Applied Information Technology And Computer Science Vol. 3 No. 2 (2022) 370-387

AITCS

Homepage: http://publisher.uthm.edu.my/periodicals/index.php/aitcs
e-ISSN :2773-5141

Development of a Web-based Optical Character Recognition System

Wong Parh Yong[1], Nayef Abdulwahab Mohammed Alduais[1*]

[1]Faculty of Computer Science and Information Technology,

University Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, 86400, MALAYSIA

DOI: https://doi.org/10.30880/aitcs.2022.03.02.025
Received 14 June 2022; Accepted 26 September 2022; Available online 30 November 2022

Abstract: Optical Character Recognition Technology, OCR in short, is one of the most crucial field of research in improving and forwarding machine learning and AI technology. The proposed system is a Python language developed web application to deliver one of the simpler and general use of such technology to solve daily problems, which is to convert texts and character in image into editable text insert into document files. The system integrates authentication system with login and register to track, record and store their uploaded data into the system's database. The main function is an OCR function that accepts image files from users to convert them into document files where the users can then download it from the server, powered by Tesseract, an open source and free command line program with an OCR engine, and can be commanded in Python with the Pytesseract library. Users can send feedback and help with the learning and training of the AI by uploading multiple files as training sample in the feedback function. The system allows editingof profile credentials through the user profile page. The end result of the project is a web app that focus primarily on the OCR function, where users can directly and indirectly help with the improving and learning of the AI. Although still having some flaws including the download function not working well with client devices, further improvements and recommendations have been suggested to better make the system works in future.

Keywords: Optical Character Recognition, Machine Learning

**Corresponding author: nayef@uthm.edu.my*

1. Introduction

The web application planned for design is an online Optical Character Recognition web application. Optical Character Recognition, or OCR in short, is the technology used to convert images of typed, handwritten and/or printed text into machine-encoded text, which can be recognized by machines, whether the source is from a photo of a document, a scanned document, a scene-photo or subtitle text superimposed on an image. This technology can be traced back to year 1914, where 2 inventors, Emanuel Goldberg and Edmund Fournier d'Albe, each developed a machine that can read characters and convert them to either standard telegraph code or voice.

Nowadays, OCR technology is widely used in different hardware and software to deliver different functions, some can convert text into voice and tone to be read to the users, such as Amazon Polly and Linguatec Voice Reader; while others are able to convert text extract from different sources to machine-readable Text to enable different function be done on the selected text, such as Adobe Acrobat PRO OmniPage Ultimate.

However, while there are quite a few OCR software out in the market, there are quite a few disadvantages that are common among these existing software, which includes accuracy of the conversion is almost always lower than expected, most of these software are quite costly to be used and the lack of function to feedback on inaccuracy and error in conversion.

Thus, there are a few suggestion that can be possible solution to the weak points, which includes setting up a community to gather more samples of different handwritten text and photographsas learning materials for the AI behind the OCR Technology, allow users to identify the converted text manually as a feedback to the application, and allow free trials for the software.

2. Related Work

OCR technology is not something new, although still having a lot of potential for a breakthrough, the system this project proposed has a lot of predecessors. This section will find among the most popular and successful existing system to be discussed and analyzed.

2.1 ILOVEPDF

ILOVEPDF is a web-based service that can manage a lot of aspects of PDF, providing various tools that can do almost everything that is related to PDF. It is free for most part, but is limitedin every of its functions. If one wants the most of it, the user can purchase the premium for it.

2.2 Adobe Acrobat Pro

Easily among the most well-known and most powerful PDF reader and editor out there, Adobe Acrobat Pro is a professional PDF All-In-One tool that should be considered by users if they frequent their uses of PDF in daily lives. The strongest advantages of it should be the linking of all Adobe software that outdo most of the available PDF tools out there as it not just covers PDFs, but it can be extended to have additional functions that will certainly help users to complete everyday document needs.

2.3 PDF-XChange Editor

Although lesser known, PDF-XChange Editor is a great PDF All-In-One tool that specialize even more than the previous mentioned Adobe Acrobat Pro and ILOVEPDF. If Adobe Acrobat Pro covers more area in width, having a wider range of tools that serve not only PDF, then PDF-XChange Editor has a more narrow but deeper thus stronger tool that only focus in PDF. If one learn to use its advance features, it can actually do literally anything to PDF, even more than ILOVEPDF and Adobe Acrobat Pro. It is also a free software that one can use almost 70 percent of its function limitless, while the other 30 percent requires only a watermarking on its page which can be removed with premium purchase. It is an advance PDF editor as it can actually embed items from other sources into

PDF, calculate the pixel and distance based on scales of PDF images, add all sorts of Annotations such as Audio and Measurement to PDF and much more, and these functions are not included in any of the previously mentioned existing systems.

3. Methodology/Framework

Prototyping model is chosen as it suits the project development of a web-application development. Since it will involve users' opinions and feedback in early days of development, it can gain more users' satisfaction. Prototyping model also helps developers to add in, edit, and remove functionalities and features, to detect, repair and amend errors and bugs, and generally test and gain insight on how their product may work in reality. Moreover, prototyping can help developers to gain more knowledge and skills in the development process, since repeated working on the model will certain allow more usage of knowledge and skills gained before, and at the end can gain more practical and hands-on experience. Lastly, prototypes not used, if not discarded, can be stored and be used in future projects.

3.1 Planning Phase

This is the planning phase, which will be the initial step of the development process. The main purpose of this phase is to determine the crucial points in developing a project, which includes WHY and HOW to develop the product/service/project. In this phase, the project's title, objectives, scopes, expected outcomes, and a few other details that will be used in enacting a system request (initial proposal of the project) are figured out. Also in this phase, the project plan, project schedule and work plan are worked out for the development process. For this project, a Gantt Chart is createdas the project schedule and plan. The Gantt Chart is placed in Appendix A.

3.2 Analysis Phase

The main goal of the analysis phase is to gather information and researching. In this phase, a few important activities are carried out, which includes writing up literature reviews, do analysis on requirements including software, hardware, functional, non-functional and user requirements. Tables and matrices are built to hold these information.

TABLE 1: Hardware Requirements

Properties Description	Properties
Device name	LAPTOP-L2TLKBGA
Processor	AMD Ryzen 7 4800H with Radeon Graphics 2.90 GHz
Installed RAM	16.0 GB (15.4 GB usable)
Device ID	45CDCB0E-B357-4083-8C6E-8B673505A073
Product ID	00342-41424-54987-AAOEM
System type	64-bit operating system, x64-based processor
Hard Drive Space(ROM Storage Space)	1.0 TB SSD

TABLE 2: Software Requirements

Applications and Software	Usage
Sublime Text, Thonny, Notepad++	Coding and Compiler software
Windows PowerShell(Admin)	Python program runner and debugger
Windows 10 Home Single Language(64 bit)	Laptop OS
Python 3.9.9	Python Environment for Web App Development
Bootstrap 5.1.3	HTML Styling Preset Pack
XAMPP with Apache and MySQL	Local Server and Database
Python Flask	Python Web App Deployment
Python Tesseract, Pytesseract and CV2	Python OCR Plugin and Data Trainer
(Various other Python Plugins)	Python App Functionalities Support
Web Browsers such as Opera GX and Google Chrome	Environment for deploy Web App and Accessing the Local Server

3.2.1 Functional Requirements

The following are the functional requirements of the system:

TABLE 3: Functional Requirements

NO	Modules	Functions
1	Login	Allow users to login into the system with correct username and password, will validate the credentials with the registered one in the database.
2	Register	Allow users to register an account, with specific format with each credentials and store them into database.
3	Home	Allow users to redirect to other function pages after user login and land at this page.
4	OCR Processing	Allow users to convert uploaded image file with text to editable text file and download them.
5	Feedback and Training Center	Allow users to contact the admin indirectly through sending messages, and help with the training of the OCR module indirectly by sending sample data.
6	User Profile	Allow users to view their profile and change their credentials.

3.2.2 Non-Functional Requirements

The following are the non-functional requirements of the system:

TABLE 4: Non-Functional Requirements

NO	Modules	Functions
1	Integrity	Data should be backed up to the database and other local storage.
2	Usability	The design of the system will be simple and easy to be used.
3	Availability	The web app should be up in the server for most of the time, only coming down for update.
4	Security	The usage of the system should be available if and only if users are logged into the system.
5	Performance	The proposed web system can be slow, but must perform its functionalities without much problem at most of the time.

3.3 Designing Phase

The next phase in the system development life cycle is the Design Phase, which where the initial design of the project is created. Before the designing process starts, determination of a design strategy (to build, to outsource or to buy) is done. The main activity here is to come up with a design of system components, including the architecture, interface, database, programs and so on, andassemble them into into System Specifications. For this project, since a web application is being designed and developed, the main interface of the web pages are designed and deployed.

3.3.1 Data Flow Diagram (DFD) Level 0

The Data Flow Diagram(DFD) at Level 0, which is also known as the Context Diagram, which is the overall and basic overview of the whole system data flow. Since this is the overall DFD,it shows how all the processes, inputs and outputs relate to each other and affect the overall process.

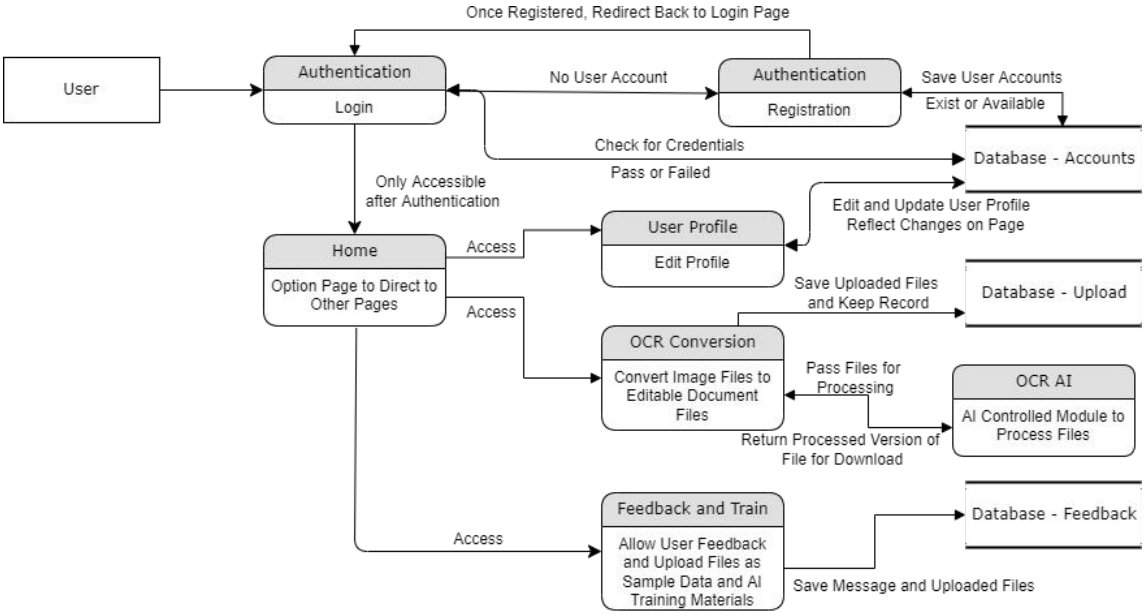

Figure 1: Data Flow Diagram (DFD) Level 0

3.3.2 Entity Relationship Diagram (ERD)

An Entity Relationship Diagram or ERD, is a type of flowchart that show the entities which includes the system, users, databases and any related human, object and concepts, and they relate with each other in a system.

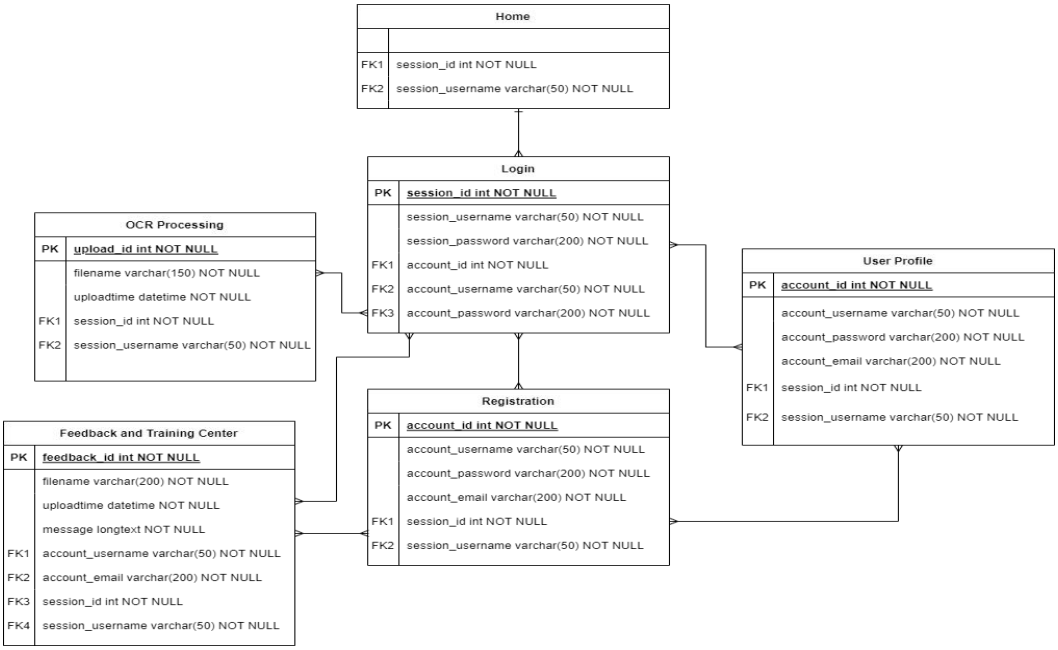

Figure 2: Entity Relationship Diagram (ERD)

3.3.3 Use Case Diagram

A use case diagram shows the interaction between user's possible cases of interaction with the system.

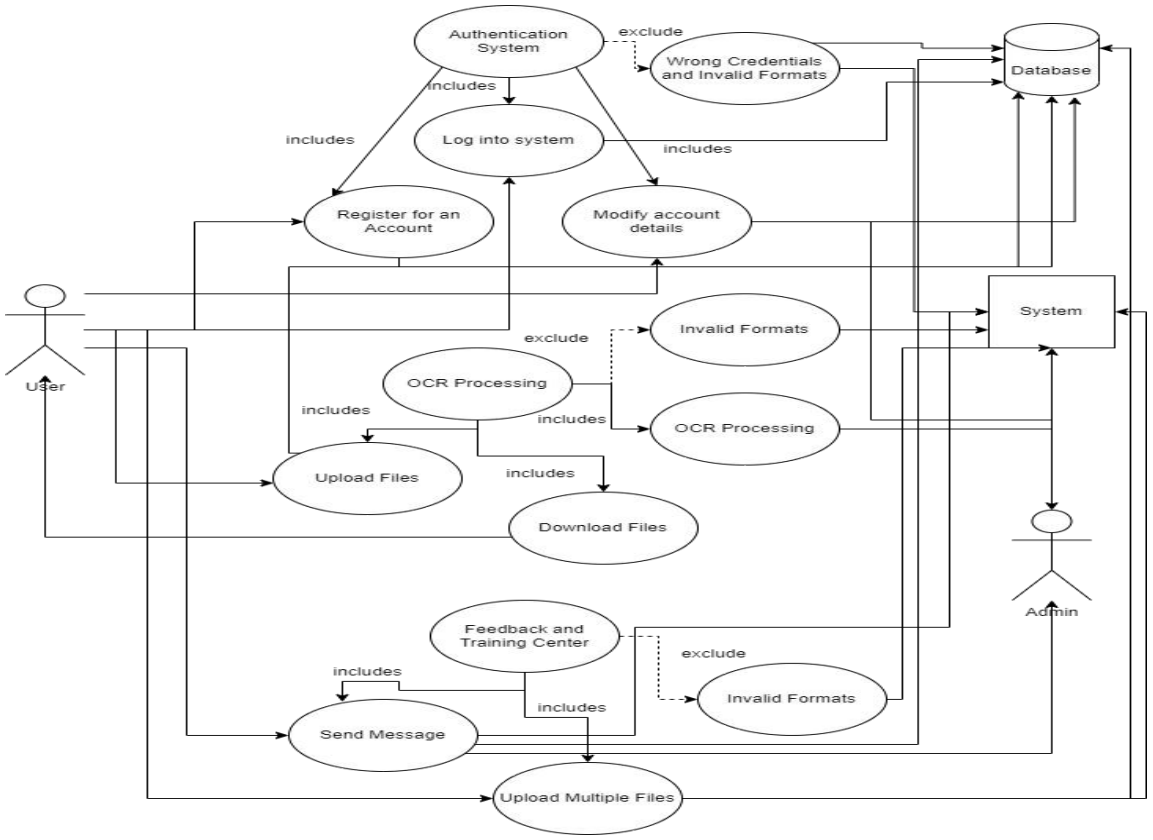

FIGURE 3: Use Case Diagram

3.4 Implementation Phase

Implementation in an IT project is the process of execution and running of designed and developed programs, coding, software, applications, websites and so on. The process will have the programs be run in normal environment to test for their functions and effectiveness, while also check for any errors and bugs before publishing the works.

3.4.1 User Interface Design

The user Interface is how users will interact with the system. Thus it is important to design and implement an user interface that is beautiful in aesthetics, simple in usage and easy to be navigated and identified. The following are the interfaces that have been designed based on each process in different figures. The Interface is designed with HTML 5, CSS, JavaScript and PHP codes.

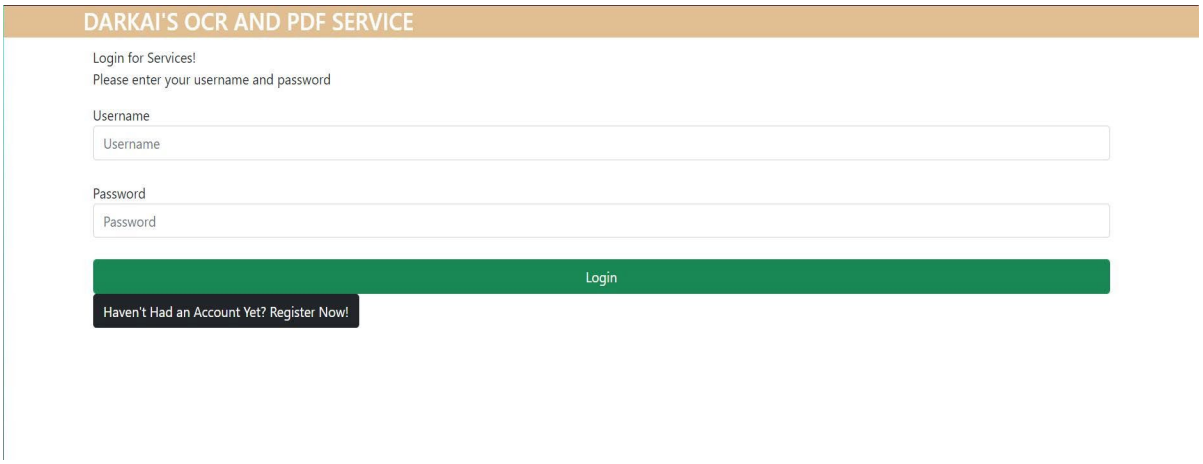

FIGURE 4: User Interface of Login Page

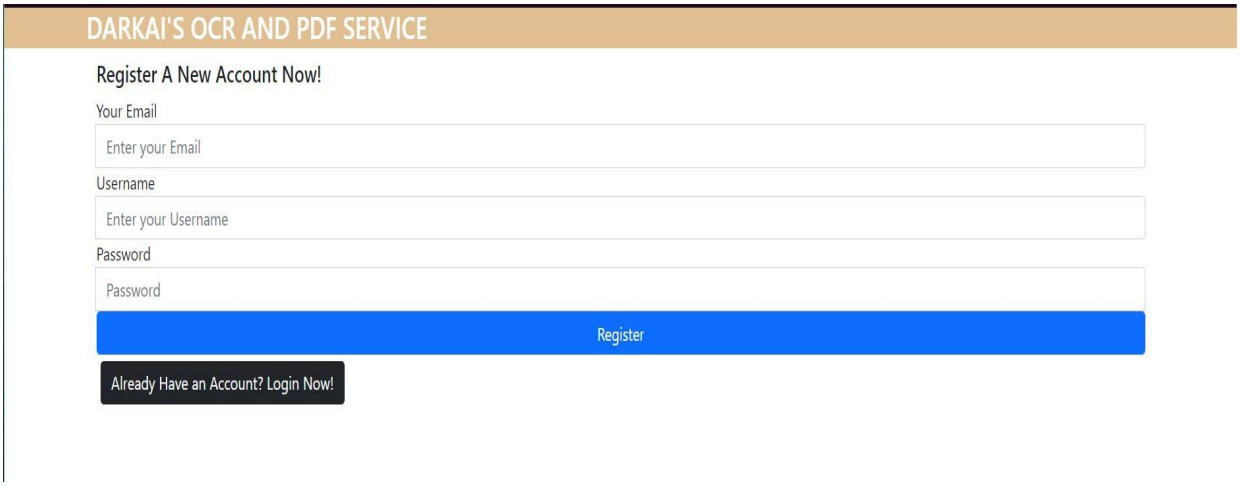

FIGURE 5: User Interface of Register Page

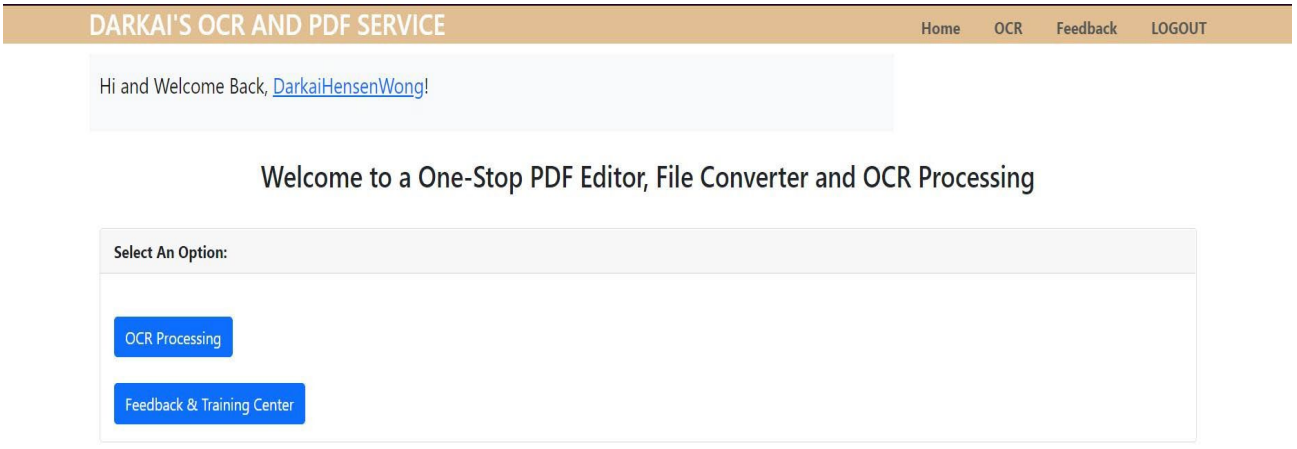

FIGURE 6: User Interface of Home Page

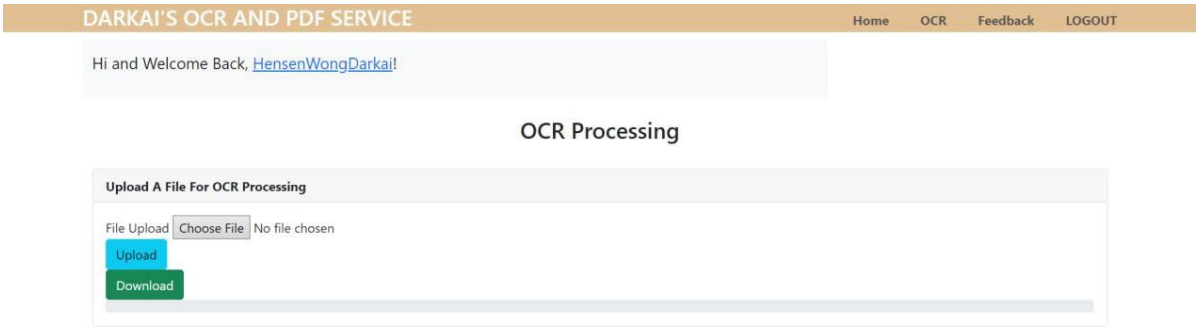

FIGURE 7: User Interface of OCR Processing Page

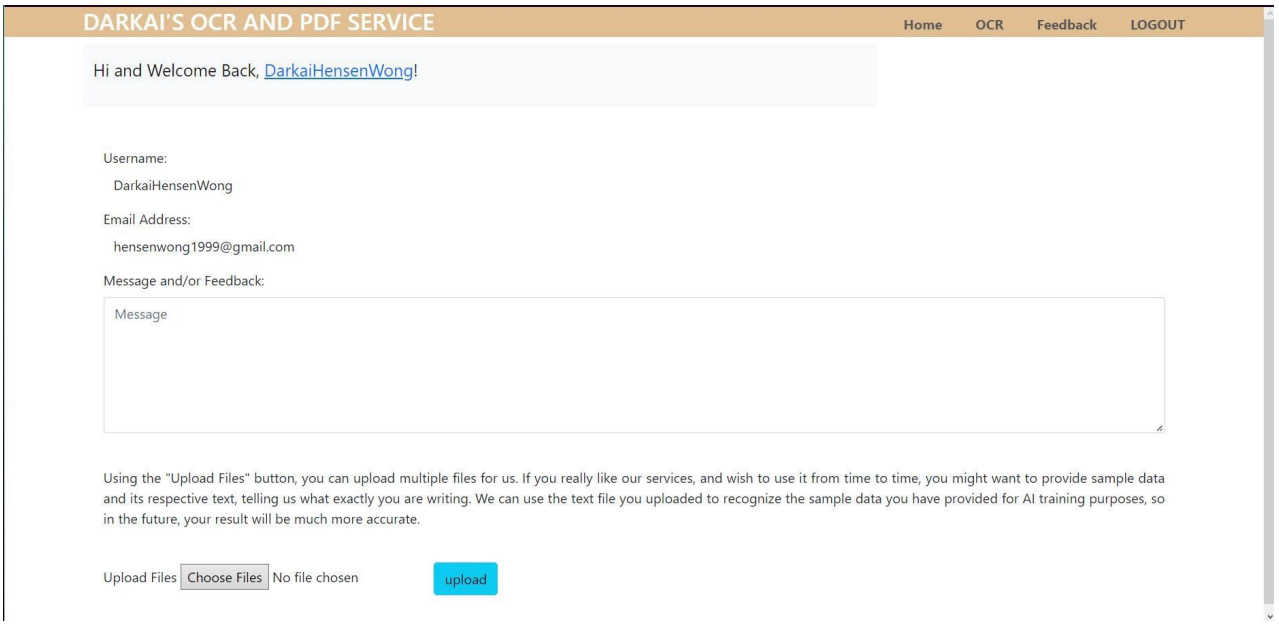

FIGURE 8: User Interface of Feedback and Training Centre Page

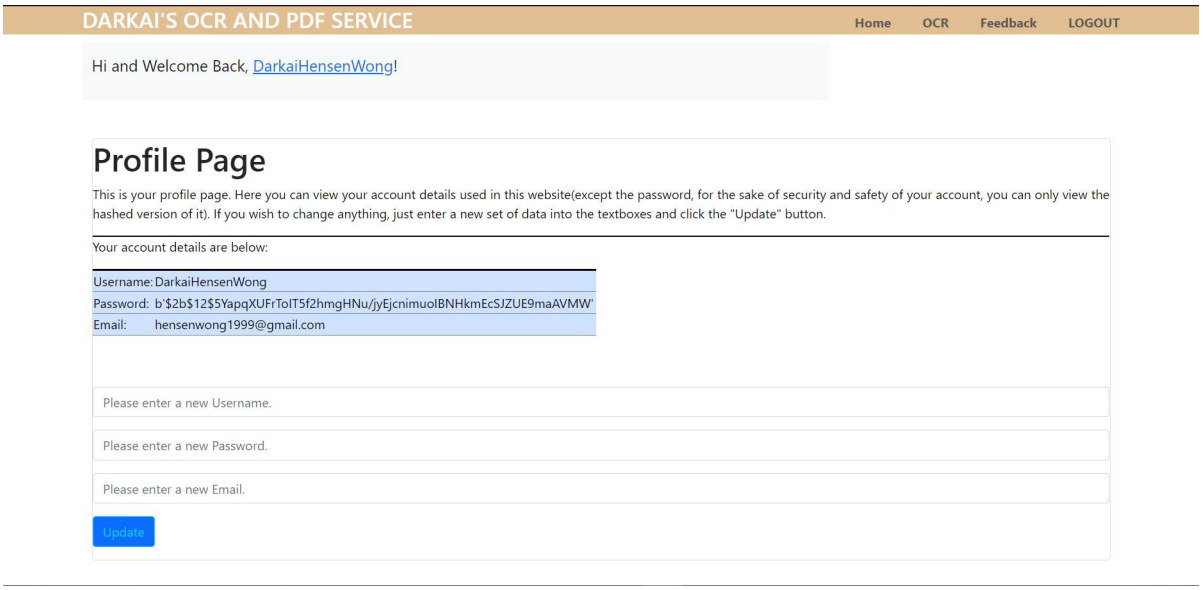

FIGURE 9: User Interface of User Profile Page

4. Results and Discussion

Testing of the system is crucial to ensure that the designed and developed system work as intended, while also to check for any possible bugs and errors that will not actually cause problems in deployment of the system, but will cause annoyance and problems in user experience. 3 types of testing have been used to evaluate the system, which includes System Testing, User Acceptance Testing and Accuracy Testing.

4.1 System Testing

System testing is a testing done by the developers to evaluate the system before full deployment to find out whether the requirements are fulfilled. The testing is based on the test plan designed in the analysis and design phase. Overall, more than 90 percent of the test passed successfully, with a few exceptions.

TABLE 5: Test Plan

Testing Title	Expected Results	Actual Result
General Properties		
The App is successfully deployed in server.	Can be deployed in any servers for usage.	Yes
The App is running without any bugs.	App can conduct without any major or minor bugs.	No
The App is able to be run online.	App can be deployed online and be searched online easily.	Yes
Login Function		
The UI of the page is displayed correctly.	All elements are displayed as intended.	Yes
The Login function is able to accept strings input by the user.	data with the ones in the database. The string inputted can be passed withoutproblems.	Yes
System is able to compare and validate the input	Remind users of incorrect and not-matching	Yes

TABLE 5: Test Plan (Cont.)

Suitable message can be shown when triggered.	Will display error message.	Yes
Redirecting to the Home Page is done correctly.	Redirected to the Home page after logging in.	Yes
Register Function		
The UI of the page is displayed correctly.	All elements are displayed as intended.	Yes
The Register function is able to accept strings input by the user.	The string inputted can be passed without problems.	Yes
System is able to compare and validate the input data with the ones in the database.	Remind users of used and not unique details.	Yes
Suitable message can be shown when triggered.	Will display error message.	Yes
Redirecting to the Home Page is done correctly.	Redirect to the Login page after successfully registered an account.	Yes
The registered information is unique and cannot be null.	Remind users of used and not unique details and not able to be used again.	Yes
The system can control and validate the entered information is according to the needed format.	Inform and decline username and email data that are not having the required format.	Yes
OCR Processing		
The UI of the page is displayed correctly.	All elements are displayed as intended.	Yes
Page cannot and should not to be able to be accessed by users without logging in.	Redirect users to login page if not logged in.	Yes
The system is able to accept and validate the needed file and file extension.	Accepts only image files	Yes
The uploaded file is successfully recognized.	Image with text can be converted to editable document files.	Yes
The uploaded file is saved in the correct specified location.	Uploaded files are saved to a specific folder to keep as training data.	Yes
A preview of the uploaded file is shown.	Successful conversion will allow display of uploaded file.	Yes
The output can be download successfully.	The file can be outputted to users' download folder or a location they choose.	No
The Accuracy of the OCR is satisfying.	The text generated from the image recognized is as accurate as possible.	Yes

Feedback and Training Center		
The UI of the page is displayed correctly.	All elements are displayed as intended.	Yes
Page cannot and should not to be able to be accessed by users without logging in.	Redirect users to login page if not logged in.	Yes
The system is able to accept and validate the needed file and file extension.	Accepts only image and document files, including PDFs	Yes
The data input is correctly uploaded and recorded in the Database.	Record uploadtime, message, username, email and filename	Yes
The uploaded file is saved in the correct specified location.	Uploaded files are saved to a specific folder to keep as training data.	Yes
Home		
The UI of the page is displayed correctly.	All elements are displayed as intended.	Yes
Page cannot and should not to be able to be accessed by users without logging in.	Redirect users to login page if not logged in.	Yes
The buttons and links are working properly.	Buttons and links should redirect users to where the links and buttons point to.	Yes
User Profile		
The UI of the page is displayed correctly.	All elements are displayed as intended.	Yes
Page cannot and should not to be able to be accessed by users without logging in.	Redirect users to login page if not logged in.	Yes
The details displayed are correct.	Take data from logged in user's database record and display	Yes
The Password shown is properly hashed and encrypted.	Displayed password should be illogical and different from users' stored password, meaning encrypted.	Yes
The input fields are able to take the data input and validate their format correctly.	Inform and decline username and email data that are not having the required format.	Yes
The uploaded details are successfully updated in the database.	Data send should be updated to the database table of "accounts".	Yes

4.2 User Acceptance Test

The User Acceptance Test is a test to collect opinions from users about how satisfying their experiences with the proposed system are. The test ultimately is to determine how well the system satisfies and fulfil non functional and user requirements. The test is done by sending survey form to users' registered email using the system when it is online. The result of the test is generated and used to built the bar chart below downloaded to administrator's computer. There are still ways to send them their results, such as emailing them to the users. However, it will slowly and inconvenient compared to be able to downloaded straight to their devices. From the collected result above, there is a minor problem

with the download function of the OCR Upload function, as the file would not download directly to their computer, but rather the files are

.

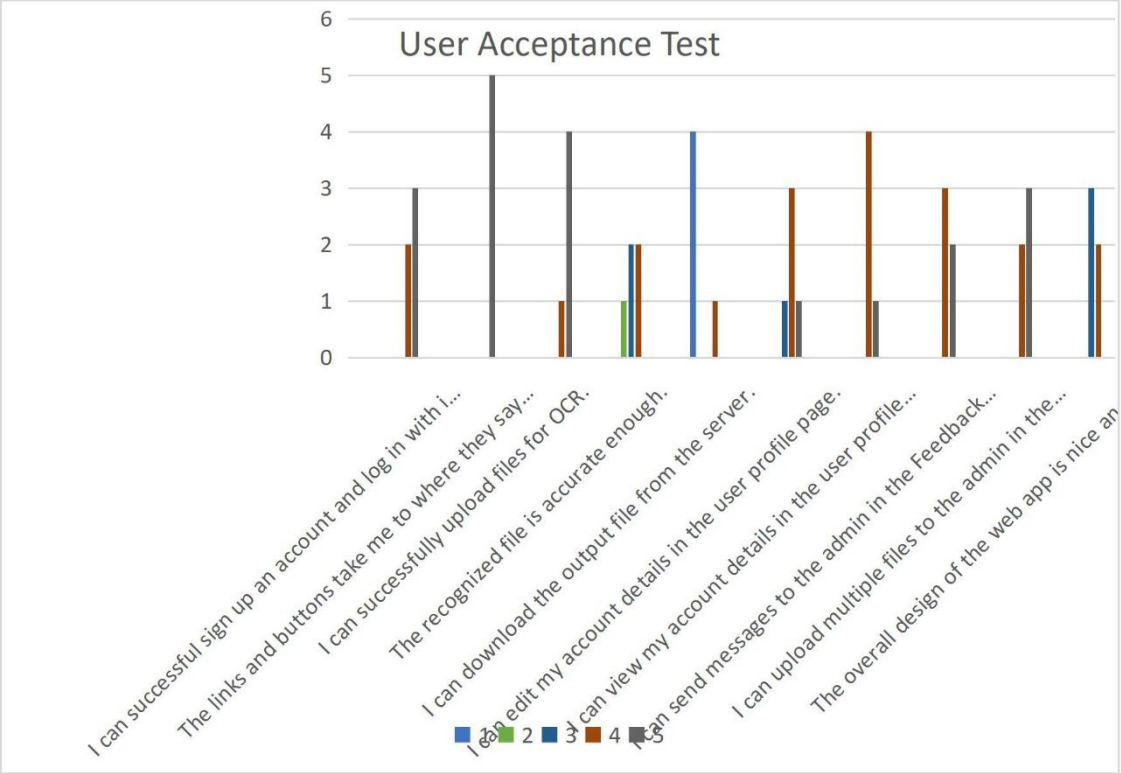

FIGURE 10: User Acceptance Test

4.3 Accuracy Testing

Accuracy testing is test for the system accuracy in certain function. The accuracy testing is done for the Optical Character Recognition's function.The accuracy of the OCR function depends a lot on the font and image uploaded into the OCR engine for processing. Easily recognized fonts (fontssuch as Times New Roman, Arial, Calibri, Bahnschrift) that are not cursive or italic in an image with clean background will almost always have a near 100% accuracy. The image background and text effect will cause a lot of disturbance on the accuracy of the OCR Processing. The original Tesseract engine also has limitation when fonts are rotated and not in a upright position.

TABLE 8: Accuracy Testing Result

Conditions	Accuracy
Easy to Recognized Fonts with Clear Image Background	~~Easy to Recognized Fonts with Noisy~~ Image Background
	Hard to Recognized Fonts with Clear Image

Background	>=95%, theoretically it is impossible to achieve 100% accuracy.
Hard to Recognized Fonts with Noisy Image Background	Less than 85%, depends on how noisy the background is, it might even have a 0% accuracy.
Fonts Not in Upright Position	Less than 80%, depends on how cursive the font and font effects, it might not be able to recognized the text even.
	Less than 50%, noisy background and artistic fonts are even harder to be recognized, most of the time it might not be able to recognize anything.
	Unable to recognize the text, engine will know it is text and characters but cannot convert them. Error will occur.
Image with No Text	No text to be recognized, convert nothing but will not trigger an error.

4.4 Machine Learning: OCR Function with Tesseract and Pytesseract

The proposed system has a function that is capable of machine learning through manual training with sample data, the Pytesseract or Tesseract module, which mainly serves the Optical Character Recognition function of the system. The Tesseract itself is a command line program with libtesseract as the OCR engine. The current version of Tesseract supports 2 types of recognition technology, the newer line recognition and the older character recognition. The code below is the implementation of Tesseract into the system, through the support of a Python Library, Pytesseract.

```
import pytesseract
from pytesseract import Output
```

FIGURE 11: The Python Library for Supporting Tesseract

```
pytesseract.pytesseract.tesseract_cmd = r'C:\Program Files\Tesseract-OCR\tesseract.exe'
```

FIGURE 12: Python Code Line for Calling Tesseract Command Line through Pytesseract

```
141 def ocr_core(filename):
142
143     text = pytesseract.image_to_string(Image.open(filename))
144     return text
```

FIGURE 13: Python Code for Using the Tesseract OCR Engine to Convert Image Line to Editable Text Line

The training of the Tesseract can be done with the Tesstrain program, however it works only on Linuxand iOS distribution of operating system. To enable training on Windows, the usage of Tesstrain

Windows GUI is needed. The GUI is an executable program that can automate the process of training Tesseract once you have gather enough sample data (around thousands or tens of thousands). The end product is a .traindata file that can be placed within the OCR engine to be used. It will not increase theaccuracy to 100%, but with the correct trained data for specific font, it can will increase the accuracy, although the higher the accuracy of a current identified font, the harder is to reach the 100% accuracy and the more data you need to increase the accuracy again.

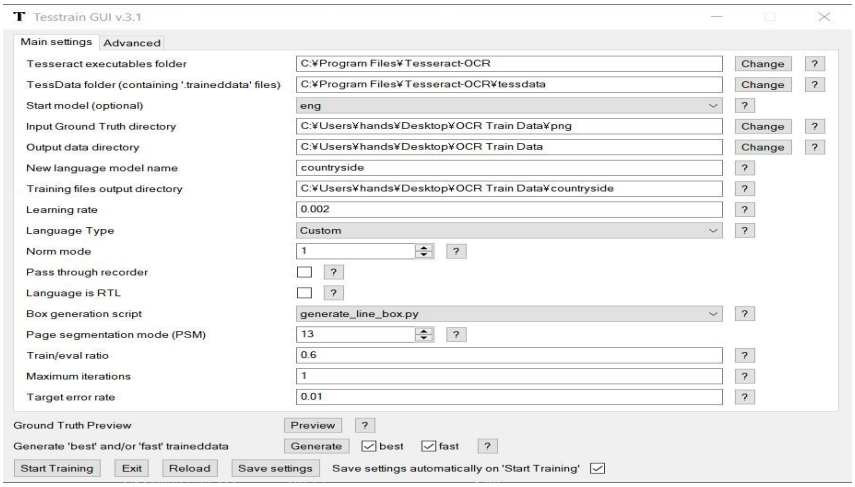

FIGURE 14: Windows Tesstrain GUI

5. Conclusion

In conclusion, the main objectives of the proposed system have been achieved and are mostly a success. The proposed system is able to help authorize users to convert their needed image with text to editable text documents, and allows them to send feedback to the admins while supporting the training of the Tesseract AI by supplying training data and sample data through the feedback and OCR function. However, there are still some weaknesses with the proposed system, including a minorbug that cause clients to not be able to download the files, slower deployment speed and less functionalities provided than applications in the market.

Thus, a few recommendations that can improve the web app further, such as working with cloud servers and services to avoid the bugs, tidying the back-end with better coding, and generally improve the knowledge of development teamto work with the project better.

Acknowledgment

The authors would like to thank the Faculty of Computer Science and Information Technology, University Tun Hussein Onn Malaysia for its support.

References

[1] Taratukhin, V., Yadgarova, Y., & Becker, J. (2018, June). The Internet of things

prototyping platform under the design thinking methodology. In 2018 ASEE Annual Conference & Exposition.

[2] Bimonte, S., Sautot, L., Journaux, L., & Faivre, B. (2017). Multidimensional model design using data mining: A rapid prototyping methodology. International Journal of Data Warehousing and Mining (IJDWM), 13(1), 1-35.

[3] Potnuru, D., & Ch, S. (2018). Design and implementation methodology for rapid control prototyping of closed loop speed control for BLDC motor. Journal of Electrical Systems and Information Technology, 5(1), 99-111.

[4] Pinto, S. C. D., Masson, D., Villeneuve, E., Boy, G., & Urfels, L. (2021). FROM REQUIREMENTS TO PROTOTYPING: APPLICATION OF HUMAN-SYSTEM INTEGRATION METHODOLOGY TO DIGITAL TWIN DESIGN. Proceedings of the Design Society, 1, 1617-1626.

[5] Kara, H., Gergen, K. J., & Gergen, M. M. (2015). Creative research methods in the social sciences: A practical guide.

[6] Pandey, P., & Pandey, M. M. (2021). Research Methodology Tools and Techniques.

[7] Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A systematic review on supervised and unsupervised machine learning algorithms for data science. Supervised and unsupervised learning for data science, 3-21.

[8] Memon, J., Sami, M., Khan, R. A., & Uddin, M. (2020). Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR). IEEE Access, 8, 142642-142668.

[9] Yelton, A. (2019). Introduction to Python programming for librarians [Internet]. Library Juice Academy [cited 6 Aug 2019].< http://libraryjuiceacademy. com/066-python. php.

[10] *Tesseract user manual. tessdoc.* (2015, July 24). Retrieved June 14, 2022, from https://tesseract-ocr.github.io/tessdoc/

APPENDIX A

	Task Name	Duration	Start	ETA
1	Project Planning	30 Days	01.09.2021	30.09.2021
2	Proposal Preparation	40 Days	01.09.2021	09.10.2021
3	Proposal Checking & Amendment	14 Days	10.10.2021	23.10.2021
4	Alpha-Prototype Designing	28 Days	24.10.2021	20.11.2021
5	Alpha-Prototype Testing & Feedback	14 Days	21.11.2021	04.12.2021
6	Beta-Prototype Designing	28 Days	05.12.2021	01.01.2022
7	Beta-Prototype Testing & Feedback 01	7 Days	02.01.2022	08.01.2022
8	Beta-Prototype Upgrade & Bugfix 01	14 Days	09.01.2022	22.01.2022
9	Beta-Prototype Testing & Feedback 02	7 Days	23.01.2022	29.01.2022
10	Beta-Prototype Upgrade & Bugfix 02	14 Days	30.01.2022	12.02.2022
11	Beta-Prototype Testing & Feedback 03	7 Days	13.02.2022	19.02.2022
12	Bera-Prototype Finalizing	28 Days	20.02.2022	19.03.2022
13	Initial Deployment	14 Days	20.03.2022	02.04.2022
14	Deployment Bugfixing and Feature Update	14 Days	03.04.2022	16.04.2022
15	Finalizing and Sharpening	7 Days	17.04.2022	23.04.2022
16	Application Release	Continuous	24.04.2022	-
17	Monitoring and Continuous Patching & Bugfix	Continuous	24.04.2022	-
18	Sample and Material Gathering	Continuous	10.10.2021	-
19				
20				
21				
22				
23				
24				
25				
26				

FIGURE 15: Project Gantt Chart

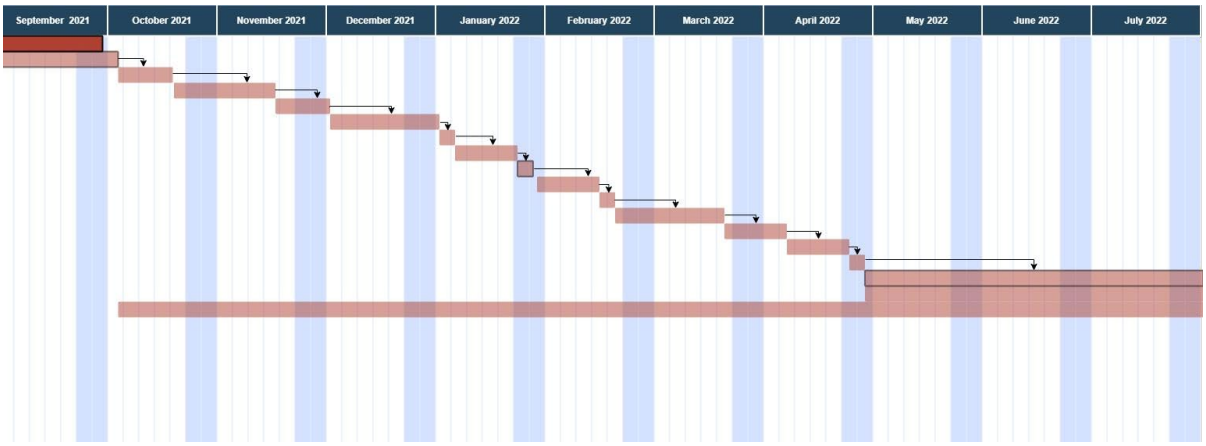

FIGURE 16: Project Gantt Chart(Cont.)

TABLE 9: User Acceptance Test

Testing Questions	Satisfaction Level (Scale from 1 to 5)				
	1	2	3	4	5
I can successful sign up an account and log in with it, without much problems.	0	0	0	2	3
The links and buttons take me to where they say they go.	0	0	0	0	5
I can successfully upload files for OCR.	0	0	0	1	4
The recognized file is accurate enough.	0	1	2	2	0
I can download the output file from the server.	4	0	0	1	0
I can edit my account details in the user profile page.	0	0	1	3	1
I can view my account details in the user profile page easily.	0	0	0	4	1
I can send messages to the admin in the Feedback page.	0	0	0	3	2
I can upload multiple files to the admin in the feedback page.	0	0	0	2	3
The overall design of the web app is nice and enjoyable	0	0	3	2	0