Department of Computer Science & Engineering
The University of Texas at Arlington

Project Charter
CSE 4316: Senior Design I
Spring 2021

Optical Profilers
Document Classifier

Muhammad Daud
Xavier Wells
Bishwamitra Sapkota
Koshish Khadka

Revision History

Revision	Date	Author(s)	Description
0.1	02.14.2021	XW	Document Creation
0.2	03.01.2021	XW,MD,KK,BS	First Draft
0.3	04.30.2021	XW,MD,KK,BS	Second Draft
1.0	8.16.2021	XW	Final Draft

Contents

List of Figures

1 Problem Statement

The primary purpose of the system is to verify and classify different kinds of documents and Identification cards. Classifying thousands of documents manually is very exhausting process. The existing process of manual classification and verification of the documents not only costs a lot of time and money to the company but also is prone to errors. Having a system that can automatically classify and verify the documents on a certain basis, can save time, reduce operational cost and moreover, help the company provide faster and better service to its customers. Additionally, our goal is to move all of the identification process online. This will allow customers to keep an archive of documents and allow the system to have an easier setup as there will be less local setup necessary.

2 Methodology

A solution for this problem is to have a program recognize and process text from the document to classify it. The team will create a user interface (UI) that allows documents to be uploaded in a variety of file formats. The UI will be connected to a web API which can write uploaded documents to a storage location. A document reading model will be created that can recognize a subset of documents. The model will be able to classify a document as recognized or not recognized, and also pull all pertnent information form the document such as name, address, etc.. Finally, the classification result will be stored to be retrieved later.

3 Value Proposition

The value of this sort of application is extremely high. It takes a while to read a document, then enter the information and then at the end of the day find out where it belongs in the storeroom. This project helps the investor one of their most important wealth, their time and money itself. It cuts the time of the information of the document be stored in an easy and timely manner. This will take some time of the employer's hands and help them focus on their other responsibilities. It will also make the money the investor pays to an employer worth it since they would not have to deal with the daily struggles of reading and storing the document. This application will make their daily work extremely efficient.

4 Development Milestones

List of milestones and completion dates:

- Project Charter first draft - Mar 1, 2021
- System Requirements Specification - Mar 22, 2021
- Demonstration of basic back-end code - Feb 2, 2021
- Architectural Design Specification - Apr 9, 2021
- Demonstration of functional user interface - April 14, 2021
- CoE Innovation Day poster presentation(May not be applicable) - Apr 19, 2021
- Detailed Design Specification - July 6, 2021
- Demonstration of website - June 2021
- Demonstration of back end classification - June 2021
- Demonstration of database communication - July 2021

- Demonstration of everything working together - August 2021
- Final Project Demonstration - Aug 16, 2021

5 Background

IDENTI-DOC APPLICATION is a project that highly relies on a technology like Optical character recognition (OCR). OCR is a technology that facilitates user to convert various types of documents, such as PDF portfolios, scanned paper documents, or images captured by a digital camera into editable and searchable data. OCR software operates with users scanner/digital camera to convert printed characters into digital text, allowing user to search for or edit your document in a word processing program. So where is OCR highly used? The most well-known use case for OCR is turning printed paper documents into machine-readable text documents. Once a scanned paper document goes through OCR processing, the text of the document can be edited or can be highlighted with word processors like google docs, Microsoft Word, and many more.
(The Expresswire) – Global "Optical Character Recognition(OCR) Software Market" research report provides key statistics of the market status in terms of Optical Character Recognition(OCR) Software market size estimates and forecasts, growth rate. This report also covers key players of the market identified through their market share, product offerings [1]. By looking at the current OCR demands, high growth rate, and digital advancement this project can make a very long run.
The specific things that the system will do are, if a user misplace or accidentally delete an important digital data, such as a program or invoice, but still have a hard copy, user can effortlessly renew it in their digital filing system by using OCR software to scan the original paper or most recent draft which means there will be no more retyping. The system will be able to get quick digital searches, the system will convert scanned text into a word processing file, letting user to search for particular documents using a keyword or phrase. For example, user can effortlessly search many numbers of invoices and locate a distinct name or report in seconds, without having to go through extensive files one at a time. Other key features of the system will be editing the text, freeing up storage space by scanning paper documents and hauling the originals off to storage.

6 Related Work

Currently, there are multiple companies that focus on document classification software with varying levels of success. For state-of-the-art enterprise software that is commercially available, a good example would be software offered by Iris or ABBYY. In terms of recognition, both of these enterprise level softwares' can recognize handwritten characters, printed characters, checkboxes, barcodes, and much more [5]. However, this software can range from $500 for the barebones, private use software up to $10000 or more for state-of-the-art enterprise software [5]. While partially this price is the cost of implementation, another reason is the technology used.

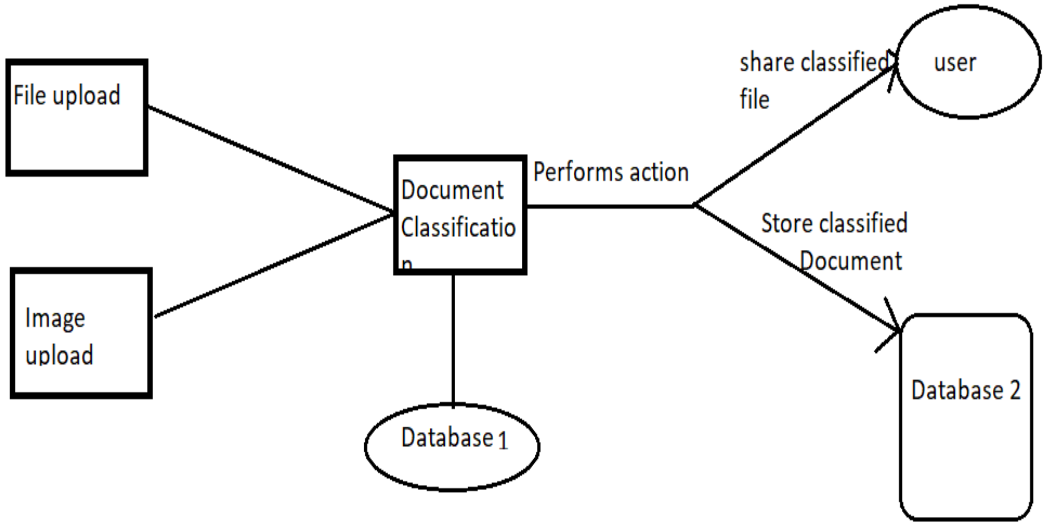
An important distinction between these document classification software technologically is what category of classification is being used; one without a model or with a model. Without a model, known as unsupervised method, using the technique clustering, a form of data mining. Clustering is commonly used in commercial software with two primary algorithm types: partitioning or hierarchical. Partitioning is easy to implement but is sensitive to background noise and has a hard to compute time complexity, though it is generally less than $O(n^2)$. Hierarchical algorithms tend to be sensitive outliers, and the time complexity is generally $O(n^2)$, making them not suitable for large datasets [3]. Both algorithm types are strongly influenced by the order of a dataset, making them less accurate than other methods [4]. Approaching classification with models, also known as the supervised method, results in a much quicker but harder to implement solution. Extensive training is needed to have model-based software work under varied conditions, resulting in the higher prices seen in professional, enterprise software [3]. An off the shelf model-based software is far too expensive for many clients, and clustering methods may be too inaccurate or slow.

Furthermore, differences in character recognition affect the accuracy of the classification discussed above. Currently there are many libraries for character classification, Tesseract being a well known API. Tesseract is free, and unfortunately that comes at the price of knowing less languages on initial deployment than software from ABBYY or Adobe [2]; however, Tesseract is open source and allows the user to modify the API for their uses. This means that while larger organizations may use ABBYY or Adobe to process language, Tesseract can be built up to process language just as well. Tesseract has the possibility of surpassing the more commercial APIs as well with the recent implementation of LSTM(Long short term) neural networks [6]. With this API it will be possible to outperform the commercial software available in character recognition.

7 System Overview

The primary goal of the system is to identify/recognize text on the file and classify it into specific documents. The system that can take file as input and give properly classified output documents is our challenge. To deal with this challenge, we are thinking of implementing a specific solution. A high-level approach of building a system for document classification is discussed in this section. First approach to the solution is figuring out how the document is input to the system and the types of the files that the system can accept for processing. System will allow user to upload files from the computer. User also should be able to take a picture of the document and proceed it for classification. After the input, the most important process is initiated. The system will read/recognize the document ,which can be in multiple form. Not only the system recognizes the document, but it will also analyze and interpret it. Natural language processing tools can be applied. One of the possible technologies for this process is implementing OCR, which converts picture of handwritten or typed text into machine encoded form. As system is basically acting as text analyst, another technology can possibly be used is Text Mining, which discover important information form the text comparing it with previously know and stored information. This approach generally involves categorization of text. The system will be capable of understanding the file, its content and providing well classified result. The system also processing images. It is necessarily reading and extracting information from images as well. After implementing one of these approaches, user will obtain the result and it can be share or store electronically. Documents have been created, studied, and shared for different purposes and widely used in almost every field. The system is undoubtedly very important to human civilization in modern world.

The entire system can be divided into three parts, first part involves input of files/document, second part involves reading/analyzing and classifying the document and the last part involves extracting well classified documents. The high-level representation of these components as follows.



8 Roles & Responsibilities

The UTA development team of document classifier has four team members Xavier Wells , Kosish Khadka, Muhammed Daud and Bishwamitra Sapkota. The stakeholder is authority, who provides resources to the team. At this point, University of Texas at Arlington is considered as stakeholder for our project. We, Xavier Wells , Kosish Khadka, Muhammed Daud and Bishwamitra Sapkota as team members already have established a good communication among us and already started to research and gather information as possible. We are four people working collaboratively and dividing works based on our interests and skillset. Team members are the ones who work on project on regular basis. We are figuring out specific approaches and methods to proceed the development in organized manner. Each of us are excited to work together and achieve the common goals. We have understood the responsibilities from individual point of view and as a team. We, as a team member must take an initiative to figure out challenges and solutions to those problems that comes along the journey. The roles of a team member also include listening to other members and respecting other perspective as well. A good team member should make sure all of us are on the same pace and continuing the work. We have not selected a scrum master yet. One of the reasons of not having a project manager is all of us must take equal initiative, responsivities, and accountability. We are optimistic to be able complete this project successfully.

9 Cost Proposal

The approximate budget of the project is $800 which is provided by the CSE department. The major expense will be on electronics like scanners, digital cameras, and software licenses.

9.1 Preliminary Budget

Description	Total
Grand Total	$800
Software	$300
Electronics	$500

9.2 Current & Pending Support

Currently, CSE Department is the only one who is funding the project and there are not any potential additional pending funding sources for this project yet.

Description	Status	Amount
CSE Department	Pending	$800
Grand Total	X	$800

10 Facilities & Equipment

We do not really plan to go require a lab space, but we might need to book a room on campus to have in person meeting and discuss future plans once every other week. We will need some sort of equipment to scan a document whether it is the phone camera or a camera attachment device for the computer. We can require our hardware from UTA if they have any laying around, if we use our phone cameras, we can just program an application on our devise to scan the documents in that way, and if we use a camera attachment then we would purchase one from the local store or amazon if needed. For the software part of the project if we are unable to meet due to covid-19 restrictions we will use visual code studio to live share the code we are working on. For the software part of the project we plan to use python and its applications since we are very similar with its APIs and packages.

11 Assumptions

- A finite number of different types of documents will need to be classified by the system.
- An open source API will be able to be used for Optical Character Recognition(OCR)
- The customer will provide templates of the documents that they wish to be characterized
- The customer will have access to a machine that can run the software
- The installation site will have a network connection available to download updated models

12 Constraints

The following list contains key constraints related to the implementation and testing of the project.

- Web page and initial software implementation must be completed for a demonstration by May 3rd, 2021
- Implementation must follow any data privacy laws, especially in regards to training of models and internet connectivity

- Images must have only a low to moderate amount of "noise"; that is, background and foreground of documents must be distinguishable
- Total development costs must not exceed $800
- Project may be required to work and be used offline

13 Risks

Top 5 risks:

Risk description	Probability	Loss (days)	Exposure (days)
Feature implementation failure	0.50	25	12.5
Inflation in requirements/changes	0.40	16	6.4
Team member's schedules do not align.	0.30	15	4.5
Member turnover	0.10	30	3.0
Assuming app to be perfect	0.20	10	2.0

Table 1: Overview of highest exposure project risks

14 Documentation & Reporting

14.1 Major Documentation Deliverables

14.1.1 Project Charter

This document will be updated as new information is made available as well as each sprint. Specifically, when a sponsor is specified for this project this document will be updated. Any major changed in team makeup will result in a change as well. For each sprint, more detail will be provided on policies for other documentation deliverables below.

- The initial version will be published on: 2/27/2021
- The 2nd version will be published on: 5/4/2021
- The final version is scheduled for an undetermined date in July, 2021

14.1.2 System Requirements Specification

This document will be updated as new features are requested, as well as when any changes to existing requirements or features are changed. This includes user interface changes as well. If there are changes to any tentative dates, those will be updated.

- The initial version will be published on: 3/22/2021
- The 2nd version will be published on: 5/4/2021
- The final version is scheduled for an undetermined date in July, 2021

14.1.3 Architectural Design Specification

This document will only be changed when major system changes occur or a new layer must be added to accomplish a new feature or requirement. Ergo, this document will be updated primarily when the System Requirements Specification is changed.

- The initial version will be published on: 4/12/2021
- The 2nd version will be published on: 5/4/2021
- The final version is scheduled for an undetermined date in July, 2021

14.1.4 Detailed Design Specification

This document will be updated periodically throughout the sprint as code changes. As this document will be completed much later in the project life-cycle, these changes should be minimal. Primary examples of code changes that would effect this document are changes to or new functions, libraries used, or data structure changes.

- The initial version of this document will be completed in early July.

14.2 Recurring Sprint Items

- Sprint Planning PowerPoints
- Sprint Review PowerPoints
- Engineering Notebooks
- Updates to the Project Charter
- Sprint Retrospectives

14.2.1 Product Backlog

The product owner will have a role in what gets moved from the product backlog, as they will determine what features they would like delivered earliest. From there, backlog items will be added from the SRS with an interface first, function second approach; however, The decision will be held to a group vote and liable to change. Currently, a virtual spreadsheet will be used to share the product backlog with both stakeholders and team members.

14.2.2 Sprint Planning

Each sprint will be planned in a team meeting to develop the backlog. This project will be finished in 7 sprints.

14.2.3 Sprint Goal

The product owner will ultimately decide the sprint goal, but the team will tentatively propose a goal during the planning meeting.

14.2.4 Sprint Backlog

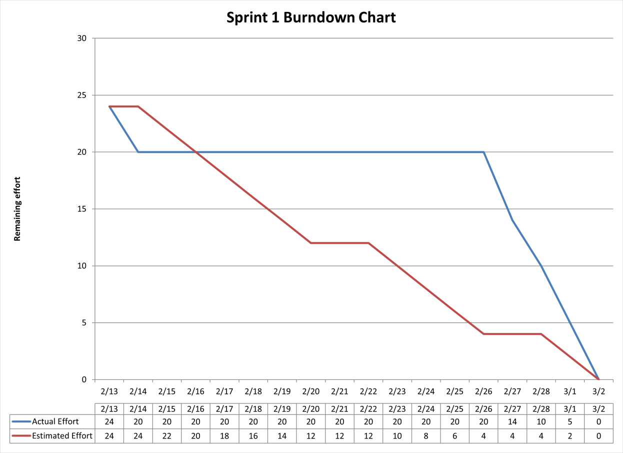
The sprint backlog will be compiled as a team in the sprint planning meeting. This will be shared and maintained through an excel spreadsheet, similar to the product backlog.

14.2.5 Task Breakdown

While team members are encouraged to voluntarily claim a task, if any tasks or persons are not assigned by the end of the meeting a team member will assign the rest of the tasks. Time spent on tasks will be documented in engineering notebooks and digitized weekly during the sprint.

14.2.6 Sprint Burn Down Charts

Each sprint, 1 team member will be assigned to make the burn down chart. This member will be able to view the hours put in by various team members digitally, as individual team members will upload their total time spent on tasks in Microsoft teams under the channel for the current sprint. The burn down chart will be based on this template: The red line represents a linear and ideal burn down; The blue line is the actual burn down. The x axis will represent the date, the y axis will represent the hours remaining this sprint.



	2/13	2/14	2/15	2/16	2/17	2/18	2/19	2/20	2/21	2/22	2/23	2/24	2/25	2/26	2/27	2/28	3/1	3/2
Actual Effort	24	20	20	20	20	20	20	20	20	20	20	20	20	20	14	10	5	0
Estimated Effort	24	24	22	20	18	16	14	12	12	12	10	8	6	4	4	4	2	0

Figure 1: Example sprint burn down chart

14.2.7 Sprint Retrospective

The sprint retrospective will be done in tandem with the sprint review PowerPoint. This will generally be completed the day the sprint is completed.

14.2.8 Individual Status Reports

At the end of each sprint there will be an official individual status report. Key items will be tasks completed by the team member, actions by the team member that can be improved, begun, or ended, and a review of peers. This report will also include the sprint backlog and burn down chart.

14.2.9 Engineering Notebooks

Engineering notebooks will be updated at the discretion of the individual team member; however, it is expected that any work done in regards to the project should be written in the notebook and notated with the current date and time. There should be at least a half page written by the end of the sprint. Team members will keep other team members accountable by reminding them during meetings to keep track of their engineering notebooks.

14.3 Closeout Materials

- The Project Charter
- The System Requirements Specification

- The Architectural Design Specification
- The Detailed Design Specification
- The Documentation PDF
- The User Manual
- The Project's Source Code
- The Installation Script

14.3.1 System Prototype

The final system prototype will include all desired, high priority features for the application. This will be demonstrated to the product owner in June of 2021. After testing on a development machine, testing will be demonstrated off site on the targeted machines.

14.3.2 Project Poster

The poster will be a 24"x36" poster. This poster will display the user interface of the software as well as a graphical representation of the steps to processing a document including an input document, the document with removed noise, a representation of the classification model, and finally the classification of the document. Additionally, we will have the methodology for the project displayed on the poster.

14.3.3 Web Page

The first version of the web page for this project will be delivered on May 3rd, 2021. This website will include the abstract and background for our project, as well as the requirements and system overview. The web page will be publicly accessible and updated throughout the project, a demonstration video will be added to the web page at closeout.

14.3.4 Demo Video

The demo video will be around 10 minutes long and give an overview of the software. It will show a document classification in real time and explore the user interface to show all available options and functionalities of the software.

14.3.5 Source Code

Source code will be maintained in a private GitHub repository. Source code will be turned over to the product owner by allowing them access to the repository. The product owner can then clone the repository as they wish. Alternatively, the source code will be provided as a .zip file.

14.3.6 Source Code Documentation

Tentatively, documentation is planned to be accomplished with doxygen exported to LaTeX and published to PDF.

14.3.7 Installation Scripts

There will be a single installation script that accomplishes the setup.

14.3.8 User Manual

The customer will be provided with a digital user manual in the form of a PDF file. No video will need to be provided to setup the software.

References

[1] Expresswire. Optical character recognition(OCR) software market 2021 top trends, size, scope, share, development status, opportunities, growth, statistical analysis and forecast to 2025, 2020.

[2] Illinois Library. An introduction to OCR - introduction to OCR and searchable PDFs - LibGuides at university of illinois at urbana-champaign, 2021.

[3] Madjid Khalilian and Shiva Hassanzadeh. Document classification methods. Technical report, Islamic Azad University, karaj branch, 2019.

[4] Marina Santini. Advantages & disadvantages of k-means and hierarchical clustering (unsupervised learning). *http://santini.se/teaching/ml/2016/Lect_10/10c_UnsupervisedMethods.pdf*, 2016.

[5] Meta Enterprises, LLC. Compare scanning software - compare forms processing | compare document imaging | compare invoice processing, 2021.

[6] tesseract-ocr. Tesseract user manual | tessdoc. *www.tesseract-ocr.github.io/tessdoc/*, 2021.