition technology has improved speedily. However there are limitations with respect to the source materials and character formatting. Most documents formatting are lost during text scanning, except for paragraph marks and tab stops.Sometimes, the output from a finished text scan will be a single column editable computer file. In most cases this computer file will at all times require.proofreading and spellchecking and in addition to reformatting to the required final layout. Using one of the open source and free OCR software for instance, Tesseract OCR is an elegant engine with various layers. It works in step by step manner. The first step in Named as adaptive thresholding, and converts the image into binary images. Second step is to do the connected component analysis of the image, which does the task of extracting character outlines.

After this the outlines extracted from image are converted into Blobs (Binary Long Objects). It is then organized as lines and regions and further analysis is for some fixed areas. After extraction the extracted components are chopped into words and delimited with spaces. Recognition in text then starts which is a two pass process.

Tesseract was originally designed to recognize English text only and there are limits as to the range of languages it can handle. It can only handle left-to-right languages. While you can get something out with a right-to-left language, the output file will be ordered as if the text were left-to-right. Top-to-bottom languages will currently be hopeless. It is unlikely to be able to handle connected scripts like Arabic. It will take some specialized algorithms to handle this case, and right now it doesn't have them. Moreover, it is likely to be so slow with large character set languages (like Chinese) that it is probably not going to be useful. There also still needs to be some code changes to accommodate languages with more than 256 characters.

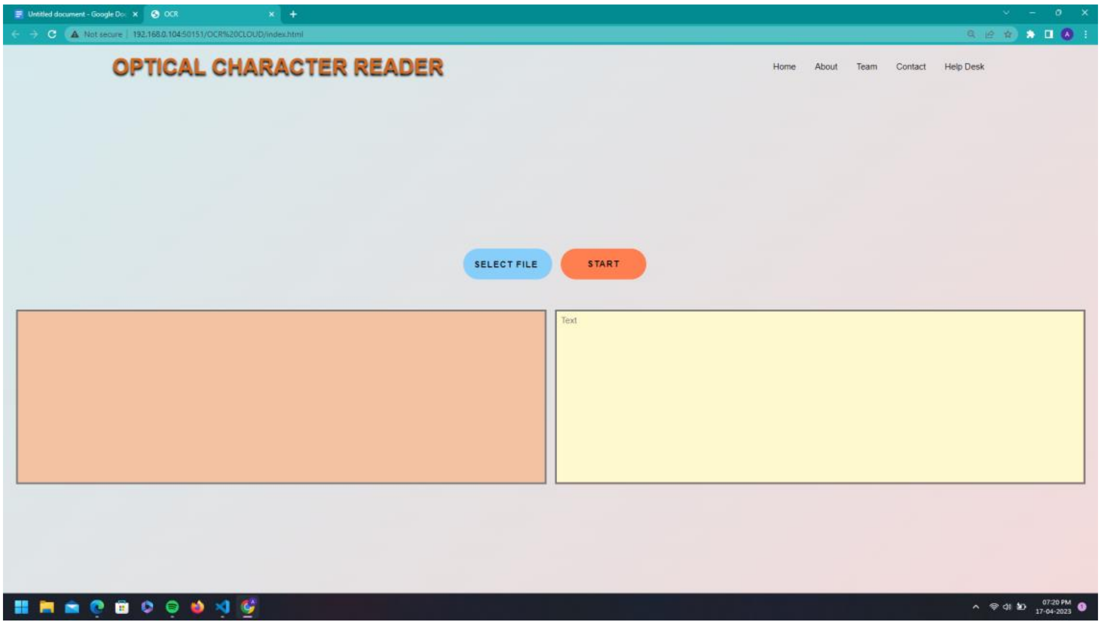
In Figure 1, the first part is when attempt to recognize each word is made. Each satisfactory word is accepted and second pass is started to gather remaining words. This brings in the role of adaptive classifier. The adaptive classifier then will classify text in more accurate manner. The adaptive classifier needs to be trained beforehand to work accurately. When the classifier receives some data, it has to resolve the issues and assign the proper place of the text.

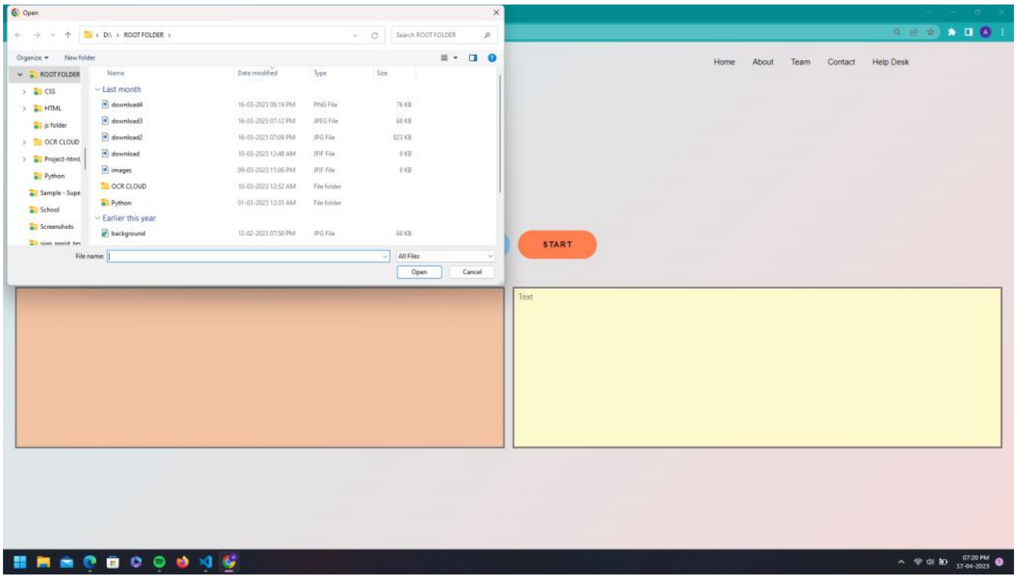


System Architecture of the Existing SystemTesseract OCR Architecture

IMPLEMENTATION AND RESULT DISCUSSIONS

Part1:
The system was implemented and tested, the JavaScript language uses a runtime environment. Also Tesseract uses OCR Development library. The user interface forms where users, Allow selecting a file to identify, the file's jpeg or png name, Open File , and start.

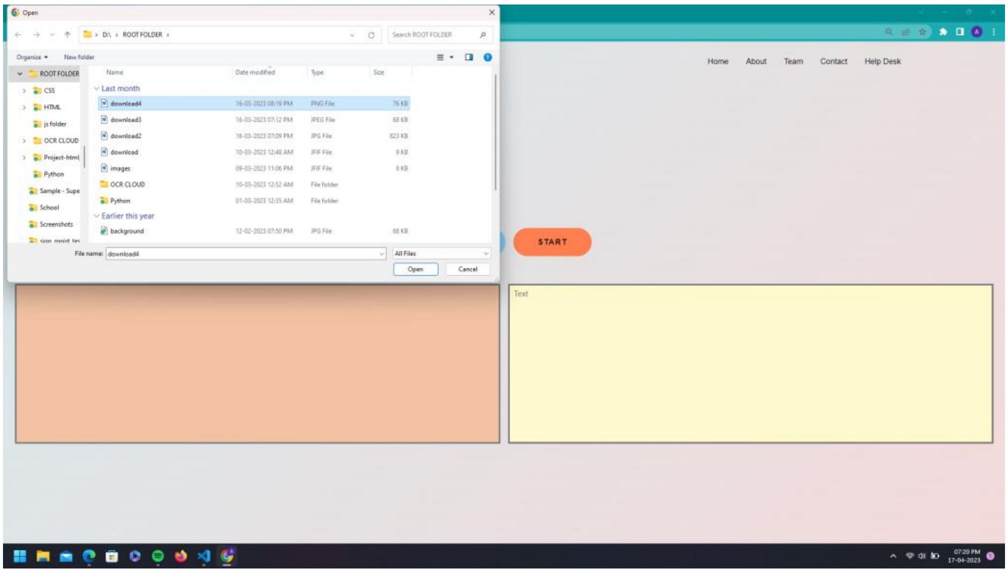




Name	Date modified	Type	Size
Last month			
download4	16-03-2023 08:19 PM	PNG File	76 KB
download3	16-03-2023 07:12 PM	JPEG File	68 KB
download2	16-03-2023 07:09 PM	JPG File	823 KB
download	10-03-2023 12:48 AM	JFIF File	9 KB
images	09-03-2023 11:06 PM	JFIF File	8 KB
OCR CLOUD	10-03-2023 12:52 AM	File folder	
Python	01-03-2023 12:35 AM	File folder	
Earlier this year			
background	12-02-2023 07:50 PM	JPG File	68 KB

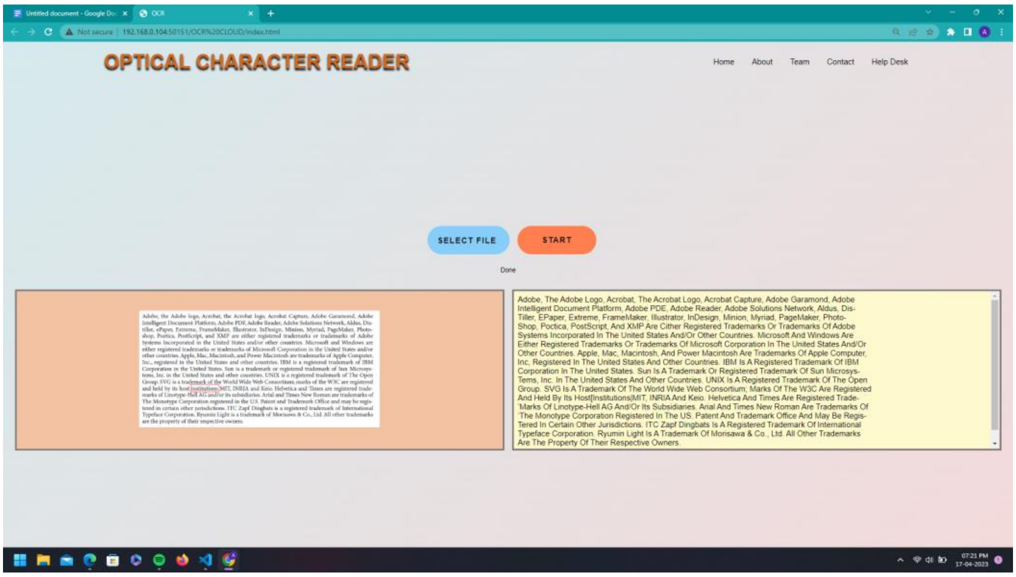
File name:

Part 2: After that the correct image file is identified for the optical character reader. For example the image file can be in jpeg or png.



Name	Date modified	Type	Size
Last month			
download4	16-03-2023 08:19 PM	PNG File	76 KB
download3	16-03-2023 07:12 PM	JPEG File	68 KB
download2	16-03-2023 07:09 PM	JPG File	823 KB
download	10-03-2023 12:48 AM	JFIF File	9 KB
images	09-03-2023 11:06 PM	JFIF File	8 KB
OCR CLOUD	10-03-2023 12:52 AM	File folder	
Python	01-03-2023 12:35 AM	File folder	
Earlier this year			
background	12-02-2023 07:50 PM	JPG File	68 KB

File name: download4



Adobe, the Adobe logo, Acrobat, the Acrobat logo, Acrobat Capture, Adobe Garamond, Adobe Intelligent Document Platform, Adobe PDF, Adobe Reader, Adobe Solutions Network, Aldus, Distiller, ePaper, Extreme, FrameMaker, Illustrator, InDesign, Minion, Myriad, PageMaker, Photoshop, Poetica, PostScript, and XMP are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States and/or other countries. Microsoft and Windows are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries. Apple, Mac, Macintosh, and Power Macintosh are trademarks of Apple Computer, Inc., registered in the United States and other countries. IBM is a registered trademark of IBM Corporation in the United States. Sun is a trademark or registered trademark of Sun Microsystems, Inc. in the United States and other countries. UNIX is a registered trademark of The Open Group. SVG is a trademark of the World Wide Web Consortium; marks of the W3C are registered and held by its host institutions MIT, INRIA and Keio. Helvetica and Times are registered trademarks of Linotype-Hell AG and/or its subsidiaries. Arial and Times New Roman are trademarks of The Monotype Corporation registered in the U.S. Patent and Trademark Office and may be registered in certain other jurisdictions. ITC Zapf Dingbats is a registered trademark of International Typeface Corporation. Ryumin Light is a trademark of Morisawa & Co., Ltd. All other trademarks are the property of their respective owners.

```
Adobe, The Adobe Logo, Acrobat, The Acrobat Logo, Acrobat Capture, Adobe Garamond, Adobe
Intelligent Document Platform, Adobe PDE, Adobe Reader, Adobe Solutions Network, Aldus, Dis-
Tiller, EPaper, Extreme, FrameMaker, Illustrator, InDesign, Minion, Myriad, PageMaker, Photo-
Shop, Poctica, PostScript, And XMP Are Cither Registered Trademarks Or Trademarks Of Adobe
Systems Incorporated In The United States And/Or Other Countries. Microsoft And Windows Are
Either Registered Trademarks Or Trademarks Of Microsoft Corporation In The United States And/Or
Other Countries. Apple, Mac, Macintosh, And Power Macintosh Are Trademarks Of Apple Computer,
Inc, Registered In The United States And Other Countries. IBM Is A Registered Trademark Of IBM
Corporation In The United States. Sun Is A Trademark Or Registered Trademark Of Sun Microsys-
Tems, Inc. In The United States And Other Countries. UNIX Is A Registered Trademark Of The Open
Group. SVG Is A Trademark Of The World Wide Web Consortium; Marks Of The W3C Are Registered
And Held By Its Host[Institutions|MIT, INRIA And Keio. Helvetica And Times Are Registered Trade-
'Marks Of Linotype-Hell AG And/Or Its Subsidiaries. Arial And Times New Roman Are Trademarks Of
'The Monotype Corporation Registered In The US. Patent And Trademark Office And May Be Regis-
Tered In Certain Other Jurisdictions. ITC Zapf Dingbats Is A Registered Trademark Of International
Typeface Corporation. Ryumin Light Is A Trademark Of Morisawa & Co., Ltd. All Other Trademarks
Are The Property Of Their Respective Owners.
```

CONCLUSION:

Cloud computing and OCR technology can greatly enhance document management and workflow automation. By combining cloud computing and OCR, users can achieve faster turnaround times, increased productivity, and cost savings. Cloud-based OCR solutions automate document processing, reduce manual data entry, and improve accuracy. This integration represents a powerful solution for businesses looking to optimize their document management processes and improve their bottom line.