University of Central Florida

Text Cropping System

Sponsored by Amy Larner Giroux, PhD.

Jacob Rodriguez	jpr@knights.ucf.edu
Brian Smith	brian.smith@knights.ucf.edu
Miguel Severino	mseverino@knights.ucf.edu
Vincent Cardaman	vcardaman@knights.ucf.edu

COP 4934
Mark Heinrich, Rick Leinecker
April 21st, 2020

1. Executive Summary

This project is called Text Cropping using Machine Learning and Computer Vision. The purpose of this project is to create a program that will automatically find an entry word in a dictionary and crop it along with its definition, then save the image file of the cropped definition and name it according to the headword that was cropped. The team members are: Miguel Severino, Vincent Cardaman, Brian Smith, and Jacob Rodriguez. The sponsor of this project is Amy Larner Giroux, PhD., The associate Director of the Center for Humanities and Digital Research.

The dictionary that the definitions are being cropped from is Samuel Johnson's dictionary, one of the most influential works of english literature. This dictionary is extremely important to the history of the English language because it was one of the first, very comprehensive dictionaries that included both every different definition for each word, and examples of the word in use in other pieces of literature. Digitizing any document is important for preservation, storage, and ease of access. The dictionary definitions are being cropped in order to retain the feeling and appearance of one of the physical copies of the dictionary. Prior to our group taking on the project, Dr. Giroux had a team hired to crop each definition by hand, and with over 42,000 words it would take a very long time to complete. Machine learning and computer vision will be implemented in order to speed up the process greatly.

The images are stored in a high quality .TIF format in order to retain as much of the quality as possible. Many of the pages are warped or rotated in many different ways, and they must be straightened before being cropped to make any subsequent image processing easier. Whitespace, guide words, signature marks, press figures, and catchwords must be removed as they are not part of any definitions. Many definitions span multiple pages or columns and must be stitched together after cropping to complete the definitions. Any entries that have been stitched together must undergo color correction to keep the text readable and neat. Finally, the files must be named according to the word that was cropped. The final project must be delivered via Docker for simple and easy replication and to remove any runtime environment errors.

2. Project Overview

In this section, we provide a brief overview of the technical aspects regarding our project. This includes a greater description on what exactly the project is and what we are trying to achieve, a statement from each of our group members on our thoughts about the project and why we chose it, the exact requirements that define our project and goals, our ideas on how we might solve some of the different issues that we expect to see throughout the course of this project, and charts and diagrams that both expand upon our project in a visual manner, and provide more depth and insight into our plans for both the current and upcoming semester of our project.

2.1 Project Description and its Significance

The software we are designing will be able to detect distinct changes in word format to identify markers for cropping images into individual entries. It will be trained via machine learning on a database of over one hundred manually cropped words already entered into the site. Additionally it will automatically rotate the text to straighten it for easier readability. The software will also remove various unneeded components of the full image to create a database of entries without excess information in them beyond the definitions within them. These unneeded components may be page numbers, guide words, or any additional text that is not a part of a word or definition. If successful this project will save hundreds of hours of work on a cultural passion project. It will make an important document of literature far more accessible to the general public and allow for searching through a historical piece of literature with nearly the same ease one expects from a modern digital dictionary. It will be an excellent example of how computer science can be applied to improve work on projects old as well as new. Not only does the significance lie in the field of computer science, our work will also help in the preservation of one of the most influential works of English literature and any and all documents from both the past and the future. Digitizing books is an extremely important task, as the digital medium has a much longer life than physical mediums. Many pieces of writing have been lost to the ages before it was possible to scan or even record photographs of them, since copying any piece of writing had to be done by hand, only scrolls or books that were considered to be of great importance were copied.

2.2 Broader Impacts

The solution to this project will be a low cost, scalable solution that can help not only in the historical preservation of documents, but also in the digitizing of text images on a large scale. With this technology, many of the large-scale museums, such as the Smithsonian Institution, could digitize their historical documents more efficiently. Historical documents need a better and faster way to be preserved and retain a good appearance. Not just historical documents are affected by this, scans of newer books are still lacking in many aspects. With this newly preserved material, interactive learning experiences could be created for museums and additionally be added to school curriculums to help teach children both nationally and internationally.

In addition to the possibility of using this for research and preservation of significant materials, this solution can be scaled to any document that needs to be scanned and preserved for any reason. It could be used for staff at doctor's offices to be able to view relevant information in a patient's files. Human resources employees at companies across the world could use the software to keep track of employee records for everyone within their company which is useful for both small companies and giant corporations. This would be immensely helpful for lawyers, being able to parse through a document to find a relevant statute, law, or even elements within a written statement to find what they need to complete their case could save hours of work and effort.

2.3 Personal Motivations

In this section each member explains their personal motivations for choosing to work on this project. They may also explain who they are and go into some level of detail about themselves personally.

2.3.1 Jacob Rodriguez

A large part of why I chose the Text Cropping System project is due to the fact that it seems both challenging and interesting to me. I wanted a project that involved something I had never done before, and machine learning seemed like just the right topic to that end. Machine learning feels like one of those topics that is very intimidating, even just the name of it sounds like it would be difficult. However, I am fairly certain I am going to find out that it might be simpler than it seems. Not only that, but the dictionary itself simply seems like a very interesting topic to work on. Digitizing a piece of history not only sounds impressive, but it seems like something that I would want to show off and be proud of to both my friends and my family. The dictionary is very important in and of itself as well, so providing a way for anyone who desires to see this dictionary but is unable to get their hands on a copy is a very nice bonus. It helps that the project is cropping the images of the pages out, because keeping the feeling of a physical book through the digital medium is a nice touch to the project and gives it more meaning. After all, if simply reading the definitions is all somebody wants to do then there would be much easier ways to go about it, but with this project I hope that they can also get a feel for the physical dictionary itself and how it was printed and written. A good example of this is the dictionary writing the letter ‘s’, because they use a symbol that looks very similar to the letter ‘f’. This is something that anyone who is simply reading OCR’d text of the dictionary will completely miss out on. My greatest motivations for this project are having something impressive for my resume, gaining a lot of knowledge about working with machine learning, and expanding my skills on working with collaborative tools such as GitHub and Trello. After all, the most important thing I can get out of school is the actual knowledge involved with the projects, how to program and make something meaningful that works. Additionally, it is very important that I have more examples of my programming ability for my employers. If I have nothing of value to show to an employer then they will have no concept of my programming ability. I want this project to be something I am proud to show off to my employers, and something I can show off to my friends and family. Machine

learning will hopefully be impressive to any employers I am trying to work with. As far as GitHub and Trello go, I feel as though I do not have enough experience working with them yet. I have done multiple projects at the University of Central Florida that involve them, but I find myself having to relearn how exactly to use them, and what commands to use to commit changes. My past projects have simply been for a good grade, but this project will be a lot more than that. It is a serious project for both our sponsor, Dr. Amy Giroux, and the team she is currently working with on it. I want this project to have an actual impact on the real world. It does not have to be millions of people, but if my work on this project significantly helps Dr. Giroux and her team, it will give my work a lot more meaning. Overall, this project, as opposed to the others, felt like it would be both a good challenge and learning experience, while also being one of the most interesting.

2.3.2 Brian Smith

My main interest in the project is in the historical preservation of such an important contribution to the English language. I believe that historical preservation is one of the most important aspects of a culture. Many people want to move on from the past and forget it, however it is extremely important that we never forget, so that we do not make the same mistakes again. I also have an interest in linguistics and the history and progression of languages. Being able to work with such an influential dictionary and see how our language has changed over time is extremely fascinating to me.

In addition to my personal affinity for historical preservation, I did not want to do a web-based application with a mobile app, backed by a database. This project will make use of certain machine learning/computer vision technologies and I have an interest in those fields. I have a very basic knowledge and understanding of machine learning/computer vision from previous and current coursework and I felt this was a great opportunity to enhance my knowledge. Machine learning and automation go hand in hand. I am learning about automating scripts and operations on a large scale at my current internship so I felt this would also help increase my knowledge on automation. With this new knowledge, I hope to be able to apply it to personal projects as I have an interest in big data and modelling that data. Lastly, this project is an impressive thing to have on a resume since machine learning is a burgeoning field and the employers I hope to work for are always looking for people with skills in those areas.

2.3.3 Miguel Severino

The first motivation for this project is to learn how to implement Machine Learning algorithms to a real-life problem and see a large project from start to finish. I also find it fascinating how computers are playing a significantly bigger role in our daily lives and how quickly technology is advancing. Further, I've used software involving Optical Character Recognition and was amazed at how it was able to extract text and digitize it. After seeing how Optical Character Recognition operated and how it was able to extract data from an image, it fascinated me, and I wanted to learn more. Lastly, I feel that this project will allow me to increase my understanding of the technologies and subsequently apply it to this project.

Further, the field of Computer Vision and Machine Learning are amongst the cutting edge of Computer Science and I wanted to inform myself more on the subject, in addition to applying Computer Vision and Machine Learning techniques to a real-world problem. Something that piqued my interest to work on this project was that it would allow me to aid in the preservation of a historical text. And therefore, become a keeper of information and preserver of knowledge. In closing, I firmly believe that we can continue advancing as a society because of small contributions by individuals that are accumulated over time. And I feel that by helping preserve this historical text, I'm able to make a noticeable contribution towards advancing historical information and informing future generations.

2.3.4 Vincent Cardaman

When I took UCF's robot vision class I was disappointed how little it forced you to learn how to actually create a program to analyze an image with the software. It was more of a guided tour through premade image detection software. This project was an opportunity to work through making my own detection code, and for something that perfectly demonstrates the advantages of robot vision and machine learning. Our work on this project will likely save countless hours of work of slow manual cropping for the members working on creating an archive of this famous set of dictionaries. It is exactly the kind of task automation was designed to help with and the completed software in a day could do well over months of work. This project will help complete a database of entries from an incredibly well known and effective dictionary. In addition it will finish a passion project taken on by many with month's of work behind it.

2.4 Legal, Ethical, and Privacy Issues

There are not any legal, ethical, or privacy issues with our solution. The software we plan to use to develop the program is entirely open source so no legal issues surround using it. Our solution will take already existing information and simply make it into a format that is more easily accessible, there is not anything about how it functions that would cause an ethical issue. The data that our code is meant to be used on is already available in the public domain. We will not be taking information kept privately at any point for this project, everything our solution works on is already public, we are merely making it far easier to find what someone wants in that information. It is already on a pre-existing site on which we will simply be putting the more accessible version of the information for others to look at. As such possible concerns in the security of the website for which the documents our code is meant to crop have already been handled.

2.5 Specifications and Requirements

The following specifications and requirements have been agreed on by both every member of our team and our project sponsor. The base requirements are the requirements that we must have in order for the project to be considered successful by our sponsor. The stretch goals are features that may be implemented if we have enough time after completing all of the base requirements.

2.5.1 Base Requirements

- The software must automatically crop images of a dictionary page into individual entries.
- Each entry must be saved into a .tif file under the naming system explained below
    - A letter representing the format, “f” for “folio” for instance
    - The year the dictionary edition is from in question e.g. “1755”
    - A hyphen(-) followed by the word this entry is defining
    - So a complete file name would be “f1755-seldom.tf”
- Entries which span multiple pages must be saved as a single .tif file one column width wide and as long as is needed to contain the fully entry.
- Each image must have it’s words rotated to reduce slanting both at the page level and the word level.
- The .tif files of entries must be saved into a directory arranged by edition (1755/1773) and in alphabetical order of the entries.
- The completed program must be sent with a docker file for the runtime environment.

2.5.2 Stretch Goals

Various pages have been distorted during scanning, making them blurred or dimmer than the rest of the text. Implement an image enhancing program to automatically find these sections and sharpen the text while adjusting the brightness to improve readability. Additionally, some of the definitions stretch from one of the pages over to the next. We would also like to be able to match the colors of the two different pages when we combine them into one cohesive definition.

2.6 Initial Ideas

Some of the pages of the dictionary are curved due to the paper sometimes not being flat on the scanner during the scanning process. This means that in order to straighten the pages, you have to do more than simply rotating the entire page as a whole. We have thought of a few potential solutions to this issue. The most obvious solution is to take out chunks of text and rotate them on that level, since the pages are typically curved very gradually as the page goes on. However, the chunks of text that we take out could be any size. You could take each individual letter and rotate them until they're straightened, then stitch them all back together, but due to the pages typically having a gradual curvature along an entire page that results in this being unnecessarily difficult and overcomplicated. We will be testing taking chunks of different sizes to see what has the best results and performance. We want the algorithm to be as smooth as possible, but also ensure that the output looks as straight as possible. For this reason, we will likely need to find a middle ground that leaves the algorithm both simple and ensures that the output looks excellent. As far as my current thoughts, since many pages do not have any tilting, and on the pages that do have it, it is very gradual, we will likely not need to go very far down. I imagine that even splitting the page into fourths will be enough to remove nearly all of the tilting issues that we have.

The website that we are cropping the dictionary pages for has its own searching features built in. Since the dictionary has a lot of space on each page dedicated to being able to physically search through the book, such as page numbers and guide words, the images need to be cropped. These are not necessary for the site, and only serve to clutter the pages of the dictionary. The definitions need to be neatly cropped out with no additional clutter or unnecessary whitespace. Our machine learning model will need to identify page numbers, guide words, and what exactly is a part of a definition in order to work.

In order to make it easy to pull out each and every definition, we must first look at what makes each definition unique from any other definition. The very first unique aspect of each definition is the headword. The headwords are both further to the left than nearly any other item on the page, and they are entirely written in capital letters. This is useful for both picking out where each definition starts, and for naming each and every one of the images of the cropped definitions, as we need to include the headword in the filename. The second thing that makes each definition unique is the definition itself. These lines are set apart from the next

headword because they are indented slightly to the right. This allows us to more easily pull out each definition, and compare it to the next headword which will not be indented at all.

Another idea is on finding a viable way to correct for warping of the scanned documents. The best technique of attaining a viable solution to this problem is by fitting an elastic curve or surface over the text line. This seems to be the best approach, since it does not require any special equipment, nor special devices. We can perform this by fitting a quadratic function over the text line and taking in account the spaces between the words. While also fitting the visible straight, not warped, text. After we fit both warped and the not warped text, we will be able to compare the two and by doing this, it will give us a robust estimate on how warped the text actually is compared to the known, and not warped, nor skewed lines.

In addition to warped documents, the documents could sometimes be skewed, which would need to be corrected. In this case, a similar procedure will be performed, but instead of individual text lines being fitted, like in the previous procedure, a multi-line segment would be accounted for and considered at once. This multi-line segment of text would then be fitted with a linear equation, which would subsequently be compared with a first-order linear equation. By making these comparisons, we would be able to measure how skewed the text body and could then

For every entry in the dictionary the books use a set of letters to denote if the word is noun, verb, or adjective. Because of this we can set the program to look for these specific sets of letters which are italicized when looking to find entry words. This classification system is quite distinct and should not appear anywhere else on a page except after an entry word to denote its type.

Whenever a new entry starts in the text it is always on a new line, this means that we can use abrupt line ends as a possible guideline for predicting when entries will start. These are places where our software would already check anyways, however if the previous line ended abruptly from a new line it could trigger the program to run additional tests on the next line to ensure it doesn’t miss an entry.

The dictionary includes several examples of the words it defines in use along with their definition. This means we can ignore any line which starts with a number as that would mean it is an example of the word in use. Hopefully by doing so we can optimize use and also decrease false positives.

On every page the text comes with guide words in a much larger font than the rest of the page, these are to easily allow the reader to know which words they can expect on this page. However for the purposes of our program these are completely unnecessary. By limiting the size of words our software will detect we can ignore these completely.

Because entries always start their own line, there isn't actually any reason to check any word after it. If our software can still accurately crop the pages into entries while only running detection on the first word of each line it will massively improve the efficiency of the program. It also serves to remove more chances of error, the more words our software needs to check in the books to find the entries the more chances it has to get a false positive result.

Another metric that we can use to help our program correctly capture text from the dictionary is to disallow signature marks that will be on the bottom of the page from being read as relevant information to the definition.

3. Block Diagram and Figures

This section describes our project flow as well as the components necessary for us to complete the project. Our plans for both, how we would like the project to run, and what aspects we would like each group member to focus on have been laid out in the graphs. Our project only relies on a few pieces of software, so we want to ensure that we are choosing the most efficient and user-friendly options. Since we will be implementing open source software and there is no shortage of open source options available for what we need to do, we need to research every available option and determine which suits our project design and workflow the best.

3.1 Program Execution Flowchart

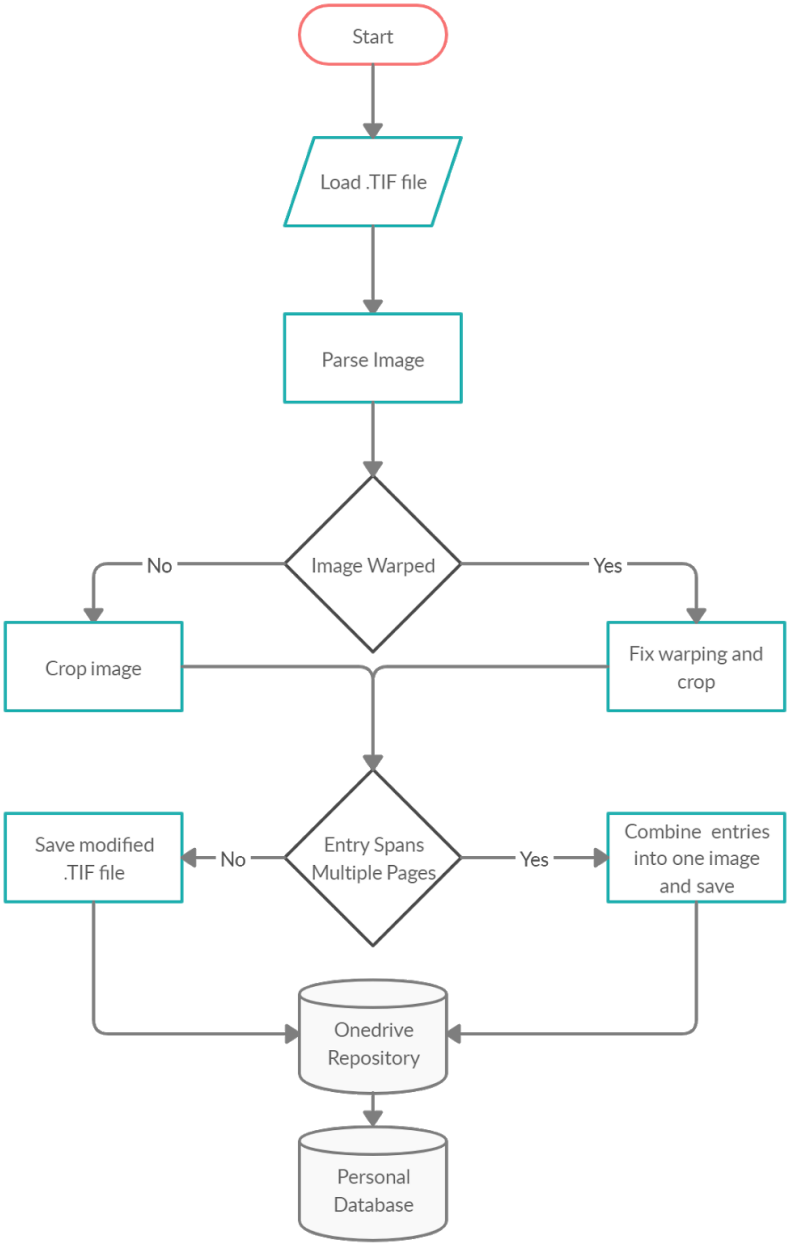


Figure 1: Program execution flowchart

This flowchart describes the basic overview of how our program is going to run. It starts when the user runs the program, which opens a GUI. The GUI prompts them if they want to process one image or process an entire directory of images. From there the image(s) are loaded and parsed. The first step is to

determine whether the image is warped or not. If the image is warped, we fix the warping using the dewarping algorithm. From this step the image is broken down into two columns, representing each side of the page. It is then further broken down into each definition. The next decision is to determine if the entry spans multiple pages or not, and finally upload the image to the repository and our own database.

3.2 Block Diagram

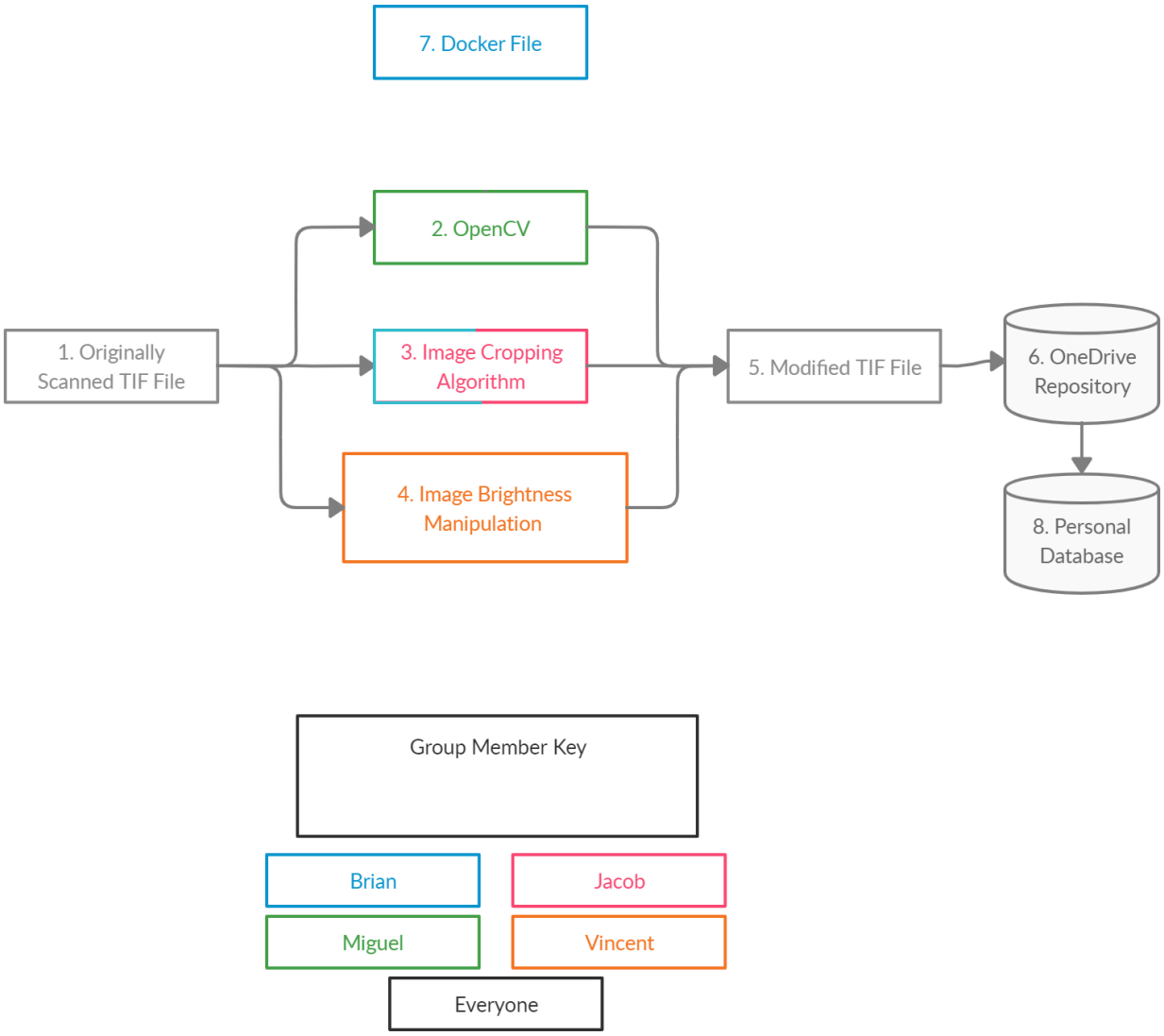


Figure 2: Project design block diagram

Block Status:

1. Acquired: The sponsor provided us with the .TIF scans from her repository.
2. Prototype: After researching, we settled on using OpenCV. At first, we believed that we had to implement some sort of machine learning library but discovered that OpenCV has everything that we need. We have come up with a process to segment the pages and Miguel is prototyping that process currently. In addition to the OpenCV implementation of segmenting the page, Miguel has designed a prototype GUI that will be presented to the user upon launch of the program.
3. Prototype:
    a. Image segmentation: Each page is segmented into two columns, with a gap in between them. Our prototype currently breaks each page into two separate images, a left and right column image. It then scans the images for the definitions and saves the definition in its own .TIF file.
    b. Image warping: At first, we believed image warping to me our most complex issue, fearing that we may not be able to implement it. We found a research project from Dr. Zucker at Swarthmore College that does most of the work already. We now need to modify the project to maintain the original image, instead of turning it into a black and white scan.
4. Research: At first, we believed this would be a stretch goal, but after making progress on the other components, we realized that this can become a core requirement. Because of this, we are still in the research phase, but hope to make enough progress to soon implement this into the current prototype we have.
5. To be acquired: Once we can create our algorithms and test modules, we can run them on the original .TIF files and obtain these modified files.
6. Acquired: The sponsor provided us with her Onedrive repository that she previously had.
7. Research: We need to deliver the final solution in a Docker file, so that for anyone wanting to run this algorithm in the future, they can easily mimic our setup and get it running without fault.
8. To be acquired: After speaking to Dr. Heinrich, we have decided to keep a small database to help organize our images. This will help when we find certain anomalies that will help us when debugging our algorithm as well as when we present our project in SD2.

3.3 Gantt Chart and Division of Labor

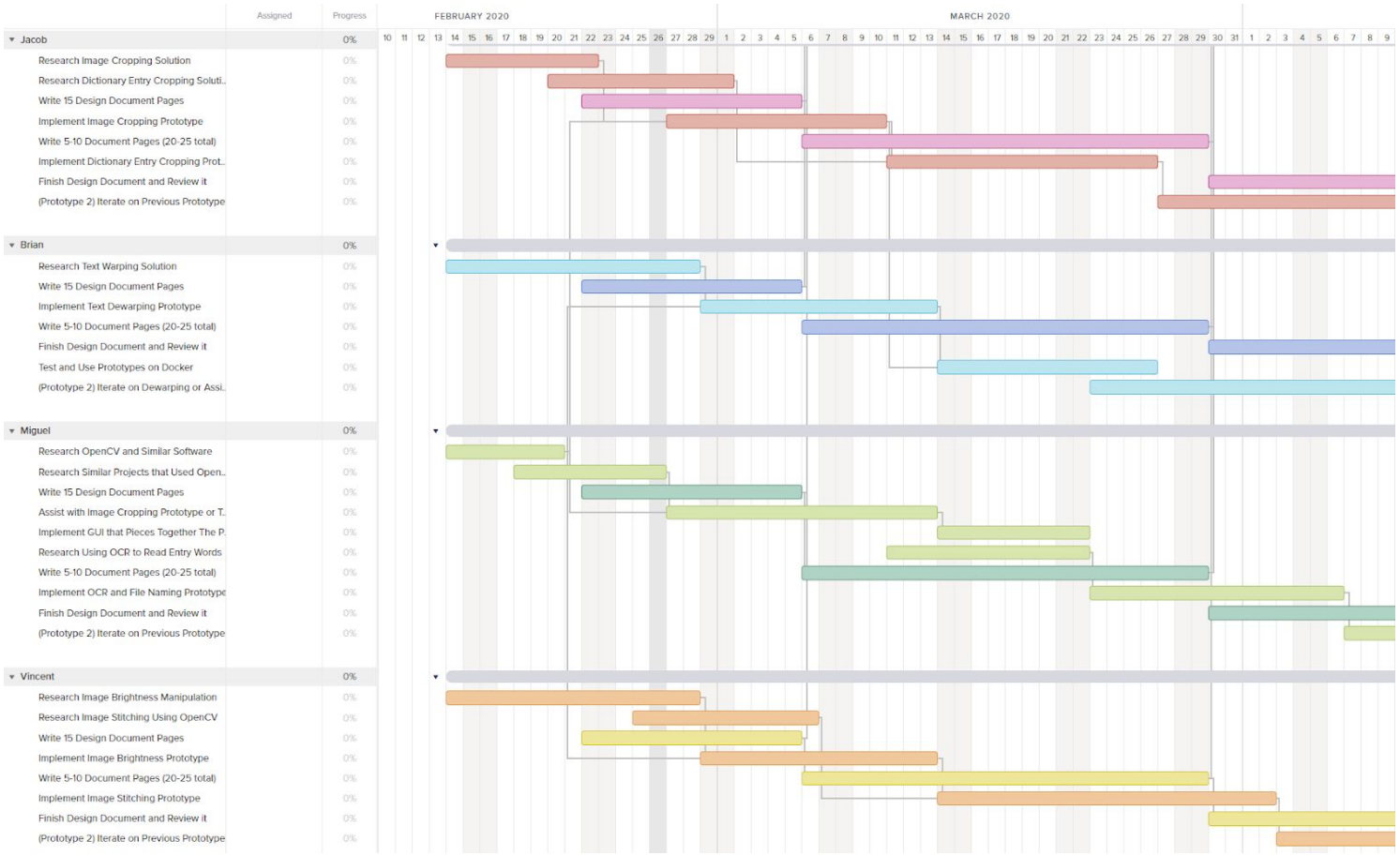


Figure 3: Gantt Chart

The Gantt chart serves three main functions:

1. To divide the labor and different sections of the project evenly to each of our group members. These sections were all agreed on by each member of the group and the difficulty assessment of each section was agreed on as well. If a section turns out to be much more difficult than previously thought, it will be discussed at our weekly meetings and anyone with an easier section will assist with the more difficult one.

2. To keep the important dates of the class somewhere easily accessible to the entire group. The important dates are both the due dates for any assignments and the dates for any group meetings. This is done by having

a link to the Gantt Chart in the Discord at all times. The link is in a separate text channel for important links specifically so that it remains prevalent and visible.

3. To tell each member of the group what they should be working on at any given time, and whether they should be working on the project itself or writing pages for the design document.

Each member of the team has also been assigned a color and segment of the project that should be working on, the division is assigned as so:

- Jacob - Red - Column cropping and dictionary entry cropping

- Brian - Blue - Text dewarping and page rotation

- Miguel - Green - Research for useful software, research OpenCV and how it can be implemented in our project, research using OCR software to read the headwords and name the output files, and implement a GUI that pieces together the other group members solutions.

- Vincent - Yellow - Research and implement the image stitching solution, and subsequently the image brightness solution on the stitched images.

The Gantt Chart shown above in Figure 3 is cut off at the end a bit, but the final lines for every member of the group are finishing the design document and reviewing it, and iterating on the previous prototypes that each member has created for their section of the project. This chart is absolutely vital for the success of this project as it provides a great amount of guidelines to each and every member of the group. Swapping between working on the programming and implementing a basic model of the project and writing the design document is the most optimal way to make it through Senior Design. If any one member of the team is stuck on what to write on the document, then they should simply switch over to continue working on the project sooner and write about their findings from there. The exact date for each and every task on the chart is absolutely not set in stone, we are expecting many of the tasks to take longer or shorter than we have anticipated. Our group is always open to communication and assisting each other with each segment of the project every step of the way. This has been shown

from very early on with our weekly meetings, where we take a look at the progress that each member of the group has made up until that point and we assist each other with any issues that we have with the code. This is shown even further by the Gantt Chart having specific sets of time where if a member of the group is finished with their section and caught up on the design document, they will reach out to the other members of the group for insight on who is struggling the most with their sections and provide them assistance. The dates on the gantt chart only go up until the end of Senior Design 1, we do not have Senior Design 2 planned out yet, as we are for the most part unaware of any important milestones within the class, and as for our milestones, it is simply too difficult to predict what exactly we want to be doing on the project that far ahead in the future. For the most part, we will continue iterating on our projects from the end of Senior Design 1 throughout Senior Design 2. The continuous iteration and learning cycle of the project is how we hope to achieve our goal of having a full and successful project in the end.

To summarize the Gantt chart, each member of the group is given their own individual tasks and each member must write about those tasks in the design document. We alternate and overlap between working on the project itself and working on the design document, so that we have more to write about each time we go back to the writing. If a member is unable to continue writing in the design document, then they should go back to working on their piece of the project to gather more ideas. We are also not restricted to our individual sections, and at our weekly meetings every member of the group helps one another with their piece of the project. All group members have access to and regularly check up on the gantt chart.

4. Technology Used

This section provides an overview of the various technologies used throughout our project, how we used each component, our reasons for choosing each component, and any difficulties we had during the process.

4.1 Methods of Communication

In order to keep the project running as smoothly as possible, the very first thing we did was set up a weekly meeting of undetermined length to work on the project together. This time was used very well, as during every meeting we discussed what we had found throughout the previous week, showed examples to the other group members, and helped each other on our respective sections of the project. We made huge strides during the group meetings, as communication and assistance from someone else in person is priceless for these types of projects. After we were unable to meet in person with COVID-19 going around, our meetings switched to being on discord. We were able to complete a large chunk of the project at these meetings alone. Our main form of communication outside of meeting in person was Discord. Discord is an invaluable tool and we use it on a daily basis. We have both a general chat for discussion of the project or upcoming due dates. We confirmed and set meeting dates using the discord, and the ability to react to messages is a good way of knowing that everyone has seen and agreed with any important reminders or announcements within the group. As the class continued communication dwindled some at some points throughout the project, but it always resumed near the important due dates that kept the group on our toes. We also had a chat that was reserved for important links. This contains links to the powerpoints that we have made together as a group on google slides, and the design document on google docs. In order to communicate with the sponsor of our group, course professor, and course teaching assistants, we used email. We plan on meeting with our sponsor every month to update and inform her of our progress and any limitations that we may have found. Communication was a strong suit of our group, and Discord is the number one reason that we were able to keep up with each other and the fellow group members progress throughout the project and the semester as a whole.

4.2 Software and Other Tools

1. OpenCV: This is the most vital piece of our project, OpenCV (Open Source Computer Vision Library) had nearly everything we needed to make this project a success. We used their algorithms and library in every piece of our project. For example, as you have read in the previous sections, the thresholding that we used to make the pages of the dictionary black and white for easier computer recognition, is an algorithm from OpenCV. Otsu's Binarization, Canny Edge Detection, our functions for implementing page dewarping and fixing page rotation, all were derived at a base level from OpenCV. The documentation on the website of OpenCV was a huge help with the early portions of our project, as each and every algorithm we used from there had a detailed definition and walkthrough for both how to use it and how exactly it works. At the end they even have exercises to perform to implement the algorithm that the specific article is on and increase your skill with using both OpenCV and Python itself. A large amount, if not the entire project will be based on OpenCV libraries along with Python.

2. Discord: The main method of communication for our group, Discord allowed us to speak to each other every day and remind each other of important events in the class. You can create both voice and text channels in this program, so this allows us to type to each other most of the time, but get in a call and speak to each other when it is necessary. We have created two separate text channels in order to better communicate with each other online. The first text channel is the general chat, this is for basic communication about anything and everything relating to this project. We use this channel to communicate to each other a majority of the time. However, anything important that we send in this text channel can get lost in the chat log very easily, as we are constantly sending messages and sometimes even images, which fills up a lot of space. Therefore, we decided to create another text channel for important links. This channel, aptly named important-links, is used to store and give easy access to any important links that the members of our group may need at any given moment. This includes links to the shared Google Docs document where we have our Final Design Document, all the way to links that seem helpful for the solution of the project. This way, the text channel

does not get filled up with any messages and we have easy access to all of these links.

3. Google Docs and Google Slides: Our group utilized both Google Docs and Google Slides to work on the design document and powerpoints remotely. Our group did not have a lot of overlapping free time, so we used Discord and Google Docs whenever we could to work together from home. This proved to be vital, as it allows us to put together the document and edit it all at the exact same time. We can see the edits that the others make in real time, and this allows us to make the document consistent and read well throughout the entire piece. This also enables us to be sure that the other members of the group are writing the different sections and we do not have too much overlap. Overlapping is a problem because we want the document to appear to have been written by a single entity, not an entire group of people. We do not want to repeat too much in the document, except where it is entirely necessary to. Information must sometimes be reiterated in a section, but too much repetition makes for a low quality design document. Upon piecing together our different sections of the design document, Google Docs was absolutely vital. We communicated through Discord and made the document flow from each of our sections.

4. Python: Python is a high level programming language that we are using to write the code for our program. This is due to the fact that OpenCV contains a library full of Python algorithms used for computer vision. It takes advantage of Numpy, a Python library used for mathematical operations and calculations to optimize and allow the usage of many of the OpenCV functions.

5. Docker: Our sponsor requested that we provide her with our project through a docker file, which is a piece of software that contains the code and all of its dependencies to ensure that any piece of software will work from one computer to the next. Essentially, we do not want the development environment to differ from the runtime environment, and have that cause the program to fail to execute. This allows the sponsor and anyone who the sponsor wants to run our program to be able to run it correctly without having to troubleshoot anything. If our program does not

run correctly on the computer of our sponsor then the project can not be considered successful. Even further down the line as well, our project will still be working and in top condition due to Docker providing the same runtime environment as when we created it.

6. Zoom: Due to COVID-19, we have begun to use Zoom to keep in contact with the Teaching Assistants for Senior Design 1. Zoom is a program where you can meet up and communicate as a group online. Many professors at the University of Central Florida have moved to using Zoom to provide lectures and have meetings with their students.

4.3 Language and Libraries

The main implementation language that we will use during this project will be Python. We all decided to use Python because it requires less overhead to run and could allow us to implement the necessary algorithms to make this program work as intended. Another reason we decided to use Python was because of its ease of use and extensive libraries at our disposal. A popular library that we will be using is called OpenCV. OpenCV is an open source computer vision and machine learning software library. Another technology that we will likely use during the implementation phase of this program will be Tesseract. Tesseract is another open source program that we are considering using for this project because it also has a lot of algorithms that have been refined over time and will suit our implementation needs.

OpenCV Modules that will be utilized either directly or indirectly:

1. core. **Core functionality**
2. imgproc. **Image Processing**
3. imgcodecs. **Image file reading and writing**
4. calib3d. **Camera Calibration and 3D Reconstruction**
5. features2d. **2D Features Framework**
6. dnn. **Deep Neural Network module**
7. ml. **Machine Learning**
8. flann. **Clustering and Search in Multi-Dimensional Spaces**
9. photo. **Computational Photography**
10. cudawarping. **Image Warping**

Python Modules that will be heavily utilized:

1. NumPy – An image is essentially a standard NumPy array containing pixels of data points. Therefore, by using basic NumPy operations, such as slicing, masking and fancy indexing, we can modify the pixel values of an image.

2. SciPy – SciPy is another of Python's core scientific modules like NumPy and can be used for basic image manipulation and processing tasks. The package currently includes functions for linear and non-linear filtering, binary morphology, B-spline interpolation, and object measurements.

3. PIL – Python Imaging Library is a free library for the Python programming language that adds support for opening, manipulating, and saving many different image file formats.

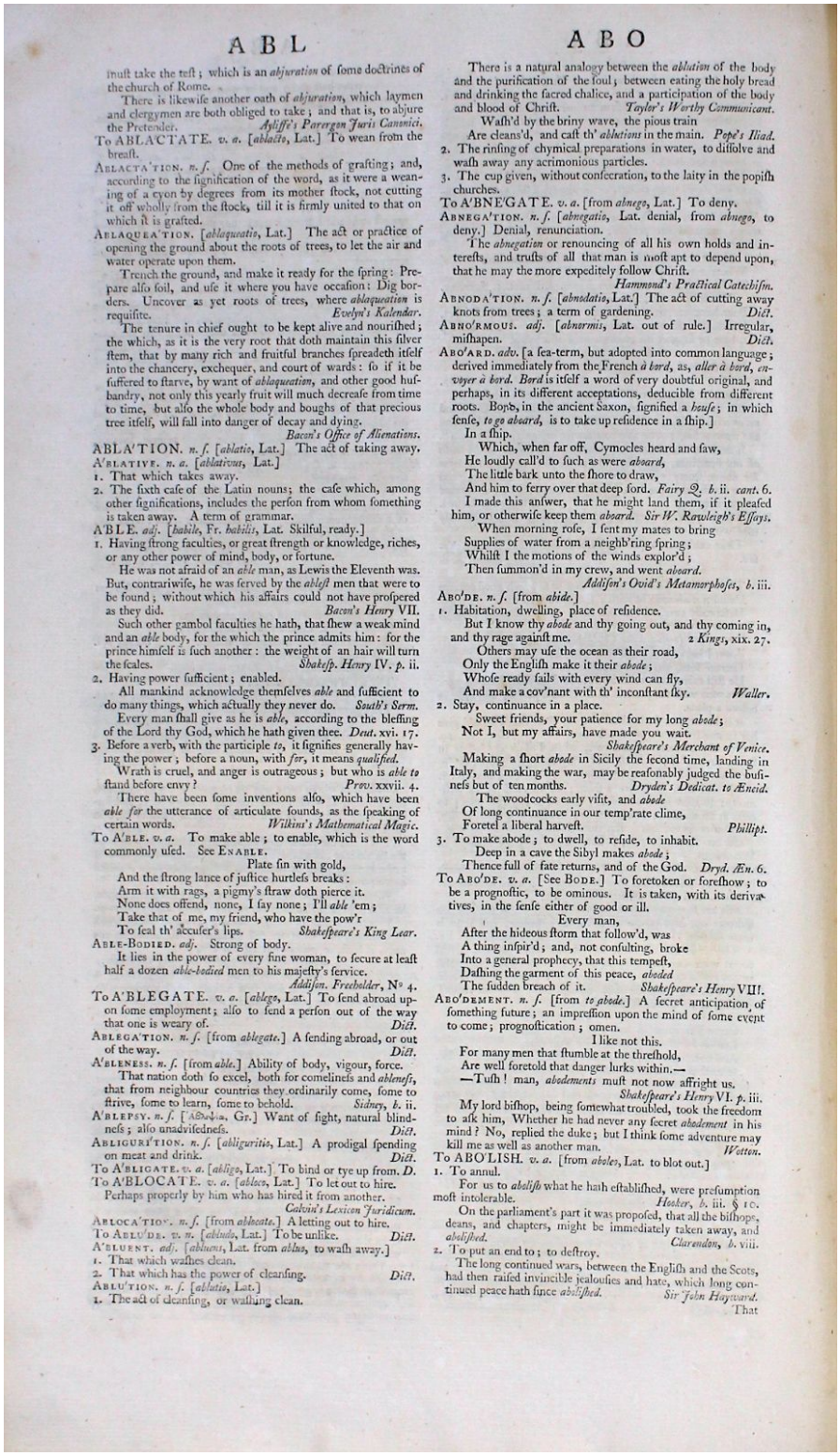
4. Matplotlib – Matplotlib is an effective library for altering images and extracting information from images

5. Research

This section outlines our research for each component of the total algorithm and explanations of the background knowledge needed to understand how each component works. In addition any modifications or initial results from our research is also included in this section.

5.1 Average Page

The following figure is what the average full page of the dictionary will look like. This is important to show as the transformations made to the pages will be shown throughout the document. There are special pages, such as when a new letter begins, or at the very start of the dictionary, but the figure shown above is a much more common page and is mostly what our algorithm will be working on. Compare the examples found throughout the document to this page for a greater understanding of what exactly is being done throughout the algorithm's process. The full page is much larger than any of the other examples show, so it is important to recognize the size of the page and the whitespace surrounding the columns for our explanations later. The figure also serves as a reference for the size of the text, the layout of the columns and how they wrap around a page, and any additional text above or below the columns. Refer back to this page for any comparison to subsequent sections containing a large amount of modification to a given page.



ABL

must take the teſt; which is an *abjuration* of ſome doctrines of the church of Rome.

There is likewiſe another oath of *abjuration*, which laymen and clergymen are both obliged to take; and that is, to abjure the Pretender. *Ayliffe's Parergon Juris Canonici.*

To ABLACTATE. *v. a.* [*ablacto*, Lat.] To wean from the breaſt.

ABLACTA'TION. *n. ſ.* One of the methods of graſting; and, according to the ſignification of the word, as it were a weaning of a cyon by degrees from its mother ſtock, not cutting it off wholly from the ſtock, till it is firmly united to that on which it is graſted.

ABLAQUEA'TION. [*ablaqueatio*, Lat.] The act or practice of opening the ground about the roots of trees, to let the air and water operate upon them.

Trench the ground, and make it ready for the ſpring: Prepare alſo ſoil, and uſe it where you have occaſion: Dig borders. Uncover as yet roots of trees, where *ablaqueation* is requiſite. *Evelyn's Kalendar.*

The tenure in chief ought to be kept alive and nouriſhed; the which, as it is the very root that doth maintain this ſilver ſtem, that by many rich and fruitful branches ſpreadeth itſelf into the chancery, exchequer, and court of wards: ſo if it be ſuffered to ſtarve, by want of *ablaqueation*, and other good huſbandry, not only this yearly fruit will much decreaſe from time to time, but alſo the whole body and boughs of that precious tree itſelf, will fall into danger of decay and dying.

Bacon's Office of Alienations.

ABLA'TION. *n. ſ.* [*ablatio*, Lat.] The act of taking away.

A'BLATIVE. *n. a.* [*ablativus*, Lat.]

1. That which takes away.
2. The ſixth caſe of the Latin nouns; the caſe which, among other ſignifications, includes the perſon from whom ſomething is taken away. A term of grammar.

A'BLE. *adj.* [*habile*, Fr. *habilis*, Lat. Skilful, ready.]

1. Having ſtrong faculties, or great ſtrength or knowledge, riches, or any other power of mind, body, or fortune.

He was not afraid of an *able* man, as Lewis the Eleventh was. But, contrariwiſe, he was ſerved by the *ableſt* men that were to be found; without which his affairs could not have proſpered as they did. *Bacon's Henry* VII.

Such other gambol faculties he hath, that ſhew a weak mind and an *able* body, for the which the prince admits him: for the prince himſelf is ſuch another: the weight of an hair will turn the ſcales. *Shakeſp. Henry* IV. *p.* ii.

2. Having power ſufficient; enabled.

All mankind acknowledge themſelves *able* and ſufficient to do many things, which actually they never do. *South's Serm.*

Every man ſhall give as he is *able*, according to the bleſſing of the Lord thy God, which he hath given thee. *Deut.* xvi. 17.

3. Before a verb, with the participle *to*, it ſignifies generally having the power; before a noun, with *for*, it means *qualified*.

Wrath is cruel, and anger is outrageous; but who is *able to* ſtand before envy? *Prov.* xxvii. 4.

There have been ſome inventions alſo, which have been *able for* the utterance of articulate ſounds, as the ſpeaking of certain words. *Wilkins's Mathematical Magic.*

To A'BLE. *v. a.* To make able; to enable, which is the word commonly uſed. See ENABLE.

Plate ſin with gold,
And the ſtrong lance of juſtice hurtleſs breaks:
Arm it with rags, a pigmy's ſtraw doth pierce it.
None does offend, none, I ſay none; I'll *able* 'em;
Take that of me, my friend, who have the pow'r
To ſeal th' accuſer's lips. *Shakeſpeare's King Lear.*

ABLE-BODIED. *adj.* Strong of body.

It lies in the power of every fine woman, to ſecure at leaſt half a dozen *able-bodied* men to his majeſty's ſervice.

Addiſon. Freeholder, N° 4.

To A'BLEGATE. *v. a.* [*ablego*, Lat.] To ſend abroad upon ſome employment; alſo to ſend a perſon out of the way that one is weary of. *Dict.*

ABLEGA'TION. *n. ſ.* [from *ablegate.*] A ſending abroad, or out of the way. *Dict.*

A'BLENESS. *n. ſ.* [from *able.*] Ability of body, vigour, force.

That nation doth ſo excel, both for comelineſs and *ableneſs*, that from neighbour countries they ordinarily come, ſome to ſtrive, ſome to learn, ſome to behold. *Sidney, b.* ii.

A'BLEPSY. *n. ſ.* [Ἀβλεψία, Gr.] Want of ſight, natural blindneſs; alſo unadvisedneſs. *Dict.*

ABLIGURI'TION. *n. ſ.* [*abliguritio*, Lat.] A prodigal ſpending on meat and drink. *Dict.*

To A'BLIGATE. *v. a.* [*abligo*, Lat.] To bind or tye up from. *D.*

To A'BLOCATE. *v. a.* [*ablocо*, Lat.] To let out to hire.

Perhaps properly by him who has hired it from another.

Calvin's Lexicon Juridicum.

ABLOCA'TION. *n. ſ.* [from *ablocate.*] A letting out to hire.

To ABLU'DE. *v. n.* [*abludo*, Lat.] To be unlike. *Dict.*

A'BLUENT. *adj.* [*abluens*, Lat. from *abluo*, to waſh away.]

1. That which waſhes clean.
2. That which has the power of cleanſing. *Dict.*

ABLU'TION. *n. ſ.* [*ablutio*, Lat.]

1. The act of cleanſing, or waſhing clean.

ABO

There is a natural analogy between the *ablution* of the body and the purification of the ſoul; between eating the holy bread and drinking the ſacred chalice, and a participation of the body and blood of Chriſt. *Taylor's Worthy Communicant.*

Waſh'd by the briny wave, the pious train
Are cleans'd, and caſt th' *ablutions* in the main. *Pope's Iliad.*

2. The rinſing of chymical preparations in water, to diſſolve and waſh away any acrimonious particles.
3. The cup given, without conſecration, to the laity in the popiſh churches.

To A'BNEGATE. *v. a.* [from *abnego*, Lat.] To deny.

ABNEGA'TION. *n. ſ.* [*abnegatio*, Lat. denial, from *abnego*, to deny.] Denial, renunciation.

The *abnegation* or renouncing of all his own holds and intereſts, and truſts of all that man is moſt apt to depend upon, that he may the more expeditely follow Chriſt.

Hammond's Practical Catechiſm.

ABNODA'TION. *n. ſ.* [*abnodatio*, Lat.] The act of cutting away knots from trees; a term of gardening. *Dict.*

ABNO'RMOUS. *adj.* [*abnormis*, Lat. out of rule.] Irregular, miſhapen. *Dict.*

ABO'ARD. *adv.* [a ſea-term, but adopted into common language; derived immediately from the French *à bord*, as, *aller à bord*, *envoyer à bord*. *Bord* is itſelf a word of very doubtful original, and perhaps, in its different acceptations, deducible from different roots. Boꝛd, in the ancient Saxon, ſignified a *houſe*; in which ſenſe, *to go aboard*, is to take up reſidence in a ſhip.]

In a ſhip.

Which, when far off, Cymocles heard and ſaw,
He loudly call'd to ſuch as were *aboard*,
The little bark unto the ſhore to draw,
And him to ferry over that deep ford. *Fairy Q. b.* ii. *cant.* 6.

I made this anſwer, that he might land them, if it pleaſed him, or otherwiſe keep them *aboard*. *Sir W. Rawleigh's Eſſays.*

When morning roſe, I ſent my mates to bring
Supplies of water from a neighb'ring ſpring;
Whilſt I the motions of the winds explor'd;
Then ſummon'd in my crew, and went *aboard*.

Addiſon's Ovid's Metamorphoſes, b. iii.

ABO'DE. *n. ſ.* [from *abide.*]

1. Habitation, dwelling, place of reſidence.

But I know thy *abode* and thy going out, and thy coming in, and thy rage againſt me. 2 *Kings*, xix. 27.

Others may uſe the ocean as their road,
Only the Engliſh make it their *abode*;
Whoſe ready ſails with every wind can fly,
And make a cov'nant with th' inconſtant ſky. *Waller.*

2. Stay, continuance in a place.

Sweet friends, your patience for my long *abode*;
Not I, but my affairs, have made you wait.

Shakeſpeare's Merchant of Venice.

Making a ſhort *abode* in Sicily the ſecond time, landing in Italy, and making the war, may be reaſonably judged the buſineſs but of ten months. *Dryden's Dedicat. to Æneid.*

The woodcocks early viſit, and *abode*
Of long continuance in our temp'rate clime,
Foretel a liberal harveſt. *Phillips.*

3. To make abode; to dwell, to reſide, to inhabit.

Deep in a cave the Sibyl makes *abode*;
Thence full of fate returns, and of the God. *Dryd. Æn.* 6.

To ABO'DE. *v. a.* [See BODE.] To foretoken or foreſhow; to be a prognoſtic, to be ominous. It is taken, with its derivatives, in the ſenſe either of good or ill.

Every man,
After the hideous ſtorm that follow'd, was
A thing inſpir'd; and, not conſulting, broke
Into a general prophecy, that this tempeſt,
Daſhing the garment of this peace, *aboded*
The ſudden breach of it. *Shakeſpeare's Henry* VIII.

ABO'DEMENT. *n. ſ.* [from *to abode.*] A ſecret anticipation of ſomething future; an impreſſion upon the mind of ſome event to come; prognoſtication; omen.

I like not this.
For many men that ſtumble at the threſhold,
Are well foretold that danger lurks within.—
—Tuſh! man, *abodements* muſt not now affright us.

Shakeſpeare's Henry VI. *p.* iii.

My lord biſhop, being ſomewhat troubled, took the freedom to aſk him, Whether he had never any ſecret *abodement* in his mind? No, replied the duke; but I think ſome adventure may kill me as well as another man. *Wotton.*

To ABO'LISH. *v. a.* [from *abolo*, Lat. to blot out.]

1. To annul.

For us to *abeliſh* what he hath eſtabliſhed, were preſumption moſt intolerable. *Hooker, b.* iii. § 10.

On the parliament's part it was propoſed, that all the biſhops, deans, and chapters, might be immediately taken away, and *abeliſhed*. *Clarendon, b.* viii.

2. To put an end to; to deſtroy.

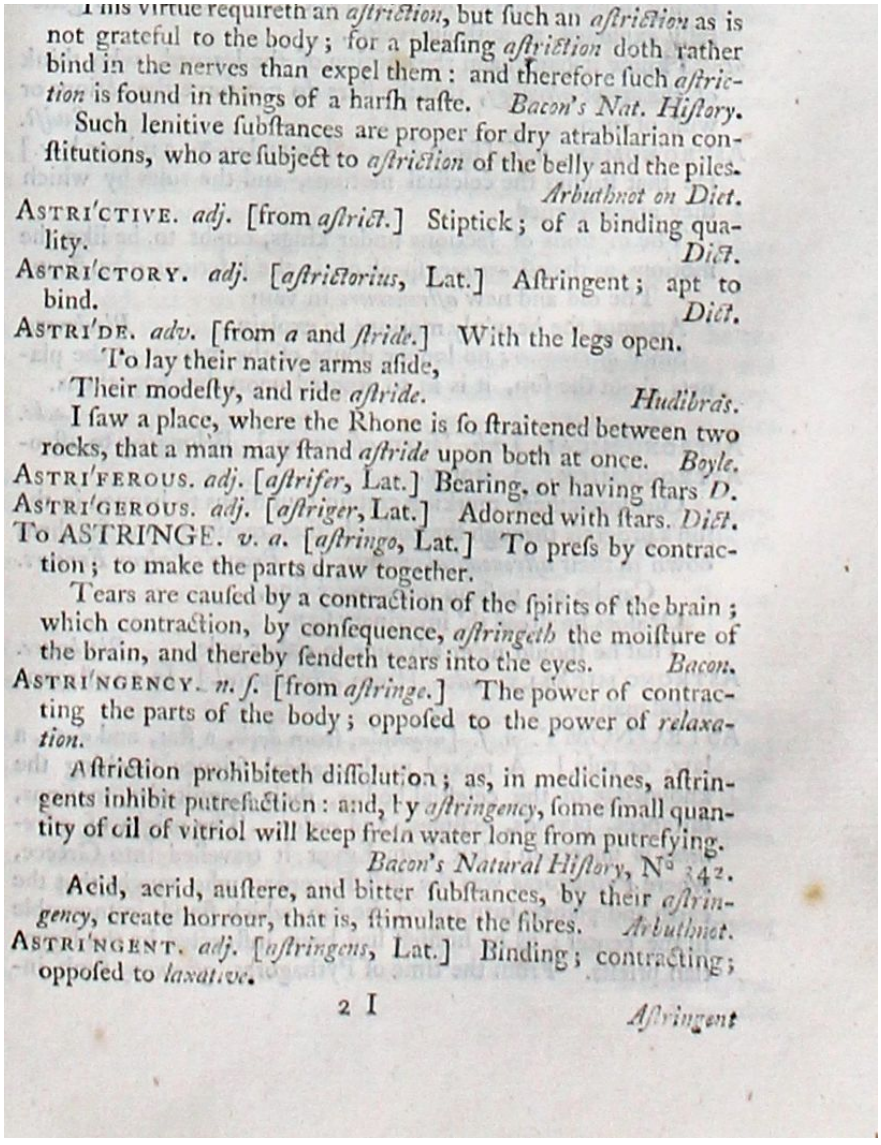
The long continued wars, between the Engliſh and the Scots, had then raiſed invincible jealouſies and hate, which long continued peace hath ſince *abeliſhed*. *Sir John Hayward.*

That

Figure 4: Average Full Page

5.2 Text Dewarping

The provided scans of the dictionary were obtained from someone placing the dictionary onto a scanner. This obviously led to imperfections in the scan such as page skewing, but the most obvious example is in the warping of text on the page as pictured below. As you can see, the page is crumpled as a result of being put on the scanner. Unlike the other parts of our solution, there is no obvious fix to this issue.



This virtue requireth an *aſtriction*, but ſuch an *aſtriction* as is not grateful to the body; for a pleaſing *aſtriction* doth rather bind in the nerves than expel them: and therefore ſuch *aſtriction* is found in things of a harſh taſte. *Bacon's Nat. Hiſtory.*

Such lenitive ſubſtances are proper for dry atrabilarian conſtitutions, who are ſubject to *aſtriction* of the belly and the piles. *Arbuthnot on Diet.*

ASTRI'CTIVE. *adj.* [from *aſtrict.*] Stiptick; of a binding quality. *Dict.*

ASTRI'CTORY. *adj.* [*aſtrictorius*, Lat.] Aſtringent; apt to bind. *Dict.*

ASTRI'DE. *adv.* [from *a* and *ſtride.*] With the legs open.

To lay their native arms aſide,
Their modeſty, and ride *aſtride.* *Hudibras.*

I ſaw a place, where the Rhone is ſo ſtraitened between two rocks, that a man may ſtand *aſtride* upon both at once. *Boyle.*

ASTRI'FEROUS. *adj.* [*aſtrifer*, Lat.] Bearing, or having ſtars *D.*

ASTRI'GEROUS. *adj.* [*aſtriger*, Lat.] Adorned with ſtars. *Dict.*

To ASTRINGE. *v. a.* [*aſtringo*, Lat.] To preſs by contraction; to make the parts draw together.

Tears are cauſed by a contraction of the ſpirits of the brain; which contraction, by conſequence, *aſtringeth* the moiſture of the brain, and thereby ſendeth tears into the eyes. *Bacon.*

ASTRI'NGENCY. *n. ſ.* [from *aſtringe.*] The power of contracting the parts of the body; oppoſed to the power of *relaxation.*

Aſtriction prohibiteth diſſolution; as, in medicines, aſtringents inhibit putrefaction: and, by *aſtringency*, ſome ſmall quantity of oil of vitriol will keep freſh water long from putrefying. *Bacon's Natural Hiſtory*, N° 342.

Acid, acrid, auſtere, and bitter ſubſtances, by their *aſtringency*, create horrour, that is, ſtimulate the fibres. *Arbuthnot.*

ASTRI'NGENT. *adj.* [*aſtringens*, Lat.] Binding; contracting; oppoſed to *laxative.*

2 I *Aſtringent*

Figure 5: Warped dictionary entry

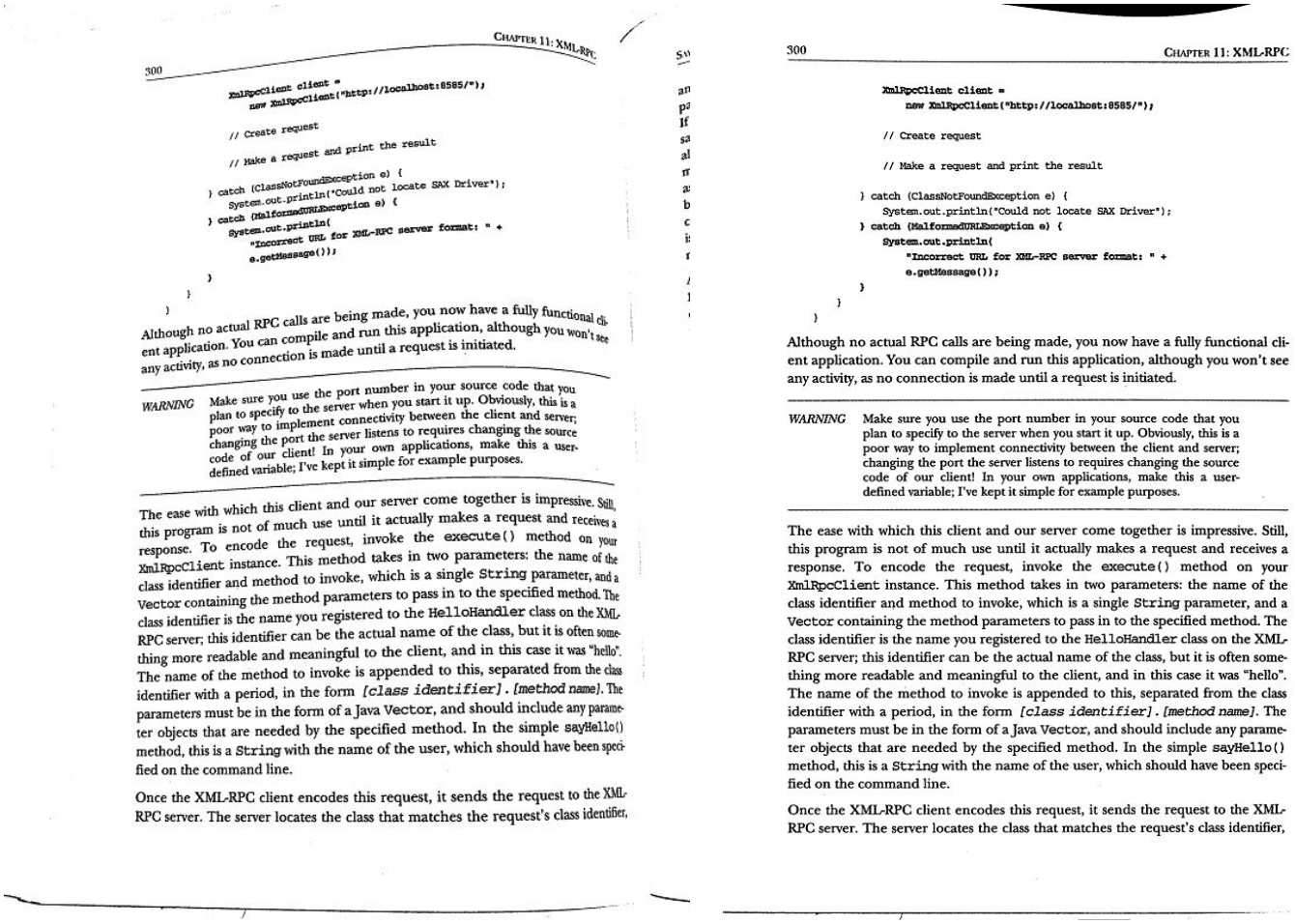
Our first proposed solution to the issue was to break the warped page down not only into each definition, but to go letter by letter and fix the warp at the smallest level we could think of. While this would be the apparent easiest and most brute force method of correcting for the warp there were several issues that arose. The first is obviously the efficiency, if we needed to break every warped page down into each letter and then correct them, the runtime would be incredibly slow. While this image manipulation will not be the most efficient in general, due to the necessary computation, we do not want to make it any worse than it has to be. The second issue that we ran into was when we finally fixed the warping for each individual letter, when pieced back together, the original appearance of the document was no longer intact. From Dr. Giroux's specifications, we need to preserve the appearance of each entry and not just the content within it.

Building on our last proposal, we decided to see how feasible it would be to implement the dewarping on each entry instead of every single letter on the page. After the original image was cropped into two columns and then into each definition, we would have a trigger to detect if the image was warped or not. For skew, it would look at the beginning and ending of the entries and based on the position of the text from the head word to the last word in the entry. If the difference in position for the entries was larger than our threshold, then correct for that accordingly. The trigger for detecting page skew would be simple enough to implement, but we ran into issues when determining how we would detect for the actual warping so we had to reevaluate entirely and do some more research.

At this point we realized that we did not understand the technical aspects of how dewarping worked. We discovered that the extent of our mathematical and computer vision knowledge was insufficient and learning the requisite skills would take many months. Given our circumstances, we would have to look into implementing an already existing open source solution and modify it to our needs.

As this problem can be common when it comes to digitization and document preservation there are already a few programs that offer a solution. The first solution that we found was a paid program that offered the ability to modify and manage a large scale of digitized documents, the LIMB product suite. Since we wanted to keep everything we used open source to ensure portability and to be available to anyone who wants to use our project, we immediately decided against it.

The next possible solution we found was from a few researchers at Stony Brook University who were trying to solve a very similar issue of unwarping document images using machine learning [1]. It was trained using documents that were first scanned to use as a baseline and then the documents were intentionally distorted. After this, their algorithm was applied to the documents to reconstruct the images of the distorted documents. Their results were extremely good as pictured below, so we decided that we wanted to see how it would work for our documents. The biggest issue was that the executable zip file provided on their Github page did not work. There was no readme file on how to run it, just an executable that did nothing and a test batch file. While at first promising, this ended up leading us to a dead end so we had to investigate even further.



300 CHAPTER 11: XML-RPC

```
            XmlRpcClient client =
                new XmlRpcClient("http://localhost:8585/");

            // Create request

            // Make a request and print the result

        } catch (ClassNotFoundException e) {
            System.out.println("Could not locate SAX Driver");
        } catch (MalformedURLException e) {
            System.out.println(
                "Incorrect URL for XML-RPC server format: " +
                e.getMessage());
        }
    }
}
```

Although no actual RPC calls are being made, you now have a fully functional client application. You can compile and run this application, although you won't see any activity, as no connection is made until a request is initiated.

WARNING Make sure you use the port number in your source code that you plan to specify to the server when you start it up. Obviously, this is a poor way to implement connectivity between the client and server; changing the port the server listens to requires changing the source code of our client! In your own applications, make this a user-defined variable; I've kept it simple for example purposes.

The ease with which this client and our server come together is impressive. Still, this program is not of much use until it actually makes a request and receives a response. To encode the request, invoke the `execute()` method on your `XmlRpcClient` instance. This method takes in two parameters: the name of the class identifier and method to invoke, which is a single `String` parameter, and a `Vector` containing the method parameters to pass in to the specified method. The class identifier is the name you registered to the `HelloHandler` class on the XML-RPC server; this identifier can be the actual name of the class, but it is often something more readable and meaningful to the client, and in this case it was "hello". The name of the method to invoke is appended to this, separated from the class identifier with a period, in the form *`[class identifier].[method name]`*. The parameters must be in the form of a Java `Vector`, and should include any parameter objects that are needed by the specified method. In the simple `sayHello()` method, this is a `String` with the name of the user, which should have been specified on the command line.

Once the XML-RPC client encodes this request, it sends the request to the XML-RPC server. The server locates the class that matches the request's class identifier,

300 CHAPTER 11: XML-RPC

```
            XmlRpcClient client =
                new XmlRpcClient("http://localhost:8585/");

            // Create request

            // Make a request and print the result

        } catch (ClassNotFoundException e) {
            System.out.println("Could not locate SAX Driver");
        } catch (MalformedURLException e) {
            System.out.println(
                "Incorrect URL for XML-RPC server format: " +
                e.getMessage());
        }
    }
}
```

Although no actual RPC calls are being made, you now have a fully functional client application. You can compile and run this application, although you won't see any activity, as no connection is made until a request is initiated.

WARNING Make sure you use the port number in your source code that you plan to specify to the server when you start it up. Obviously, this is a poor way to implement connectivity between the client and server; changing the port the server listens to requires changing the source code of our client! In your own applications, make this a user-defined variable; I've kept it simple for example purposes.

The ease with which this client and our server come together is impressive. Still, this program is not of much use until it actually makes a request and receives a response. To encode the request, invoke the `execute()` method on your `XmlRpcClient` instance. This method takes in two parameters: the name of the class identifier and method to invoke, which is a single `String` parameter, and a `Vector` containing the method parameters to pass in to the specified method. The class identifier is the name you registered to the `HelloHandler` class on the XML-RPC server; this identifier can be the actual name of the class, but it is often something more readable and meaningful to the client, and in this case it was "hello". The name of the method to invoke is appended to this, separated from the class identifier with a period, in the form *`[class identifier].[method name]`*. The parameters must be in the form of a Java `Vector`, and should include any parameter objects that are needed by the specified method. In the simple `sayHello()` method, this is a `String` with the name of the user, which should have been specified on the command line.

Once the XML-RPC client encodes this request, it sends the request to the XML-RPC server. The server locates the class that matches the request's class identifier,

Figure 6 [2]: Results from the research described above [1]

The next most promising solution that we had found up to this point was Leoptonica, an open source image processing code suite [3]. In addition to many other useful tools and the fact that it was open source, it offered a way to dewarp document text from camera images. The results shown in their documentation were promising, but the biggest issue with Leptonica is that it is written in C. Our main dependency, OpenCV is no longer available for C, only C++. While there may be a way to hack together different working pieces of our project just to get this specific program to work we did not deem it necessary. Not only this, but we decided on using Python since it lends itself very nicely to writing scripts and automating processes, which is the main crux of our solution. We ultimately landed on our current solution, based off of a project we found by Dr. Matt Zucker at Swarthmore College which will be described in section 5.2.4 [9].

5.2.1 Preliminary Definitions for Dewarping

In the following section there will be a deep dive into how the text dewarping algorithm works. This however requires many definitions, models, and mathematical equations to be explained so this section will cover those necessities before we dive into the workings of the algorithm.

When it comes to dealing with text and text segmentation, a very common approach in computer vision is to apply an adaptive threshold. An adaptive threshold is a mask applied to an image to obtain a binary image. A binary is an image where all of the pixels in the image array are either 0 or 1, indicating whether that pixel should be switched on (white) or switched off (black). There are different ways to achieve a threshold, but we are using a threshold mean value. The choice for whether a pixel should be switched on or off is calculated through the mean of its neighboring pixels.

In the realm of adaptive thresholding comes dilation and erosion. These are two image processing (morphological) operations that modify the image that was run through an adaptive threshold. Both of these operations require use of a structuring element, which is a matrix that helps to define the rule for both dilation and erosion. It is simply a matrix that represents the neighboring pixels of each pixel within the image. The following figure helps show different structuring elements.

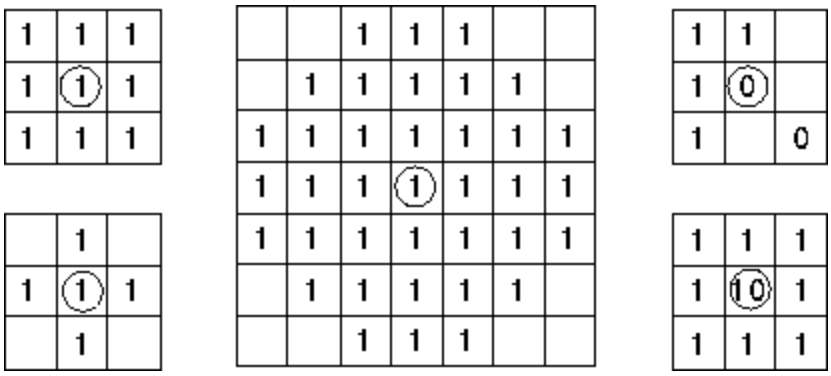


Figure 7 [5]: Multiple sizes of structuring elements

When these structuring elements are moved along a binary image array, the pixel that is circled in Figure 7 is the pixel that is going to be either switched on, in the case of dilation or switched off in erosion.

Dilation is the morphological process of adding pixels to the boundaries of objects within an image. The amount of pixels added is dependent upon the structuring element applied, in our case a matrix of size $9x1$ for text on the page and a matrix of size $3x1$ for the page itself. There are separate rules for dilation and erosion, for dilation a pixel is dilated i.e. switched from 0 to 1 in the image array if any of its neighboring pixels within the structuring element are also 1. As shown in figure 2 below, after the adaptive threshold is applied and then subsequently the dilation is applied, the j becomes much larger.

Erosion is the morphological process of removing pixels from the boundaries of objects within an image. It is the opposite of the dilation process. Once again, the amount of pixels removed is dependent upon the structuring element and in our case we a have a matrix of size $1x3$ for the text and a matrix of size $3x1$ for the page. For erosion the rule for removing pixels is that if any of the neighboring pixels to another pixel within the structuring element are 0, that pixel is then switched to 0, i.e. it is removed from the final image.

Figure 8 [10]: The image is first thresholded to obtain a binary image, dilation is applied to make the object larger

Figure 9 [10]: The image is first thresholded to obtain a binary image, erosion is applied to make the object smaller.

OpenCV's process of dealing with image mapping/remapping and projections is built on a camera model known as the pinhole camera model. This model is based on the titular pinhole camera, which is one of the most basic types of camera consisting only of an aperture and no lens as shown in Figure 10 below [4].

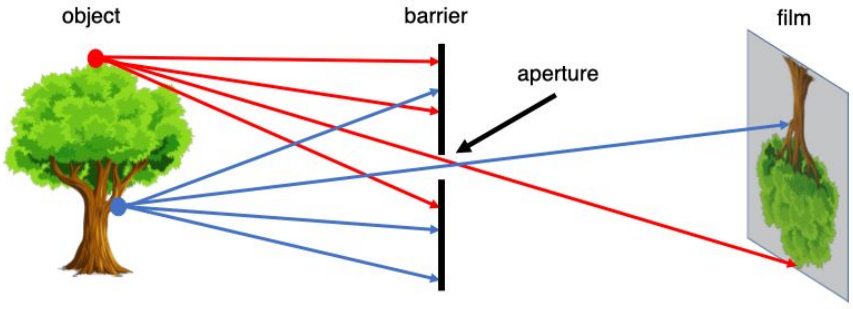


Figure 10 [4]: A pinhole camera model as described above.

In this model, not every point of light projected from the object passes through onto the film, only a few points pass through leading to a nearly one-to-one mapping between the 3D points and the points on the film. While Figure 10 shows a basic overview of this concept, below Figure 11 shows a formal construction of this model.

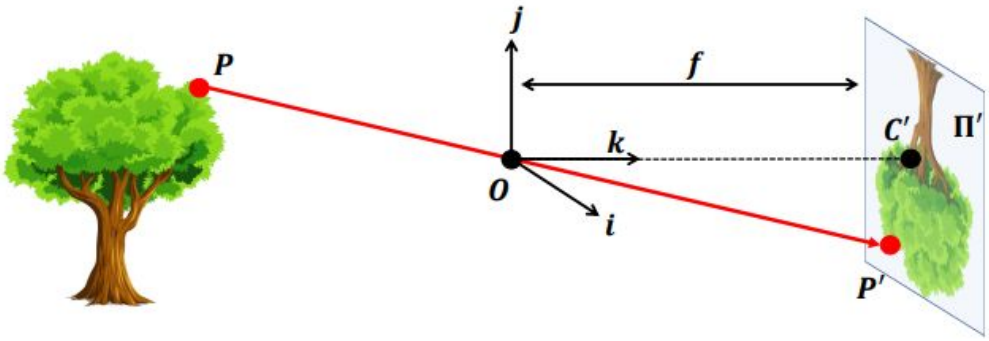


Figure 11 [4]: A pinhole camera model as described above in more detail

There are many working parts to even such a simple model and they are as follows. The image plane, this is what was projected onto the film as shown in the previous figure. The aperture, otherwise known as the pinhole represented by O, which represents the center of our camera. The distance between the image and the aperture is the focal length f. Lastly we have our coordinate vectors of 3D space $\vec{i}$, $\vec{j}$, and $\vec{k}$. The point P on our object can be represented by a $1x3$ matrix, $\begin{bmatrix} x & y & z \end{bmatrix}^T$ which is visible to our camera. This point will be projected onto the image plane Π', giving us a point on our image plane of P'= $\begin{bmatrix} x' & y' \end{bmatrix}^T$. Since our point is projected from a space in $\mathbb{R}^3$ to one in $\mathbb{R}^2$, we lose our z component. Using our point P' we now have a basis for a coordinate system that is centered at our pinhole O, which is known as the camera coordinate system [4].

Used in many image and audio processing applications is dithering. Dithering is noise applied to an image which reduces the color range of an image and helps to eliminate various inaccurate color representations in image display. This process was done in newspapers and comics since in the most basic example, a black and white image can have dithering applied to it to create the illusion that there are gray values present. One of the most common inaccuracies in computer graphics is known as color banding which can be seen in the following image.

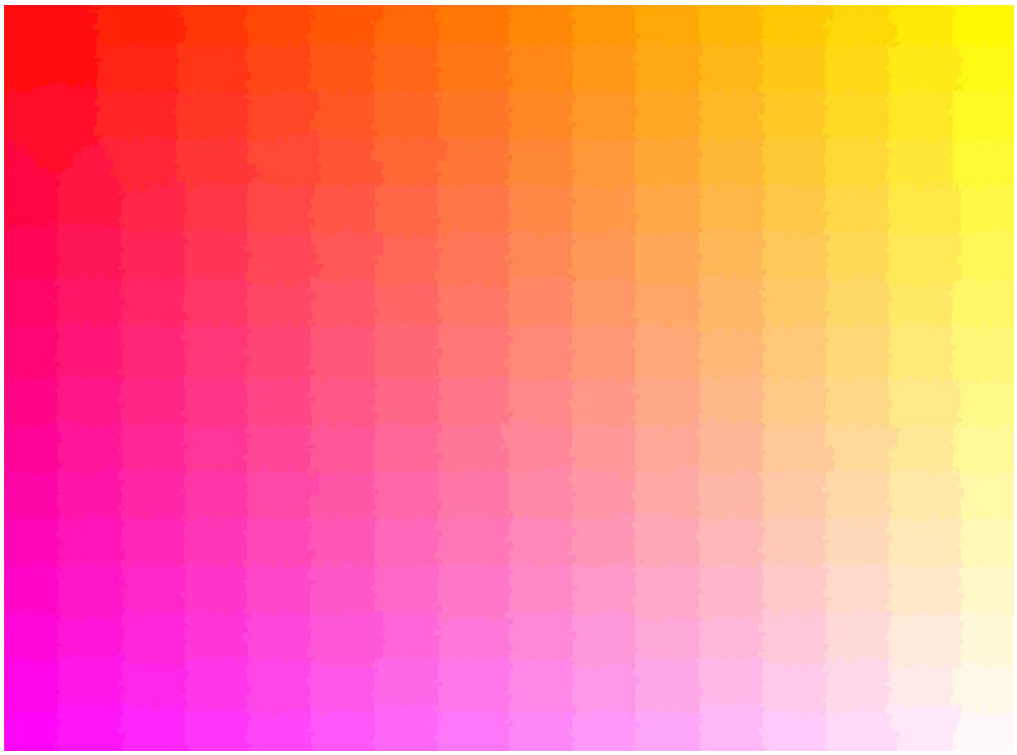


Figure 12 [8]: A gradient with color banding

This gradient has obvious transitions in color, with each transition being a new “band” of color, thus the name color banding. Dithering is then applied to this image to obtain the final smooth gradient pictured below.

Figure 13 [8]: A smooth gradient with dithering applied

Once dithering is applied, the bands of color are no longer present. The whole idea of dithering is to restrict the palette from a full 24-bit RGB color range to a restricted 8-bit range with 256 values, a process known as color quantization. This process was used very frequently in the early days of web design when high speed internet was not as prevalent. Applying dithering to images in a web browser helps reduce the size of the image to help reduce the amount of bandwidth used to load pages.

Similar to the structuring element, there is pixel connectivity. This is a way to classify pixels in an image and how they compare to their neighbors.The two main two-dimensional pixel neighborhoods are 4-connected and 8-connected. In a group of pixels that are 4-connected any pixel that shares an edge is a neighbor. In a group of pixels that is 8-connected any pixel that shares either an edge or a vertex is a neighbor.

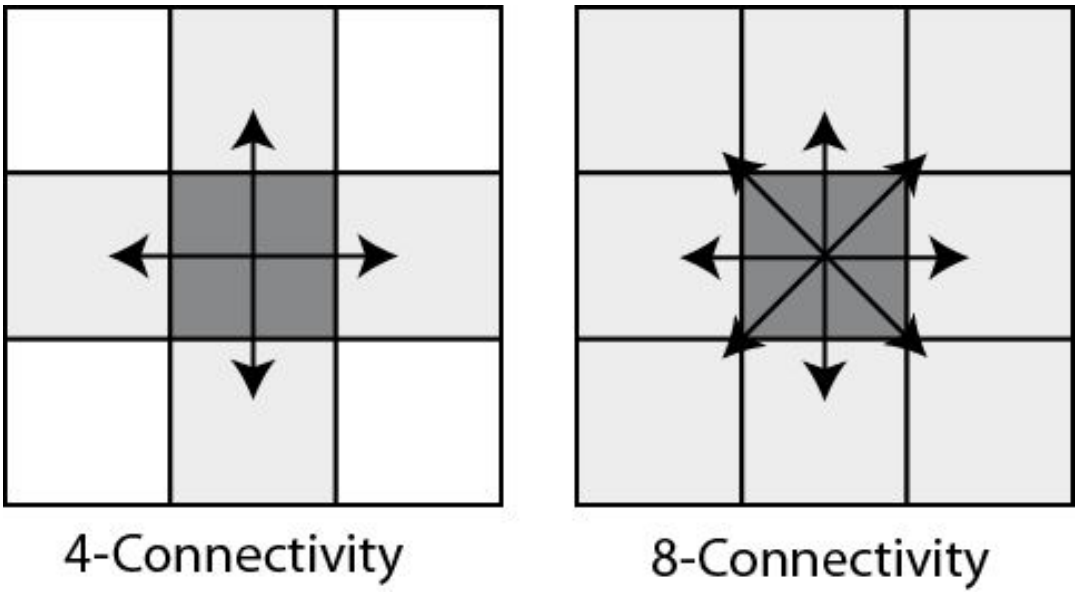


Figure 14 [6]: 4- and 8-connected pixel neighborhoods

Connected-component labelling is an application of graph theory into the world of image processing and computer vision. It is used to group components of an image that share similar qualities and label them to help differentiate them from other distinct groups [5]. There are two different implementations of connected-component labelling, but the method that is the easiest to implement and is most efficient is known as the two-pass algorithm. The steps to the algorithm are as follows:

First Pass:

```
image = binary_image
connected = []
label = 1
     for row in image
       for col in image
         # pixel is not background, we want to look at it
         if image(row, col) == 1 (white)
           neighbors[] = 8-connected pixels

             # new component of the image
             if all neighbors == 0 (black)
               # give new label, increment, and find set
               # containing the current label
               connected[label] = set containing label
               image(row, col) = label
               label += 1

            # pixel has one or more neighbors
            else
              # find indices of neighbors and update
              indices[] = indices_of_neighbors
              image(row, col) = min(indices_of_neighbors)
              for label in image
                combine sets that are neighbors but do not
                have the same label
```

Second Pass:

```
     for row in image
       for col in image
         if image(row, col) is not background
           image(row, col) = find parent label of the set
```

The first pass of the algorithm is very straight-forward, we are simply iterating through the image array and labelling connected components. In the case where there is a pixel with one or more neighbors, we get the minimum of its neighbors in this first pass just to give it a preliminary label. Then we combine the set of each label with its neighboring pixels that do not have the same label, assigning the parent node to that label. Set 1 is the parent of set 2 and Set 3 is the parent of sets 4, 5, 6, and 7 in Figure 15. In the second pass of the algorithm is when the labels are updated to correspond to their actual values.

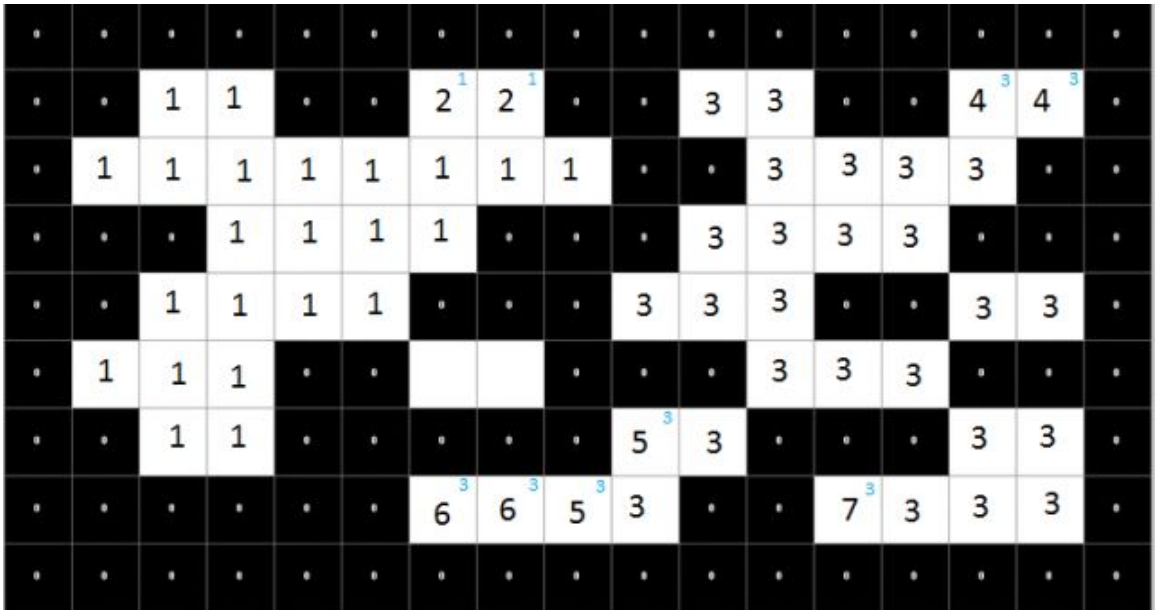


Figure 15 [1]: First pass of Connected-Component Labelling algorithm

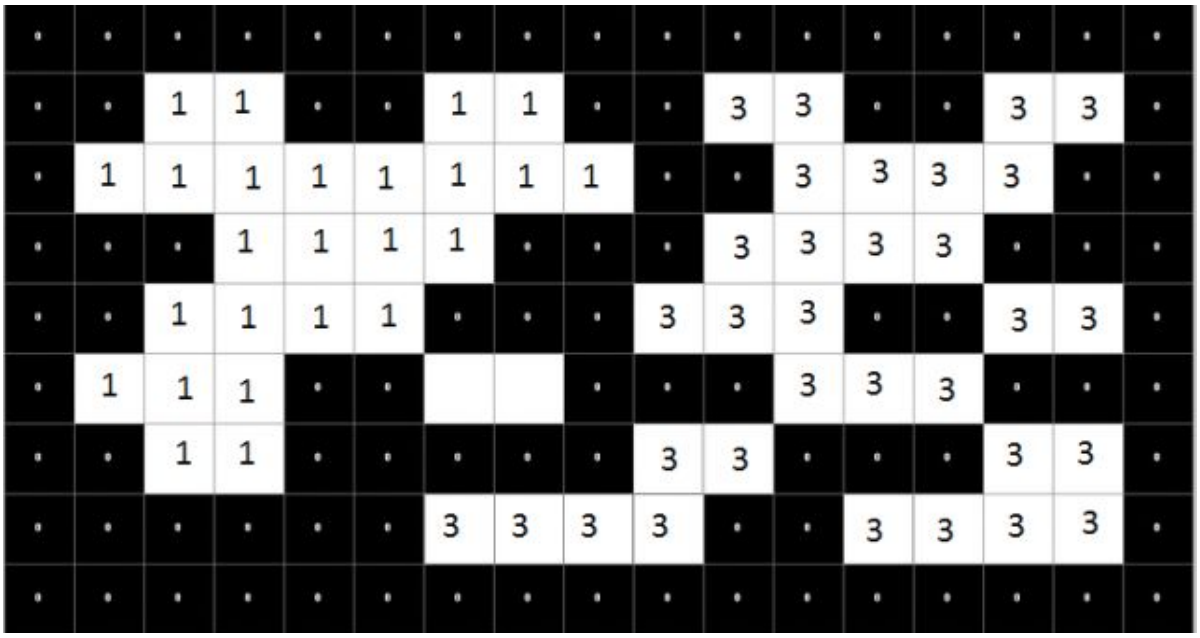


Figure 16 [1]: Second pass of Connected-Component Labelling algorithm

5.2.2 Dewarping Subprocess Flowchart

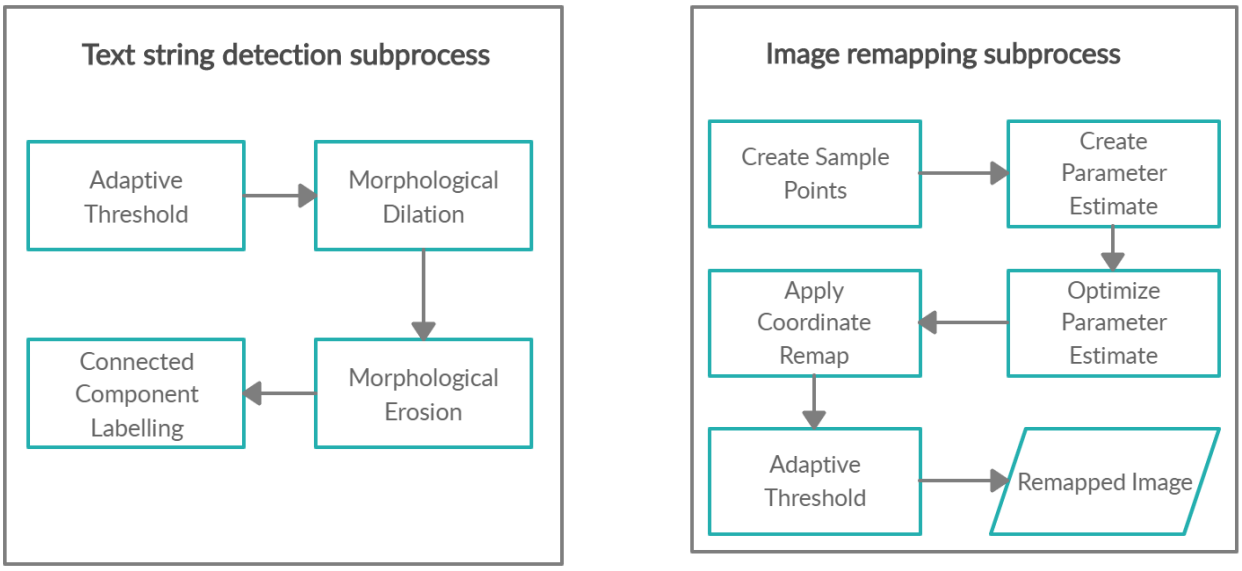


Figure 17: Image remapping and string detection subprocesses in further detail

These two flowcharts break down the subprocesses that are present in our dewarping algorithm. As described earlier, before completing the image remapping the strings of text on each page must first be detected and connected into similar groups. After this process is complete, the final remapping can be calculated and our image is dewarped.

5.2.3 Dewarping Algorithm Flowchart

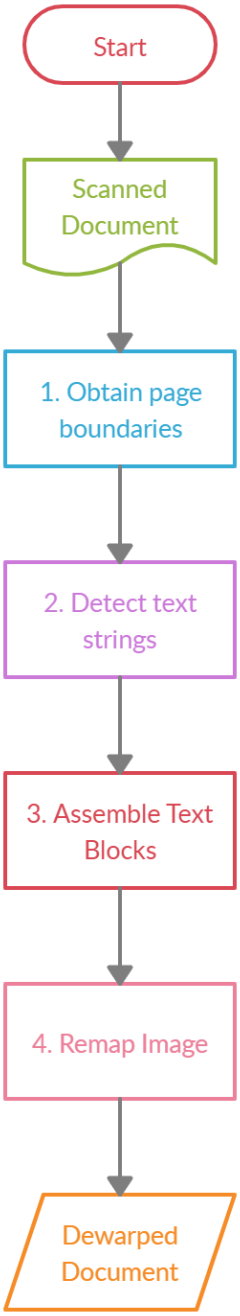


Figure 18: Dewarping algorithm flowchart

This flowchart describes the basic approach for the algorithm for dewarping text. We start with our warped image, detect the page boundaries, then detect its text strings, assemble those strings into blocks of text as they are represented on the page, remap the warped image, and save the final dewarped image.

5.2.4 Text Dewarping Algorithm

The problem is broken down into two main steps, first the text must be split into lines and then a warp/coordinate transformation must be found to ensure that the text lines are parallel and horizontal. The procedure of this algorithm are as follows:

1. Obtain page boundaries: This is done by specifying a fixed margin of your initial image that you consider your page. For example, one of our images is 713 x 2801, we can specify our margins to be 20 pixels in each direction to only obtain the text and ignore the page borders.

2. Detect strings of text: This is done in multiple steps, and it is a similar approach to how we are delineating each entry in the dictionary. First, an adaptive threshold is applied to the image. Morphological dilation is then applied to this binary image, which combines each string of text into a single contour, where a contour is a curve that joins the continuous points of the same color, in this case our text. From here, erosion is applied to the dilated binary image to make the boundaries of the text smaller and obtain a better fit. While it may seem unnecessary to both dilate and erode the image, this process is necessary to ensure that we are actually obtaining text and not any extraneous data on the page. After these two processes are applied, connected component labelling is applied to further eliminate any extraneous items on the page that may not be text.

3. Assemble text blocks: In our previous step, we broke up each section of text into its own self-contained box. Any text on the same line that is not within the same contour, that is not part of one word, is not counted as being on the same line so we need to connect them. This is done by checking edge costs between each contour and if the cost is less than infinity, i.e., if they are on the same line, we append the two contours. The

current solution for this is polynomial runtime of $O(n^2)$, but if possible we would like to reduce that to be more efficient.

4. Since our original image is taken from a lense, there are many different factors influencing the final result of the scanned image, which is the crux of this entire problem. Remapping must be done in multiple steps:

    a. Create sample points: The algorithm for detecting the curvature of the page is built using a parametric model, defined by: rotation vector $\overrightarrow{r}$, translation vector $\overrightarrow{t}$ which parameterize the 3D orientation and position of our page, slopes **α** and ***β*** which are the curvature of the page, vertical offsets y_n of the text spans on our page, and horizontal offsets x_n which represent points within each text span. This model is based on discrete points, so there needs to be fixed points in each text contour, our x_n.

    b. Create parameter estimate: Using OpenCV's `solvePnP` [11] method, which estimates the initial pose of an object in this case our page, its image projections, and the camera matrix. After this, OpenCV's `projectPoints` [12] method projects the new image onto the image plane.

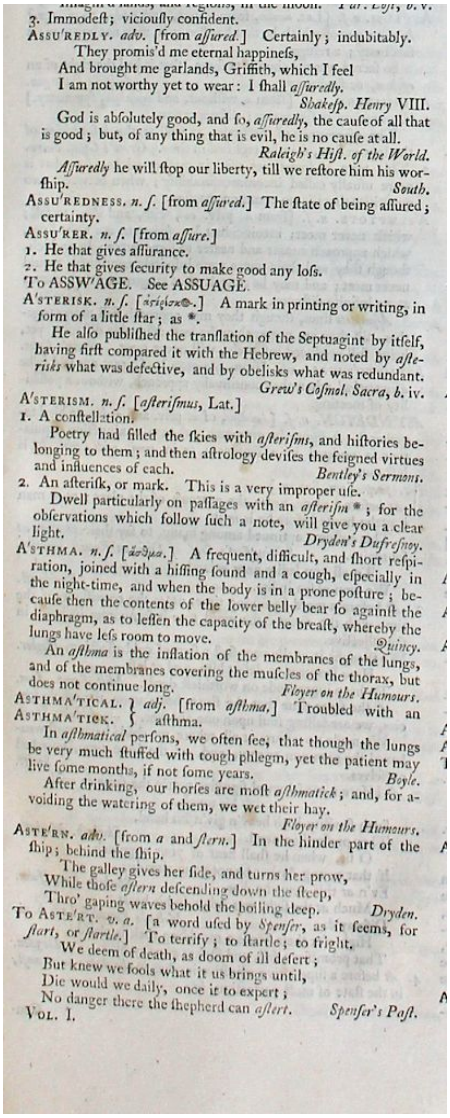
    c. Optimize: Since our last step was just an estimate, we need to minimize the error using scipy to minimize the functions. In this step, the "Powell" method [13] is used, which finds the minimum of a multi variable function without taking derivatives. This method is very useful in our case since it allows for the local minimum to be acquired without the need for any derivatives. We do not know what the parameters of our page equations will be, so we do not always know if they will be differentiable or not.

    d. Remap image and apply final threshold: Once the optimization is complete, the final dewarped image can be obtained. Using a coordinate remap with the rotation and translation vectors, as well as the two slopes, the image can be projected from $\mathbb{R}^3$ to our image plane $\mathbb{R}^2$. This is achieved by using OpenCV's

`projectPoints` [12] and `remap` methods. Finally after projecting and remapping the coordinates, the final image is run through an adaptive threshold once more.

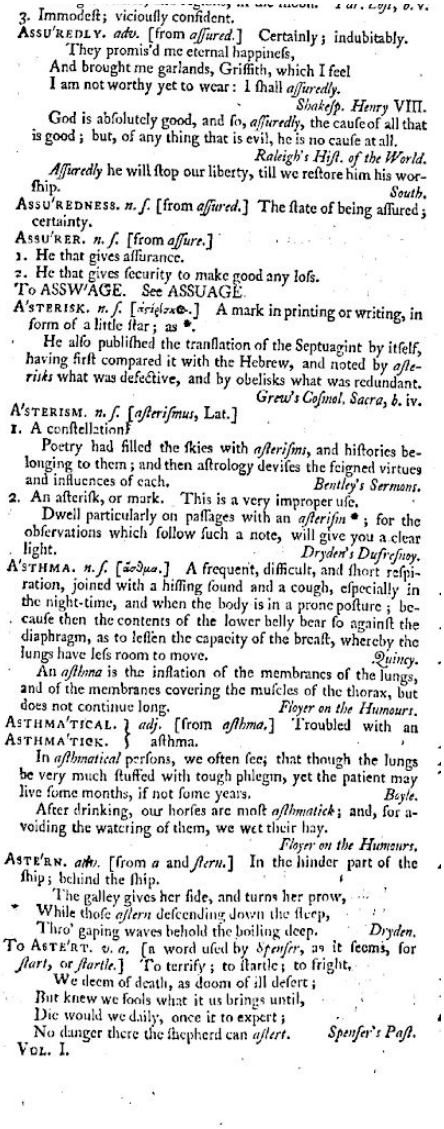
5.2.5 Initial results and modifications

The original algorithm [9] we found worked almost flawlessly, when we ran our image through it, we had incredible results. We originally believed that the best approach for our problem was to first break the page in two and then fix the warping on each side before detecting the final boundaries and the first result is shown below.



3. Immodeſt; viciouſly confident.
ASSU'REDLY. *adv.* [from *aſſured.*] Certainly; indubitably.
They promis'd me eternal happineſs,
And brought me garlands, Griffith, which I feel
I am not worthy yet to wear: I ſhall *aſſuredly.*
Shakeſp. Henry VIII.
God is abſolutely good, and ſo, *aſſuredly*, the cauſe of all that is good; but, of any thing that is evil, he is no cauſe at all.
Raleigh's Hiſt. of the World.
Aſſuredly he will ſtop our liberty, till we reſtore him his worſhip. *South.*
ASSU'REDNESS. *n. ſ.* [from *aſſured.*] The ſtate of being aſſured; certainty.
ASSU'RER. *n. ſ.* [from *aſſure.*]
1. He that gives aſſurance.
2. He that gives ſecurity to make good any loſs.
To ASSW'AGE. See ASSUAGE.
A'STERISK. *n. ſ.* [ἀστερίσκος.] A mark in printing or writing, in form of a little ſtar; as *.
He alſo publiſhed the tranſlation of the Septuagint by itſelf, having firſt compared it with the Hebrew, and noted by *aſteriſks* what was defective, and by obelisks what was redundant.
Grew's Coſmol. Sacra, b. iv.
A'STERISM. *n. ſ.* [*aſteriſmus*, Lat.]
1. A conſtellation.
Poetry had filled the ſkies with *aſteriſms*, and hiſtories belonging to them; and then aſtrology deviſes the feigned virtues and influences of each. *Bentley's Sermons.*
2. An aſteriſk, or mark. This is a very improper uſe.
Dwell particularly on paſſages with an *aſteriſm* *; for the obſervations which follow ſuch a note, will give you a clear light. *Dryden's Dufreſnoy.*
A'STHMA. *n. ſ.* [ἄσθμα.] A frequent, difficult, and ſhort reſpiration, joined with a hiſſing ſound and a cough, eſpecially in the night-time, and when the body is in a prone poſture; becauſe then the contents of the lower belly bear ſo againſt the diaphragm, as to leſſen the capacity of the breaſt, whereby the lungs have leſs room to move. *Quincy.*
An *aſthma* is the inflation of the membranes of the lungs, and of the membranes covering the muſcles of the thorax, but does not continue long. *Floyer on the Humours.*
ASTHMA'TICAL. } *adj.* [from *aſthma.*] Troubled with an
ASTHMA'TICK. } aſthma.
In *aſthmatical* perſons, we often ſee; that though the lungs be very much ſtuffed with tough phlegm, yet the patient may live ſome months, if not ſome years. *Boyle.*
After drinking, our horſes are moſt *aſthmatick*; and, for avoiding the watering of them, we wet their hay.
Floyer on the Humours.
ASTE'RN. *adv.* [from *a* and *ſtern.*] In the hinder part of the ſhip; behind the ſhip.
The galley gives her ſide, and turns her prow,
While thoſe *aſtern* deſcending down the ſteep,
Thro' gaping waves behold the boiling deep. *Dryden.*
To ASTE'RT. *v. a.* [a word uſed by *Spenſer*, as it ſeems, for *ſtart*, or *ſtartle.*] To terrify; to ſtartle; to fright.
We deem of death, as doom of ill deſert;
But knew we fools what it us brings until,
Die would we daily, once it to expert;
No danger there the ſhepherd can *aſtert.* *Spenſer's Paſt.*
VOL. I.

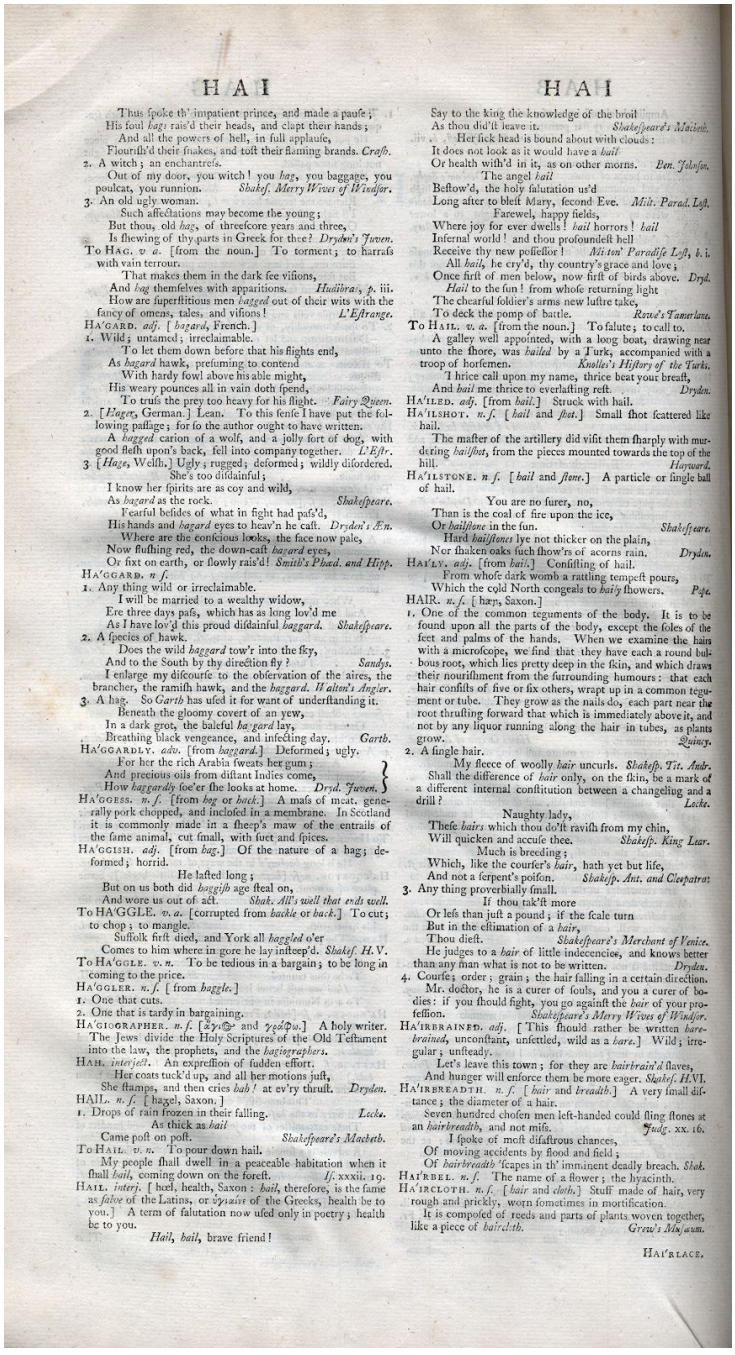
Figure 19: Original warped image



3. Immodeſt; viciouſly confident.
ASSU'REDLY. *adv.* [from *aſſured.*] Certainly; indubitably.
They promis'd me eternal happineſs,
And brought me garlands, Griffith, which I feel
I am not worthy yet to wear: I ſhall *aſſuredly.*
Shakeſp. Henry VIII.
God is abſolutely good, and ſo, *aſſuredly*, the cauſe of all that is good; but, of any thing that is evil, he is no cauſe at all.
Raleigh's Hiſt. of the World.
Aſſuredly he will ſtop our liberty, till we reſtore him his worſhip. *South.*
ASSU'REDNESS. *n. ſ.* [from *aſſured.*] The ſtate of being aſſured; certainty.
ASSU'RER. *n. ſ.* [from *aſſure.*]
1. He that gives aſſurance.
2. He that gives ſecurity to make good any loſs.
To ASSW'AGE. See ASSUAGE.
A'STERISK. *n. ſ.* [ἀστερίσκος.] A mark in printing or writing, in form of a little ſtar; as *.
He alſo publiſhed the tranſlation of the Septuagint by itſelf, having firſt compared it with the Hebrew, and noted by *aſteriſks* what was defective, and by obelisks what was redundant.
Grew's Coſmol. Sacra, b. iv.
A'STERISM. *n. ſ.* [*aſteriſmus*, Lat.]
1. A conſtellation.
Poetry had filled the ſkies with *aſteriſms*, and hiſtories belonging to them; and then aſtrology deviſes the feigned virtues and influences of each. *Bentley's Sermons.*
2. An aſteriſk, or mark. This is a very improper uſe.
Dwell particularly on paſſages with an *aſteriſm* *; for the obſervations which follow ſuch a note, will give you a clear light. *Dryden's Dufreſnoy.*
A'STHMA. *n. ſ.* [ἄσθμα.] A frequent, difficult, and ſhort reſpiration, joined with a hiſſing ſound and a cough, eſpecially in the night-time, and when the body is in a prone poſture; becauſe then the contents of the lower belly bear ſo againſt the diaphragm, as to leſſen the capacity of the breaſt, whereby the lungs have leſs room to move. *Quincy.*
An *aſthma* is the inflation of the membranes of the lungs, and of the membranes covering the muſcles of the thorax, but does not continue long. *Floyer on the Humours.*
ASTHMA'TICAL. } *adj.* [from *aſthma.*] Troubled with an
ASTHMA'TICK. } aſthma.
In *aſthmatical* perſons, we often ſee; that though the lungs be very much ſtuffed with tough phlegm, yet the patient may live ſome months, if not ſome years. *Boyle.*
After drinking, our horſes are moſt *aſthmatick*; and, for avoiding the watering of them, we wet their hay.
Floyer on the Humours.
ASTE'RN. *adv.* [from *a* and *ſtern.*] In the hinder part of the ſhip; behind the ſhip.
The galley gives her ſide, and turns her prow,
While thoſe *aſtern* deſcending down the ſteep,
Thro' gaping waves behold the boiling deep. *Dryden.*
To ASTE'RT. *v. a.* [a word uſed by *Spenſer*, as it ſeems, for *ſtart*, or *ſtartle.*] To terrify; to ſtartle; to fright.
We deem of death, as doom of ill deſert;
But knew we fools what it us brings until,
Die would we daily, once it to expert;
No danger there the ſhepherd can *aſtert.* *Spenſer's Paſt.*
VOL. I.

Figure 20: Initial fixed output

From here we realized that it would be easier to first unwarp the entire page before splitting the page into two columns. If the entire page was warped, our algorithm could not properly detect boundaries; those results are below.



HAI

Thus ſpoke th' impatient prince, and made a pauſe;
His foul *hags* rais'd their heads, and clapt their hands;
And all the powers of hell, in full applauſe,
Flouriſh'd their ſnakes, and toſs their flaming brands. *Craſh.*

2. A witch; an enchantreſs.
Out of my door, you witch! you *hag*, you baggage, you
poulcat, you runnion. *Shakeſ. Merry Wives of Windſor.*

3. An old ugly woman.
Such affections may become the young;
But thou, old *hag*, of threeſcore years and three,
Is ſhewing of thy parts in Greek for thee? *Dryden's Juven.*

To HAG. *v. a.* [from the noun.] To torment; to harraſs with vain terrour.
That makes them in the dark ſee viſions,
And *hag* themſelves with apparitions. *Hudibras, p.* iii.
How are ſuperſtitious men *hagged* out of their wits with the fancy of omens, tales, and viſions! *L'Eſtrange.*

HA'GGARD. *adj.* [*hagard*, French.]
1. Wild; untamed; irreclaimable.
To let them down before that his flights end,
As *hagard* hawk, preſuming to contend
With hardy fowl above his able might,
His weary pounces all in vain doth ſpend,
To truſs the prey too heavy for his flight. *Fairy Queen.*

2. [*Hager*, German.] Lean. To this ſenſe I have put the following paſſage; for ſo the author ought to have written.
A *hagged* cation of a wolf, and a jolly ſort of dog, with good fleſh upon's back, fell into company together. *L'Eſtr.*

3. [*Hage*, Welſh.] Ugly; rugged; deformed; wildly diſordered.
She's too diſdainful;
I know her ſpirits are as coy and wild,
As *hagard* as the rock. *Shakeſpeare.*
Fearful beſides of what in fight had paſs'd,
His hands and *hagard* eyes to heav'n he caſt. *Dryden's Æn.*
Where are the conſcious looks, the face now pale,
Now bluſhing red, the down-caſt *hagard* eyes,
Or fixt on earth, or ſlowly rais'd! *Smith's Phad. and Hipp.*

HA'GGARD. *n. ſ.*
1. Any thing wild or irreclaimable.
I will be married to a wealthy widow,
Ere three days paſs, which has as long lov'd me
As I have lov'd this proud diſdainful *haggard*. *Shakeſpeare.*

2. A ſpecies of hawk.
Does the wild *haggard* tow'r into the ſky,
And to the South by thy direction fly? *Sandys.*
I enlarge my diſcourſe to the obſervation of the aires, the brancher, the ramiſh hawk, and the *haggard*. *Walton's Angler.*

3. A hag. So *Garth* has uſed it for want of underſtanding it.
Beneath the gloomy covert of an yew,
In a dark grot, the baleful *haggard* lay,
Breathing black vengeance, and infecting day. *Garth.*

HA'GGARDLY. *adv.* [from *haggard*.] Deformed; ugly.
For her the rich Arabia ſweats her gum;
And precious oils from diſtant Indies come,
How *haggardly* ſoe'er the looks at home. *Dryd. Juvenal.*

HA'GGESS. *n. ſ.* [from *hog* or *back*.] A maſs of meat, generally pork chopped, and incloſed in a membrane. In Scotland it is commonly made in a ſheep's maw of the entrails of the ſame animal, cut ſmall, with ſuet and ſpices.

HA'GGISH. *adj.* [from *hag*.] Of the nature of a hag; deformed; horrid.
He laſted long;
But on us both did *haggiſh* age ſteal on,
And wore us out of act. *Shak. All's well that ends well.*

To HA'GGLE. *v. a.* [corrupted from *haikle* or *hack*.] To cut; to chop; to mangle.
Suffolk firſt died, and York all *haggled* o'er
Comes to him where in gore he lay inſteep'd. *Shakeſ. H. V.*

To HA'GGLE. *v. n.* To be tedious in a bargain; to be long in coming to the price.

HA'GGLER. *n. ſ.* [from *haggle*.]
1. One that cuts.
2. One that is tardy in bargaining.

HAGIOGRAPHER. *n. ſ.* [ἅγιος and γράφω.] A holy writer. The Jews divide the Holy Scriptures of the Old Teſtament into the law, the prophets, and the *hagiographers*.

HAH. *interject.* An expreſſion of ſudden effort.
Her coats tuck'd up, and all her motions juſt,
She ſtamps, and then cries *hah!* at ev'ry thruſt. *Dryden.*

HAIL. *n. ſ.* [hagel, Saxon.]
1. Drops of rain frozen in their falling. *Locke.*
As thick as *hail*
Came poſt on poſt. *Shakeſpeare's Macbeth.*

To HAIL. *v. n.* To pour down hail.
My people ſhall dwell in a peaceable habitation when it ſhall *hail*, coming down on the foreſt. *Iſ.* xxxii. 19.

HAIL. *interj.* [hoel, health, Saxon: *hail*, therefore, is the ſame as *ſalve* of the Latins, or ὑγιαίνε of the Greeks, health be to you.] A term of ſalutation now uſed only in poetry; health be to you.
Hail, hail, brave friend!

HAI

Say to the king the knowledge of the broil
As thou did'ſt leave it. *Shakeſpeare's Macbeth.*
Her ſick head is bound about with clouds:
It does not look as it would have a *hail*
Or health with'd in it, as on other morns. *Ben. Johnſon.*
The angel *hail*
Beſtow'd, the holy ſalutation us'd
Long after to bleſt Mary, ſecond Eve. *Milt. Parad. Loſt.*
Farewel, happy fields,
Where joy for ever dwells! *hail* horrors! *hail*
Infernal world! and thou profoundeſt hell
Receive thy new poſſeſſor! *Milton' Paradiſe Loſt*, b. i.
All *hail*, he cry'd, thy country's grace and love;
Once firſt of men below, now firſt of birds above. *Dryd.*
Hail to the ſun! from whoſe returning light
The chearful ſoldier's arms new luſtre take,
To deck the pomp of battle. *Rowe's Tamerlane.*

To HAIL. *v. a.* [from the noun.] To ſalute; to call to.
A galley well appointed, with a long boat, drawing near unto the ſhore, was *hailed* by a Turk, accompanied with a troop of horſemen. *Knolles's Hiſtory of the Turks.*
Thrice call upon my name, thrice beat your breaſt,
And *hail* me thrice to everlaſting reſt. *Dryden.*

HA'ILED. *adj.* [from *hail*.] Struck with hail.

HA'ILSHOT. *n. ſ.* [*hail* and *ſhot*.] Small ſhot ſcattered like hail.
The maſter of the artillery did viſit them ſharply with murdering *hailſhot*, from the pieces mounted towards the top of the hill. *Hayward.*

HA'ILSTONE. *n. ſ.* [*hail* and *ſtone*.] A particle or ſingle ball of hail.
You are no furer, no,
Than is the coal of fire upon the ice,
Or *hailſtone* in the ſun. *Shakeſpeare.*
Hard *hailſtones* lye not thicker on the plain,
Nor ſhaken oaks ſuch ſhow'rs of acorns rain. *Dryden.*

HA'ILY. *adj.* [from *hail*.] Conſiſting of hail.
From whoſe dark womb a rattling tempeſt pours,
Which the cold North congeals to *hail*y ſhowers. *Pope.*

HAIR. *n. ſ.* [hær, Saxon.]
1. One of the common teguments of the body. It is to be found upon all the parts of the body, except the ſoles of the feet and palms of the hands. When we examine the hairs with a microſcope, we find that they have each a round bulbous root, which lies pretty deep in the ſkin, and which draws their nouriſhment from the ſurrounding humours: that each hair conſiſts of five or ſix others, wrapt up in a common tegument or tube. They grow as the nails do, each part near the root thruſting forward that which is immediately above it, and not by any liquor running along the hair in tubes, as plants grow. *Quincy.*

2. A ſingle hair.
My fleece of woolly *hair* uncurls. *Shakeſp. Tit. Andr.*
Shall the difference of *hair* only, on the ſkin, be a mark of a different internal conſtitution between a changeling and a drill? *Locke.*
Naughty lady,
Theſe *hairs* which thou do'ſt raviſh from my chin,
Will quicken and accuſe thee. *Shakeſp. King Lear.*
Much is breeding;
Which, like the courſer's *hair*, hath yet but life,
And not a ſerpent's poiſon. *Shakeſp. Ant. and Cleopatra.*

3. Any thing proverbially ſmall.
If thou tak'ſt more
Or leſs than juſt a pound; if the ſcale turn
But in the eſtimation of a *hair*,
Thou dieſt. *Shakeſpeare's Merchant of Venice.*
He judges to a *hair* of little indecencies, and knows better than any man what is not to be written. *Dryden.*

4. Courſe; order; grain; the hair falling in a certain direction.
Mr. doctor, he is a curer of ſouls, and you a curer of bodies: if you ſhould fight, you go againſt the *hair* of your profeſſion. *Shakeſpeare's Merry Wives of Windſor.*

HA'IRBRAINED. *adj.* [This ſhould rather be written *harebrained*, unconſtant, unſettled, wild as a *hare*.] Wild; irregular; unſteady.
Let's leave this town; for they are *hairbrain'd* ſlaves,
And hunger will enforce them be more eager. *Shakeſ. H. VI.*

HA'IRBREADTH. *n. ſ.* [*hair* and *breadth*.] A very ſmall diſtance; the diameter of a hair.
Seven hundred choſen men left-handed could ſling ſtones at an *hairbreadth*, and not miſs. *Judg.* xx. 16.
I ſpoke of moſt diſaſtrous chances,
Of moving accidents by flood and field;
Of *hairbreadth* 'ſcapes in th' imminent deadly breach. *Shak.*

HAI'RBELL. *n. ſ.* The name of a flower; the hyacinth.

HA'IRCLOTH. *n. ſ.* [*hair* and *cloth*.] Stuff made of hair, very rough and prickly, worn ſometimes in mortification.
It is compoſed of reeds and parts of plants woven together, like a piece of *haircloth*. *Grew's Muſæum.*

HAI'RLACE.

Figure 21: Original warped image

Thus ſpoke th' impatient prince, and made a pauſe;
His foul *hags* rais'd their heads, and clapt their hands;
And all the powers of hell, in full applauſe,
Flouriſh'd their ſnakes, and toſt their flaming brands. *Craſh.*

2. A witch; an enchantreſs.
Out of my door, you witch! you *hag*, you baggage, you poulcat, you runnion. *Shakeſ. Merry Wives of Windſor.*

3. An old ugly woman.
Such affectations may become the young;
But thou, old *hag*, of threeſcore years and three,
Is ſhewing of thy parts in Greek for thee? *Dryden's Juven.*

To HAG. *v. a.* [from the noun.] To torment; to harraſs with vain terrour.
That makes them in the dark ſee viſions,
And *hag* themſelves with apparitions. *Hudibras, p.* iii.
How are ſuperſtitious men *hagged* out of their wits with the fancy of omens, tales, and viſions! *L'Eſtrange.*

HA'GARD. *adj.* [*hagard*, French.]
1. Wild; untamed; irreclaimable.
To let them down before that his flights end,
As *hagard* hawk, preſuming to contend
With hardy fowl above his able might,
His weary pounces all in vain doth ſpend,
To truſs the prey too heavy for his flight. *Fairy Queen.*

2. [*Hager*, German.] Lean. To this ſenſe I have put the following paſſage; for ſo the author ought to have written.
A *hagged* carion of a wolf, and a jolly ſort of dog, with good fleſh upon's back, fell into company together. *L'Eſtr.*

3. [*Hage*, Welſh.] Ugly; rugged; deformed; wildly diſordered.
She's too diſdainful;
I know her ſpirits are as coy and wild,
As *hagard* as the rock. *Shakeſpeare.*
Fearful beſides of what in fight had paſs'd,
His hands and *hagard* eyes to heav'n he caſt. *Dryden's Æn.*
Where are the conſcious looks, the face now pale,
Now flushing red, the down-caſt *hagard* eyes,
Or fixt on earth, or ſlowly rais'd! *Smith's Phæd. and Hipp.*

HA'GGARD. *n. ſ.*
1. Any thing wild or irreclaimable.
I will be married to a wealthy widow,
Ere three days paſs, which has as long lov'd me
As I have lov'd this proud diſdainful *haggard*. *Shakeſpeare.*

2. A ſpecies of hawk.
Does the wild *haggard* tow'r into the ſky,
And to the South by thy direction fly? *Sandys.*
I enlarge my diſcourſe to the obſervation of the aires, the brancher, the ramiſh hawk, and the *haggard*. *Walton's Angler.*

3. A hag. So *Garth* has uſed it for want of underſtanding it.
Beneath the gloomy covert of an yew,
In a dark grot, the baleful *haggard* lay,
Breathing black vengeance, and infecting day. *Garth.*

HA'GGARDLY. *adv.* [from *haggard*.] Deformed; ugly.
For her the rich Arabia ſweats her gum;
And precious oils from diſtant Indies come,
How *haggardly* ſoe'er ſhe looks at home. *Dryd. Juven.*

HA'GGESS. *n. ſ.* [from *hog* or *hack*.] A maſs of meat, generally pork chopped, and incloſed in a membrane. In Scotland it is commonly made in a ſheep's maw of the entrails of the ſame animal, cut ſmall, with ſuet and ſpices.

HA'GGISH. *adj.* [from *hag*.] Of the nature of a hag; deformed; horrid.
He laſted long;
But on us both did *haggiſh* age ſteal on,
And wore us out of act. *Shak. All's well that ends well.*

To HA'GGLE. *v. a.* [corrupted from *hackle* or *hack*.] To cut; to chop; to mangle.
Suffolk firſt died, and York all *haggled* o'er
Comes to him where in gore he lay inſteep'd. *Shakeſ. H. V.*

To HA'GGLE. *v. n.* To be tedious in a bargain; to be long in coming to the price.

HA'GGLER. *n. ſ.* [from *haggle*.]
1. One that cuts.
2. One that is tardy in bargaining.

HAGIOGRAPHER. *n. ſ.* [ἁγιογράφος and γράφω.] A holy writer. The Jews divide the Holy Scriptures of the Old Teſtament into the law, the prophets, and the *hagiographers*.

HAH. *interjeſt.* An expreſſion of ſudden effort.
Her coats tuck'd up, and all her motions juſt,
She ſtamps, and then cries *hah!* at ev'ry thruſt. *Dryden.*

HAIL. *n. ſ.* [hagel, Saxon.]
1. Drops of rain frozen in their falling. *Locke.*
As thick as *hail*
Came poſt on poſt. *Shakeſpeare's Macbeth.*

To HAIL. *v. n.* To pour down hail.
My people ſhall dwell in a peaceable habitation when it ſhall *hail*, coming down on the foreſt. *Iſ.* xxxii. 19.

HAIL. *interj.* [hœl, health, Saxon: *hail*, therefore, is the ſame as *ſalve* of the Latins, or ὑγιαίνε of the Greeks, health be to you.] A term of ſalutation now uſed only in poetry; health be to you.
Hail, hail, brave friend!

Say to the king the knowledge of the broil
As thou did'ſt leave it. *Shakeſpeare's Macbeth.*
Her ſick head is bound about with clouds:
It does not look as it would have a *hail*
Or health with'd in it, as on other morns. *Ben. Johnſon.*
The angel *hail*
Beſtow'd, the holy ſalutation us'd
Long after to bleſt Mary, ſecond Eve. *Milt. Parad. Loſt.*
Farewel, happy fields,
Where joy for ever dwells! *hail* horrors! *hail*
Infernal world! and thou profoundeſt hell
Receive thy new poſſeſſor! *Milton' Paradiſe Loſt, b.* i.
All *hail*, he cry'd, thy country's grace and love;
Once firſt of men below, now firſt of birds above. *Dryd.*
Hail to the ſun! from whoſe returning light
The chearful ſoldier's arms new luſtre take,
To deck the pomp of battle. *Rowe's Tamerlane.*

To HAIL. *v. a.* [from the noun.] To ſalute; to call to.
A galley well appointed, with a long boat, drawing near unto the ſhore, was *hailed* by a Turk, accompanied with a troop of horſemen. *Knolles's Hiſtory of the Turks.*
Thrice call upon my name, thrice beat your breaſt,
And *hail* me thrice to everlaſting reſt. *Dryden.*

HA'ILED. *adj.* [from *hail*.] Struck with hail.

HA'ILSHOT. *n. ſ.* [*hail* and *ſhot*.] Small ſhot ſcattered like hail.
The maſter of the artillery did viſit them ſharply with murdering *hailſhot*, from the pieces mounted towards the top of the hill. *Hayward.*

HA'ILSTONE. *n. ſ.* [*hail* and *ſtone*.] A particle or ſingle ball of hail.
You are no furer, no,
Than is the coal of fire upon the ice,
Or *hailſtone* in the ſun. *Shakeſpeare.*
Hard *hailſtones* lye not thicker on the plain,
Nor ſhaken oaks ſuch ſhow'rs of acorns rain. *Dryden.*

HAI'LY. *adj.* [from *hail*.] Conſiſting of hail.
From whoſe dark womb a rattling tempeſt pours,
Which the cold North congeals to *haily* ſhowers. *Pope.*

HAIR. *n. ſ.* [hær, Saxon.]
1. One of the common teguments of the body. It is to be found upon all the parts of the body, except the ſoles of the feet and palms of the hands. When we examine the hairs with a microſcope, we find that they have each a round bulbous root, which lies pretty deep in the ſkin, and which draws their nouriſhment from the ſurrounding humours: that each hair conſiſts of five or ſix others, wrapt up in a common tegument or tube. They grow as the nails do, each part near the root thruſting forward that which is immediately above it, and not by any liquor running along the hair in tubes, as plants grow. *Quincy.*

2. A ſingle hair.
My fleece of woolly *hair* uncurls. *Shakeſp. Tit. Andr.*
Shall the difference of *hair* only, on the ſkin, be a mark of a different internal conſtitution between a changeling and a drill? *Locke.*
Naughty lady,
Theſe *hairs* which thou do'ſt raviſh from my chin,
Will quicken and accuſe thee. *Shakeſp. King Lear.*
Much is breeding;
Which, like the courſer's *hair*, hath yet but life,
And not a ſerpent's poiſon. *Shakeſp. Ant. and Cleopatra.*

3. Any thing proverbially ſmall.
If thou tak'ſt more
Or leſs than juſt a pound; if the ſcale turn
But in the eſtimation of a *hair*,
Thou dieſt. *Shakeſpeare's Merchant of Venice.*
He judges to a *hair* of little indecencies, and knows better than any man what is not to be written. *Dryden.*

4. Courſe; order; grain; the hair falling in a certain direction.
Mr. doctor, he is a curer of ſouls, and you a curer of bodies: if you ſhould fight, you go againſt the *hair* of your profeſſion. *Shakeſpeare's Merry Wives of Windſor.*

HA'IRBRAINED. *adj.* [This ſhould rather be written *harebrained*, unconſtant, unſettled, wild as a *hare*.] Wild; irregular; unſteady.
Let's leave this town; for they are *hairbrain'd* ſlaves,
And hunger will enforce them be more eager. *Shakeſ. H. VI.*

HA'IRBREADTH. *n. ſ.* [*hair* and *breadth*.] A very ſmall diſtance; the diameter of a hair.
Seven hundred choſen men left-handed could ſling ſtones at an *hairbreadth*, and not miſs. *Judg.* xx. 16.
I ſpoke of moſt diſaſtrous chances,
Of moving accidents by flood and field;
Of *hairbreadth* 'ſcapes in th' imminent deadly breach. *Shak.*

HAI'RBELL. *n. ſ.* The name of a flower; the hyacinth.

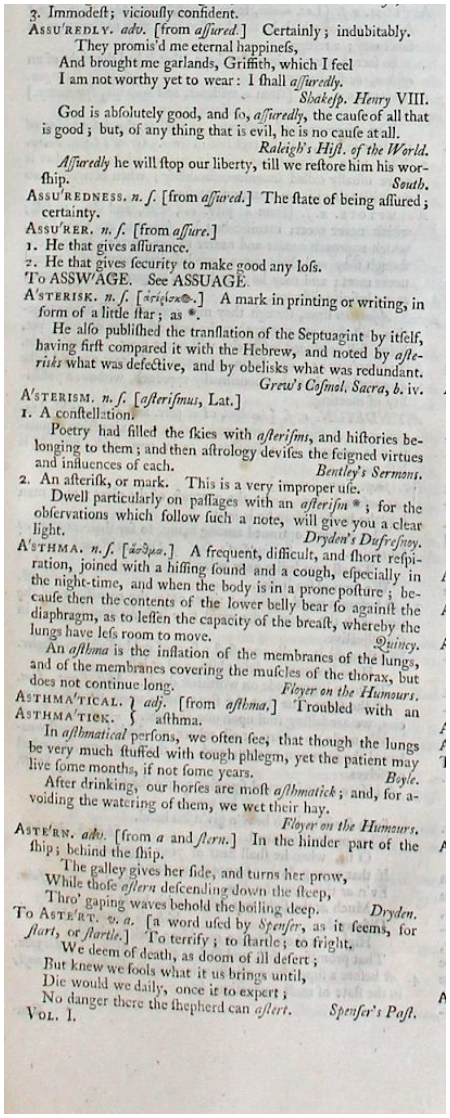
HA'IRCLOTH. *n. ſ.* [*hair* and *cloth*.] Stuff made of hair, very rough and prickly, worn ſometimes in mortification.
It is compoſed of reeds and parts of plants woven together, like a piece of *hair-cloth*. *Grew's Muſæum.*

Figure 22: Initial Fixed Output

Our example page has slight warping on the text along with many imperfections in the page that are no longer present in our unwarped document. While there were some anomalies in our tests of this method, we ultimately decided that this would be the way to go. To avoid any of those anomalies in the final project we will implement a threshold for warping, if our text is not warped at all we will simply skip the dewarping step and go right to the cropping step.

While these results are very promising, there are some present issues, so we analyzed the code to figure out what was happening. The first issue being that this program outputs a binary image. This program was originally designed to quickly scan and save images of pages of text that were sent by students when turning in homework. So for the original purpose, there was no need to maintain the original color picture as only the text of the image was important to see.

In step 2, an adaptive threshold is applied to the image to help detect text contours, but this is only temporary. Right before the image is saved, a final threshold is applied, to give the image the appearance of a scanned document. This is not an issue when exact document preservation is not necessary, as in the original use case, however we need the images to be preserved as per Dr. Giroux's request. In a very quick attempt to fix this, we simply stopped the program from changing the image to black and white before saving. To our surprise, this actually worked, but the resulting image had dithering applied to it, as shown below.



3. Immodeſt; viciouſly confident.

ASSU'REDLY. *adv.* [from *aſſured.*] Certainly; indubitably.

They promis'd me eternal happineſs,
And brought me garlands, Griffith, which I feel
I am not worthy yet to wear: I ſhall *aſſuredly.*
Shakeſp. Henry VIII.

God is abſolutely good, and ſo, *aſſuredly*, the cauſe of all that is good; but, of any thing that is evil, he is no cauſe at all.
Raleigh's Hiſt. of the World.

Aſſuredly he will ſtop our liberty, till we reſtore him his worſhip. *South.*

ASSU'REDNESS. *n. ſ.* [from *aſſured.*] The ſtate of being aſſured; certainty.

ASSU'RER. *n. ſ.* [from *aſſure.*]
1. He that gives aſſurance.
2. He that gives ſecurity to make good any loſs.

To ASSWAGE. See ASSUAGE.

A'STERISK. *n. ſ.* [ἀστερίσκος.] A mark in printing or writing, in form of a little ſtar; as *.

He alſo publiſhed the tranſlation of the Septuagint by itſelf, having firſt compared it with the Hebrew, and noted by *aſteriſks* what was defective, and by obeliſks what was redundant.
Grew's Coſmol. Sacra, b. iv.

A'STERISM. *n. ſ.* [*aſteriſmus*, Lat.]
1. A conſtellation.

Poetry had filled the ſkies with *aſteriſms*, and hiſtories belonging to them; and then aſtrology deviſes the feigned virtues and influences of each. *Bentley's Sermons.*

2. An aſteriſk, or mark. This is a very improper uſe.

Dwell particularly on paſſages with an *aſteriſm* *; for the obſervations which follow ſuch a note, will give you a clear light. *Dryden's Dufreſnoy.*

A'STHMA. *n. ſ.* [ἄσθμα.] A frequent, difficult, and ſhort reſpiration, joined with a hiſſing ſound and a cough, eſpecially in the night-time, and when the body is in a prone poſture; becauſe then the contents of the lower belly bear ſo againſt the diaphragm, as to leſſen the capacity of the breaſt, whereby the lungs have leſs room to move. *Quincy.*

An *aſthma* is the inflation of the membranes of the lungs, and of the membranes covering the muſcles of the thorax, but does not continue long. *Floyer on the Humours.*

ASTHMA'TICAL. } *adj.* [from *aſthma.*] Troubled with an
ASTHMA'TICK. } aſthma.

In *aſthmatical* perſons, we often ſee, that though the lungs be very much ſtuffed with tough phlegm, yet the patient may live ſome months, if not ſome years. *Boyle.*

After drinking, our horſes are moſt *aſthmatick*; and, for avoiding the watering of them, we wet their hay.
Floyer on the Humours.

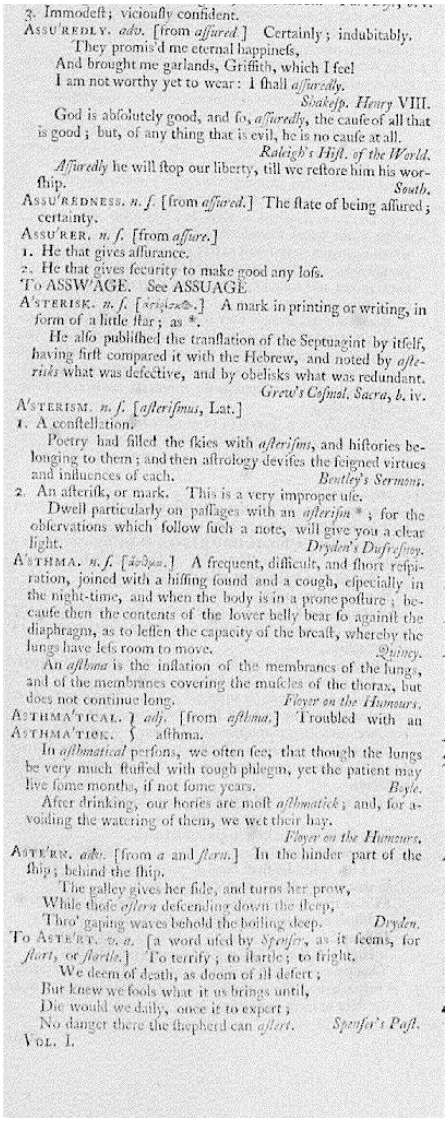
ASTE'RN. *adv.* [from *a* and *ſtern.*] In the hinder part of the ſhip; behind the ſhip.

The galley gives her ſide, and turns her prow,
While thoſe *aſtern* deſcending down the ſteep,
Thro' gaping waves behold the boiling deep. *Dryden.*

To ASTE'RT. *v. a.* [a word uſed by *Spenſer*, as it ſeems, for *ſtart*, or *ſtartle.*] To terrify; to ſtartle; to fright.

We deem of death, as doom of ill deſert;
But knew we fools what it us brings until,
Die would we daily, once it to expert;
No danger there the ſhepherd can *aſtert.* *Spenſer's Paſt.*

VOL. I.



3. Immodeſt; viciouſly confident.

ASSU'REDLY. *adv.* [from *aſſured.*] Certainly; indubitably.

They promis'd me eternal happineſs,
And brought me garlands, Griffith, which I feel
I am not worthy yet to wear: I ſhall *aſſuredly.*
Shakeſp. Henry VIII.

God is abſolutely good, and ſo, *aſſuredly*, the cauſe of all that is good; but, of any thing that is evil, he is no cauſe at all.
Raleigh's Hiſt. of the World.

Aſſuredly he will ſtop our liberty, till we reſtore him his worſhip. *South.*

ASSU'REDNESS. *n. ſ.* [from *aſſured.*] The ſtate of being aſſured; certainty.

ASSU'RER. *n. ſ.* [from *aſſure.*]
1. He that gives aſſurance.
2. He that gives ſecurity to make good any loſs.

To ASSWAGE. See ASSUAGE.

A'STERISK. *n. ſ.* [ἀστερίσκος.] A mark in printing or writing, in form of a little ſtar; as *.

He alſo publiſhed the tranſlation of the Septuagint by itſelf, having firſt compared it with the Hebrew, and noted by *aſteriſks* what was defective, and by obeliſks what was redundant.
Grew's Coſmol. Sacra, b. iv.

A'STERISM. *n. ſ.* [*aſteriſmus*, Lat.]
1. A conſtellation.

Poetry had filled the ſkies with *aſteriſms*, and hiſtories belonging to them; and then aſtrology deviſes the feigned virtues and influences of each. *Bentley's Sermons.*

2. An aſteriſk, or mark. This is a very improper uſe.

Dwell particularly on paſſages with an *aſteriſm* *; for the obſervations which follow ſuch a note, will give you a clear light. *Dryden's Dufreſnoy.*

A'STHMA. *n. ſ.* [ἄσθμα.] A frequent, difficult, and ſhort reſpiration, joined with a hiſſing ſound and a cough, eſpecially in the night-time, and when the body is in a prone poſture; becauſe then the contents of the lower belly bear ſo againſt the diaphragm, as to leſſen the capacity of the breaſt, whereby the lungs have leſs room to move. *Quincy.*

An *aſthma* is the inflation of the membranes of the lungs, and of the membranes covering the muſcles of the thorax, but does not continue long. *Floyer on the Humours.*

ASTHMA'TICAL. } *adj.* [from *aſthma.*] Troubled with an
ASTHMA'TICK. } aſthma.

In *aſthmatical* perſons, we often ſee, that though the lungs be very much ſtuffed with tough phlegm, yet the patient may live ſome months, if not ſome years. *Boyle.*

After drinking, our horſes are moſt *aſthmatick*; and, for avoiding the watering of them, we wet their hay.
Floyer on the Humours.

ASTE'RN. *adv.* [from *a* and *ſtern.*] In the hinder part of the ſhip; behind the ſhip.

The galley gives her ſide, and turns her prow,
While thoſe *aſtern* deſcending down the ſteep,
Thro' gaping waves behold the boiling deep. *Dryden.*

To ASTE'RT. *v. a.* [a word uſed by *Spenſer*, as it ſeems, for *ſtart*, or *ſtartle.*] To terrify; to ſtartle; to fright.

We deem of death, as doom of ill deſert;
But knew we fools what it us brings until,
Die would we daily, once it to expert;
No danger there the ſhepherd can *aſtert.* *Spenſer's Paſt.*

VOL. I.

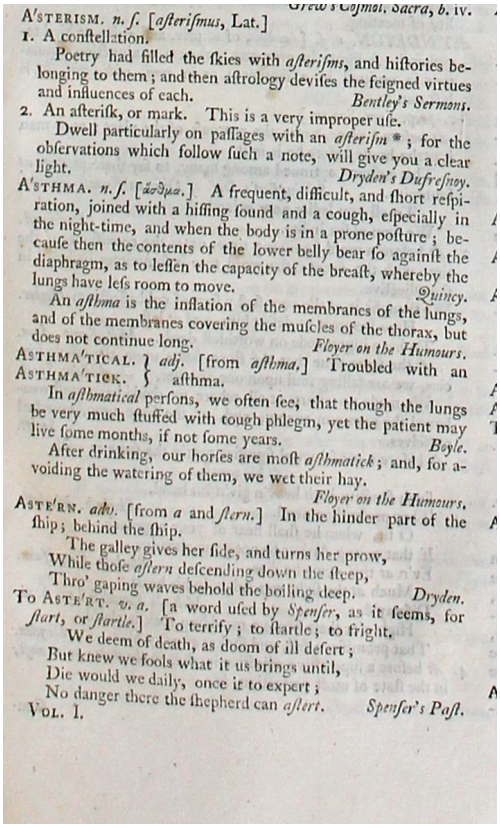
Figure 23: Original warped image

Figure 24: Corrected image with dithering

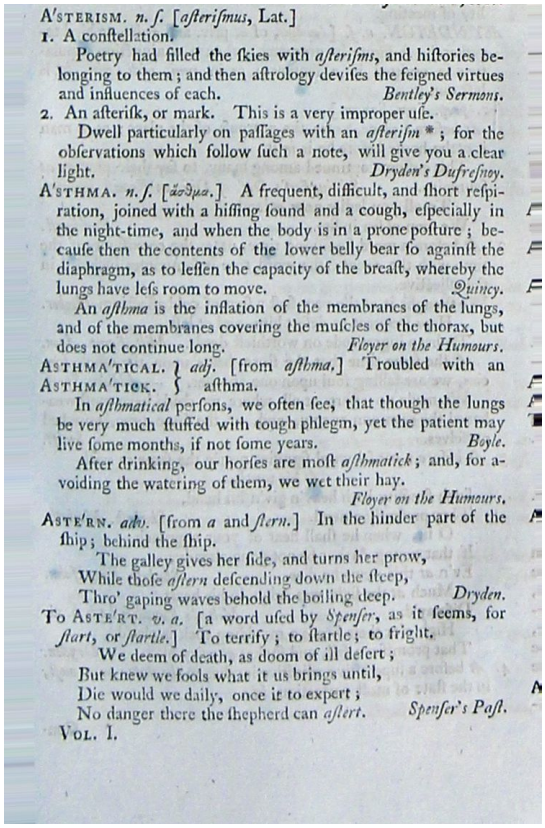
This result still keeps some of the original colors from the image, but has visual noise plastered all over it. Lastly, when saving the image we just removed the dithering and kept the original image, the results of this modification are pictured in the next section.

5.2.6 Initial Results

After modifying the program to output a color image, we had very successful results. Figures 25 and 26 show the results from running an image with noticeable page skew and text warp through the dewarping algorithm. However there are a few anomalies present.



A'STERISM. *n. ſ.* [*aſteriſmus*, Lat.]
1. A conſtellation.
Poetry had filled the ſkies with *aſteriſms*, and hiſtories belonging to them; and then aſtrology deviſes the feigned virtues and influences of each. *Bentley's Sermons.*
2. An aſteriſk, or mark. This is a very improper uſe.
Dwell particularly on paſſages with an *aſteriſm* *; for the obſervations which follow ſuch a note, will give you a clear light. *Dryden's Dufreſnoy.*
A'STHMA. *n. ſ.* [ἄσθμα.] A frequent, difficult, and ſhort reſpiration, joined with a hiſſing ſound and a cough, eſpecially in the night-time, and when the body is in a prone poſture; becauſe then the contents of the lower belly bear ſo againſt the diaphragm, as to leſſen the capacity of the breaſt, whereby the lungs have leſs room to move. *Quincy.*
An *aſthma* is the inflation of the membranes of the lungs, and of the membranes covering the muſcles of the thorax, but does not continue long. *Floyer on the Humours.*
ASTHMA'TICAL. } *adj.* [from *aſthma.*] Troubled with an
ASTHMA'TICK. } aſthma.
In *aſthmatical* perſons, we often ſee, that though the lungs be very much ſtuffed with tough phlegm, yet the patient may live ſome months, if not ſome years. *Boyle.*
After drinking, our horſes are moſt *aſthmatick*; and, for avoiding the watering of them, we wet their hay.
Floyer on the Humours.
ASTE'RN. *adv.* [from *a* and *ſtern.*] In the hinder part of the ſhip; behind the ſhip.
The galley gives her ſide, and turns her prow,
While thoſe *aſtern* deſcending down the ſteep,
Thro' gaping waves behold the boiling deep. *Dryden.*
To ASTE'RT. *v. a.* [a word uſed by *Spenſer*, as it ſeems, for *ſtart*, or *ſtartle.*] To terrify; to ſtartle; to fright.
We deem of death, as doom of ill deſert;
But knew we fools what it us brings until,
Die would we daily, once it to expert;
No danger there the ſhepherd can *aſtert.* *Spenſer's Paſt.*
VOL. I.



A'STERISM. *n. ſ.* [*aſteriſmus*, Lat.]
1. A conſtellation.
Poetry had filled the ſkies with *aſteriſms*, and hiſtories belonging to them; and then aſtrology deviſes the feigned virtues and influences of each. *Bentley's Sermons.*
2. An aſteriſk, or mark. This is a very improper uſe.
Dwell particularly on paſſages with an *aſteriſm* *; for the obſervations which follow ſuch a note, will give you a clear light. *Dryden's Dufreſnoy.*
A'STHMA. *n. ſ.* [ἄσθμα.] A frequent, difficult, and ſhort reſpiration, joined with a hiſſing ſound and a cough, eſpecially in the night-time, and when the body is in a prone poſture; becauſe then the contents of the lower belly bear ſo againſt the diaphragm, as to leſſen the capacity of the breaſt, whereby the lungs have leſs room to move. *Quincy.*
An *aſthma* is the inflation of the membranes of the lungs, and of the membranes covering the muſcles of the thorax, but does not continue long. *Floyer on the Humours.*
ASTHMA'TICAL. } *adj.* [from *aſthma.*] Troubled with an
ASTHMA'TICK. } aſthma.
In *aſthmatical* perſons, we often ſee, that though the lungs be very much ſtuffed with tough phlegm, yet the patient may live ſome months, if not ſome years. *Boyle.*
After drinking, our horſes are moſt *aſthmatick*; and, for avoiding the watering of them, we wet their hay.
Floyer on the Humours.
ASTE'RN. *adv.* [from *a* and *ſtern.*] In the hinder part of the ſhip; behind the ſhip.
The galley gives her ſide, and turns her prow,
While thoſe *aſtern* deſcending down the ſteep,
Thro' gaping waves behold the boiling deep. *Dryden.*
To ASTE'RT. *v. a.* [a word uſed by *Spenſer*, as it ſeems, for *ſtart*, or *ſtartle.*] To terrify; to ſtartle; to fright.
We deem of death, as doom of ill deſert;
But knew we fools what it us brings until,
Die would we daily, once it to expert;
No danger there the ſhepherd can *aſtert.* *Spenſer's Paſt.*
VOL. I.

Figure 25: Original warped image

Figure 26: Dewarped image

From the changes previously discussed, we were able to achieve an image of practically identical quality to the original, but as you can see, it has a cooler tone to it. The page itself no longer has yellow-ish hue, but rather a blue-ish tint and wherever there were brown age stains on the page they are now dark blue. We believe this is due to how the algorithm modifies the color channels when implementing the adaptive threshold. A possible solution to this discoloration is discussed in the next section.

5.2.7 Modifications

The biggest struggle is eliminating the present discoloration. We found a similar project done by Dropbox [14] in their automatic document scanning app and much like our solution, the color of the image is ignored. Instead of only applying the algorithm to a binary image, we can apply it to each RGB color channel, but this can lead to an issue called color constancy.

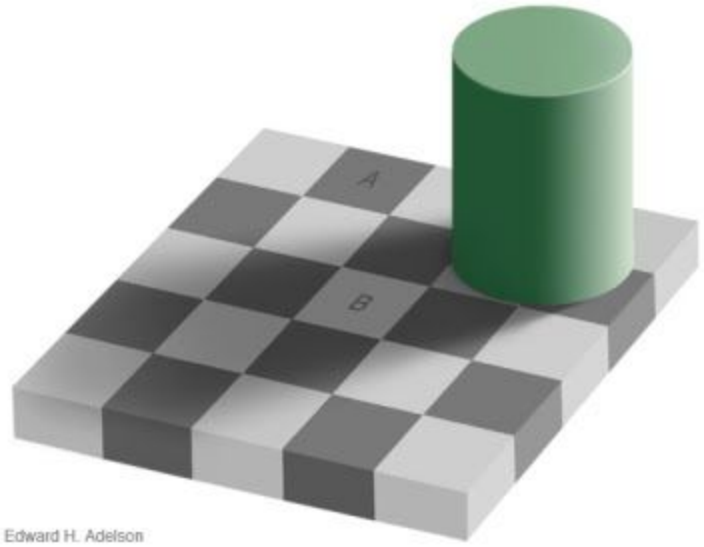


Figure 27 [14]: Color Constancy

Color constancy is a classic example of an optical illusion. The two squares A and B are the same color, but due to the shadow from the cylinder they appear to be different colors. We do not believe we will run into this issue since there are not any apparent heavy shadows in the scans, but also the only two colors that appear in the scans are the yellow-ish background of the pages themselves and the black text. If we do encounter this issue, we can elevate the saturation and hue of either the background or the text and this should help account for those cases.

The next issue is the image quality. Most of the images are high quality, but some are older scans and are not as high of quality. There is not really a solution to make sure the lower quality images are as high quality as the higher quality ones, but we ensure that our algorithm does not degrade the quality of the

image too much so the lower quality images do not end up looking any worse than they already do.

With the current implementation, there is no estimate of the page boundaries, in the program before running, we give a rough estimate of where we think the page boundaries are going to be based on our input size. As discussed previously, this led to warping occurring at the boundaries of our images. There is a solution to this issue proposed inAutomatic Borders Detection of Camera Document Images [15] that automatically determines the borders of a camera document image, which we will implement if we believe it is necessary. It works by carrying out the following steps:

1. Noisy black border detection and removal: In document images obtained from cameras, there are often things called noisy black borders. These are parts of the image that do not belong to the page and are extraneous portions that can be discarded. This is done in multiple steps.
    a. Horizontal and vertical smoothing are applied to the images using what is known as the Run Length Smoothing Algorithm (RLSA). This algorithm is mainly used for text block segmentation and an example is shown in figure 28 below.
    b. Connected Component Labelling is then applied to find the connected text components to keep within the page's boundaries.
    c. A vertical histogram is calculated to find the sum of the black pixels in each column of the image array. From this vertical histogram, the limit for the left side of the image is calculated. This is calculated by searching for the entire width of the vertical black border; however, if no vertical black border exists, this process terminates early. A similar process is applied to find the vertical black border on the right side of the image.
    d. A horizontal histogram is calculated to find the sum of the black pixels in each row of the image array. This process is similar to the border detection described in the previous step.
    e. Lastly, the black borders are removed. Any black pixels that are in a connected component that is out of the limits of the connected components of the text is transformed into a white pixel.
2. Noisy text region detection and removal: This step is useful for scans of images that span multiple pages. In this step, any text that is outside the boundary of our present page is detected and removed. In our scenario

this is not necessary so we do. We are only dealing with one page at a time and do not see a need to implement this.

The proposed method above will be very useful to help ensure that we are completely accurate in our cropping and dewarping, but there will need to be some modifications made for our implementation. For our page scans, only the page is present, there are no noisy black borders. To make this work for our solution, we will need to do almost the opposite of what is being proposed. Instead of removing black borders, we should remove the white borders on both sides of the text, so we can be left with just the text.

Figure 28 [84]: RLSA Applied to a news article

5.2.8 Noisy Black Border Removal Flowchart

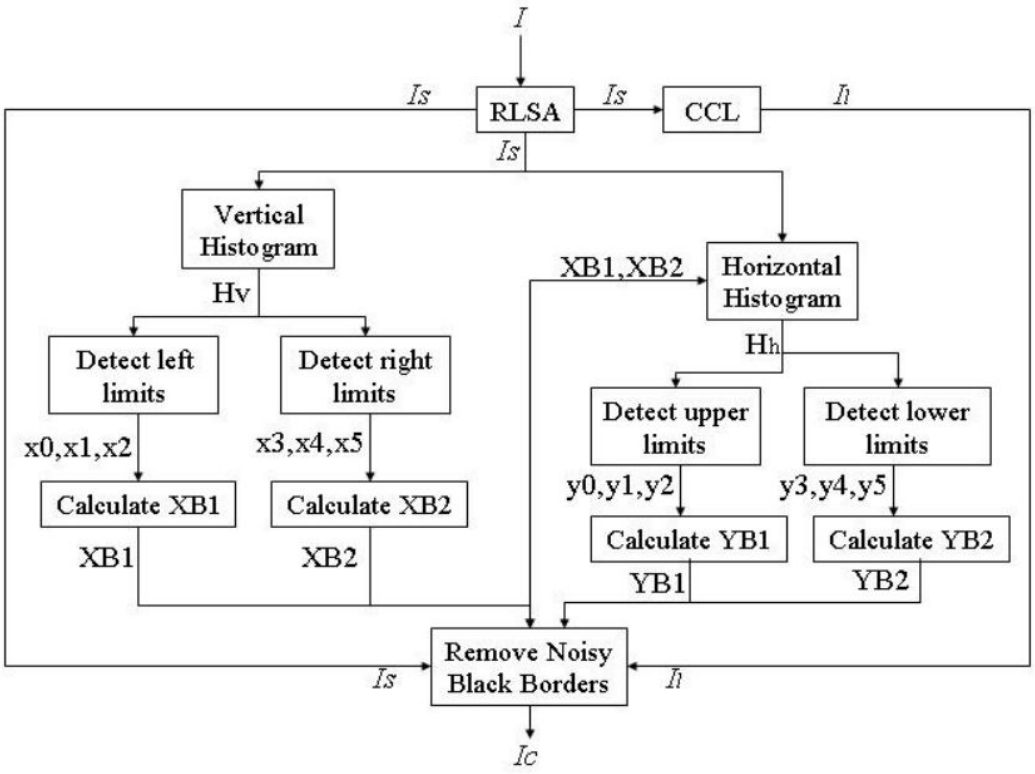


Figure 29 [15]: Black border removal flowchart

From the border removal process described earlier, it is much clearer to see here what is going on. The incoming image I has RLSA applied to it as well as connected-component labelling which breaks it down into I_s which are the sets of labels and then further into I_l which is the labelled image. I_s is also fed into the vertical and horizontal histograms to detect the page limits as well as $XB1, XB2, YB1,$ and $YB2$, which are the limits of the page itself. Lastly, the complete image I_c is produced.

5.3 Column Cropping

This project has several unique challenges that need to be addressed. The first step would be to recognize an entry word. Entry words are observed to have all capital letters and larger font size than the surrounding characters. In addition to the larger font size, entry words could sometimes be capitalized. We will use this fact as a criterion in determining what is an entry word and what likely is not. Additionally, we must take into consideration that the scanned documents also have guide words and have similar characteristics to entry words which can be observed below.

In order to make the algorithm be as successful as possible, we want to first trim any excess whitespace and unneeded text. The unneeded text consists of:

- A large letter at the top of the page to indicate where a new letters section begins

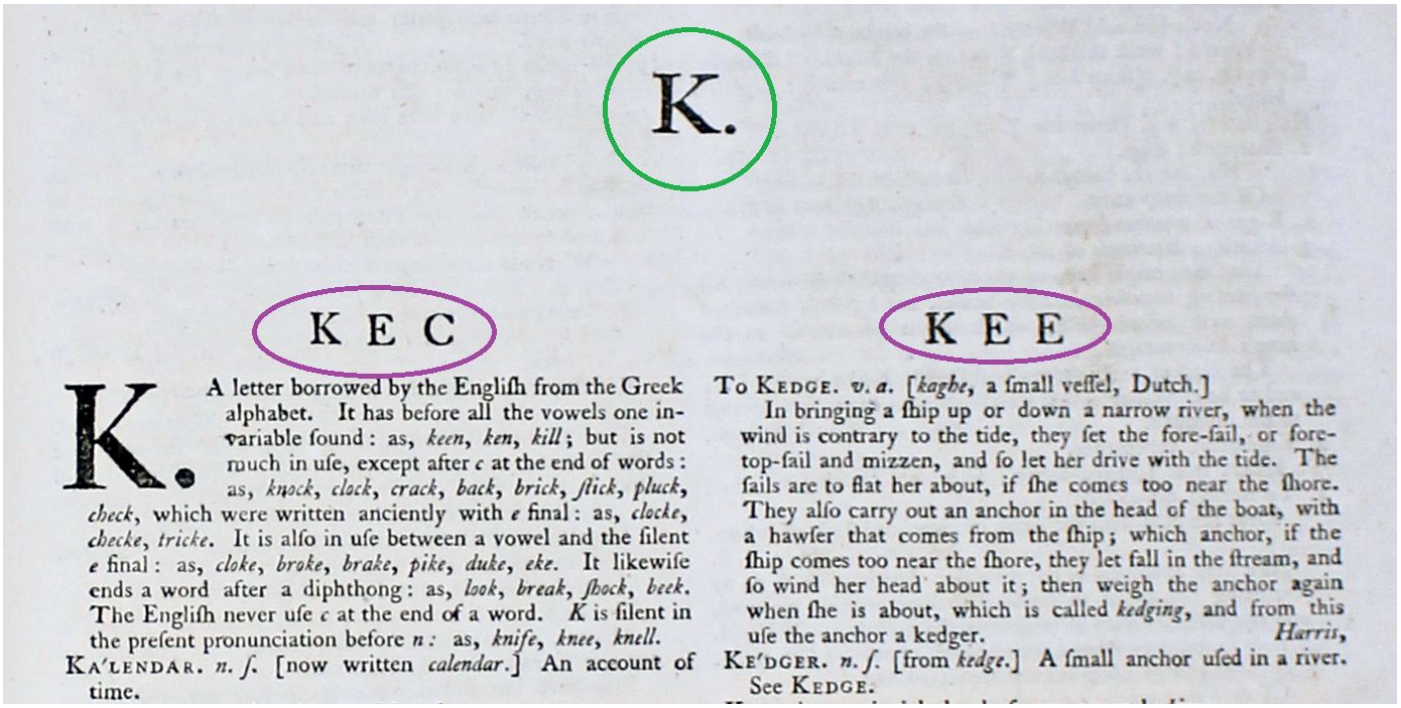
- Guide Words on the top of the page, which act as an index to the first and last words on that given page.

- Signature Marks on the bottom of the page, which tell where the page fits into the dictionary, or what volume it is in.

- Press Figure on the bottom of the page, which identifies the worker who set the type.

- Catchwords on the bottom right of the page, which tell you the first word that is on the next page. These are to help put the pages in order.

The following page contains images with examples of some of these obstructions. Circled in green is the letter section guide, purple are the guide words, blue are the press figures, and red are the catchwords. The signature mark is not present in this example, but it tells you the volume on the lower left side of the page.



K.

K E C

K E E

K. A letter borrowed by the Engliſh from the Greek alphabet. It has before all the vowels one invariable ſound: as, *keen*, *ken*, *kill*; but is not much in uſe, except after *c* at the end of words: as, *knock*, *clock*, *crack*, *back*, *brick*, *ſtick*, *pluck*, *check*, which were written anciently with *e* final: as, *clocke*, *checke*, *tricke*. It is alſo in uſe between a vowel and the ſilent *e* final: as, *cloke*, *broke*, *brake*, *pike*, *duke*, *eke*. It likewiſe ends a word after a diphthong: as, *look*, *break*, *ſhock*, *beek*. The Engliſh never uſe *c* at the end of a word. *K* is ſilent in the preſent pronunciation before *n*: as, *knife*, *knee*, *knell*.

KA'LENDAR. *n. ſ.* [now written *calendar*.] An account of time.

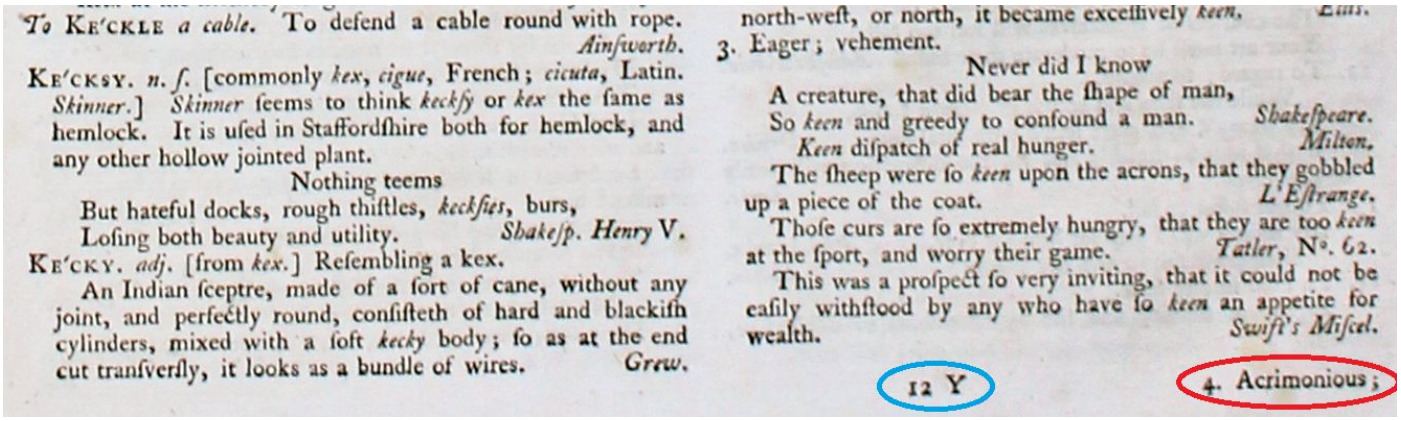
To KEDGE. *v. a.* [*kaghe*, a ſmall veſſel, Dutch.]

In bringing a ſhip up or down a narrow river, when the wind is contrary to the tide, they ſet the fore-ſail, or fore-top-ſail and mizzen, and ſo let her drive with the tide. The ſails are to ſlat her about, if ſhe comes too near the ſhore. They alſo carry out an anchor in the head of the boat, with a hawſer that comes from the ſhip; which anchor, if the ſhip comes too near the ſhore, they let fall in the ſtream, and ſo wind her head about it; then weigh the anchor again when ſhe is about, which is called *kedging*, and from this uſe the anchor a kedger. *Harris*,

KE'DGER. *n. ſ.* [from *kedge*.] A ſmall anchor uſed in a river. See KEDGE.

Figure 30: Letter Section and Guide Words

The guide words, in this instance, must be ignored as they are not part of the definition of any of the words being defined on the page. Another item that must also be ignored, during the processing of the image, are signature marks on the bottom of the page, press figures, and finally catch words. The items can be observed here.



To KE'CKLE *a cable.* To defend a cable round with rope. *Ainſworth.*

KE'CKSY. *n. ſ.* [commonly *kex*, *cigue*, French; *cicuta*, Latin. *Skinner*.] *Skinner* ſeems to think *keckſy* or *kex* the ſame as hemlock. It is uſed in Staffordſhire both for hemlock, and any other hollow jointed plant.

Nothing teems
But hateful docks, rough thiſtles, *keckſies*, burs,
Loſing both beauty and utility. *Shakeſp. Henry V.*

KE'CKY. *adj.* [from *kex*.] Reſembling a kex.

An Indian ſceptre, made of a ſort of cane, without any joint, and perfectly round, conſiſteth of hard and blackiſh cylinders, mixed with a ſoft *kecky* body; ſo as at the end cut tranſverſly, it looks as a bundle of wires. *Grew.*

north-weſt, or north, it became exceſſively *keen*.

3. Eager; vehement.

Never did I know
A creature, that did bear the ſhape of man,
So *keen* and greedy to confound a man. *Shakeſpeare.*
Keen diſpatch of real hunger. *Milton.*

The ſheep were ſo *keen* upon the acrons, that they gobbled up a piece of the coat. *L'Eſtrange.*

Thoſe curs are ſo extremely hungry, that they are too *keen* at the ſport, and worry their game. *Tatler*, N°. 62.

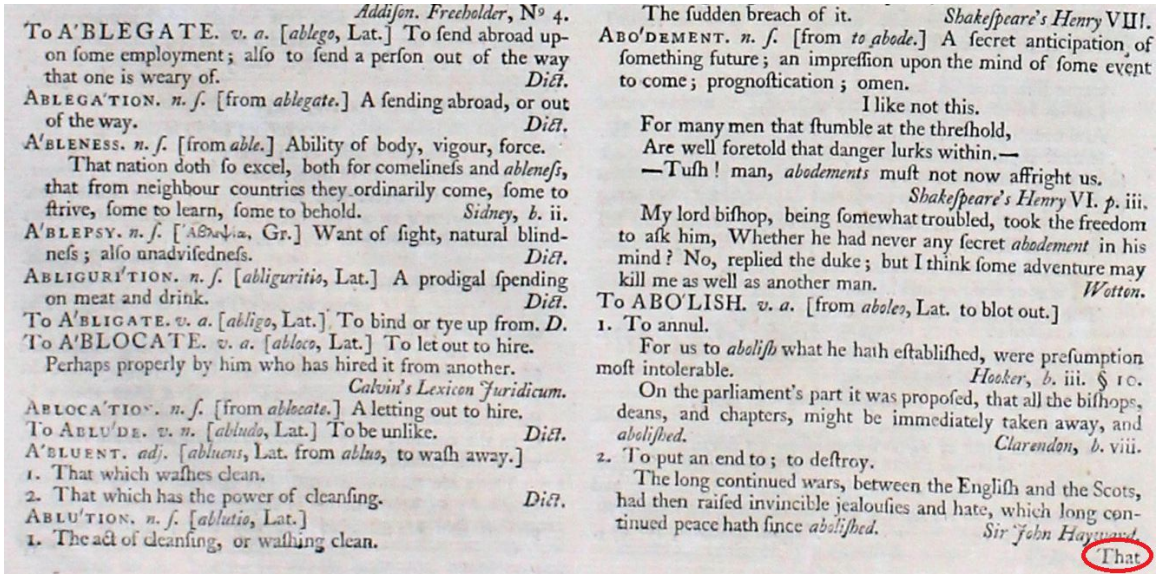
This was a proſpect ſo very inviting, that it could not be eaſily withſtood by any who have ſo *keen* an appetite for wealth. *Swift's Miſcel.*

12 Y

4. Acrimonious;

Figure 31: Press Figure and Catchword Example

These items were placed in the document to help identify the worker who scanned the document, where the page fits within the book, and the word on the far right, shows the first headword on the following page. Even though we were recommended by the sponsor to ignore the catchword on the far right, I believe that we can potentially use it to help us find the starting point on the following page. We could probably isolate that catchword, process it, and immediately look for that specific word in the next page.



Addiſon. Freeholder, N° 4.

To A'BLEGATE. *v. a.* [*ablego*, Lat.] To ſend abroad upon ſome employment; alſo to ſend a perſon out of the way that one is weary of. *Dict.*

ABLEGA'TION. *n. ſ.* [from *ablegate.*] A ſending abroad, or out of the way. *Dict.*

A'BLENESS. *n. ſ.* [from *able.*] Ability of body, vigour, force.

That nation doth ſo excel, both for comelineſs and *ableneſs*, that from neighbour countries they ordinarily come, ſome to ſtrive, ſome to learn, ſome to behold. *Sidney, b.* ii.

A'BLEPSY. *n. ſ.* [Ἀβλεψία, Gr.] Want of ſight, natural blindneſs; alſo unadviſedneſs. *Dict.*

ABLIGURI'TION. *n. ſ.* [*abliguritio*, Lat.] A prodigal ſpending on meat and drink. *Dict.*

To A'BLIGATE. *v. a.* [*abligo*, Lat.] To bind or tye up from. *D.*

To A'BLOCATE. *v. a.* [*ablocco*, Lat.] To let out to hire.

Perhaps properly by him who has hired it from another. *Calvin's Lexicon Juridicum.*

ABLOCA'TION. *n. ſ.* [from *ablocate.*] A letting out to hire.

To ABLU'DE. *v. n.* [*abludo*, Lat.] To be unlike. *Dict.*

A'BLUENT. *adj.* [*abluens*, Lat. from *abluo*, to waſh away.]

1. That which waſhes clean.
2. That which has the power of cleanſing. *Dict.*

ABLU'TION. *n. ſ.* [*ablutio*, Lat.]

1. The act of cleanſing, or waſhing clean.

The ſudden breach of it. *Shakeſpeare's Henry* VIII.

ABO'DEMENT. *n. ſ.* [from *to abode.*] A ſecret anticipation of ſomething future; an impreſſion upon the mind of ſome event to come; prognoſtication; omen.

I like not this.
For many men that ſtumble at the threſhold,
Are well foretold that danger lurks within.—
—Tuſh! man, *abodements* muſt not now affright us.
Shakeſpeare's Henry VI. *p.* iii.

My lord biſhop, being ſomewhat troubled, took the freedom to aſk him, Whether he had never any ſecret *abodement* in his mind? No, replied the duke; but I think ſome adventure may kill me as well as another man. *Wotton.*

To ABO'LISH. *v. a.* [from *aboleo*, Lat. to blot out.]

1. To annul.

For us to *aboliſh* what he hath eſtabliſhed, were preſumption moſt intolerable. *Hooker, b.* iii. § 10.

On the parliament's part it was propoſed, that all the biſhops, deans, and chapters, might be immediately taken away, and *aboliſhed.* *Clarendon, b.* viii.

2. To put an end to; to deſtroy.

The long continued wars, between the Engliſh and the Scots, had then raiſed invincible jealouſies and hate, which long continued peace hath ſince *aboliſhed.* *Sir John Hayward.*

That

Figure 32: Catch Word Obstruction Example

The first few things that we noticed could potentially be very useful are:

- The headwords that lead each definition are entirely capital letters.

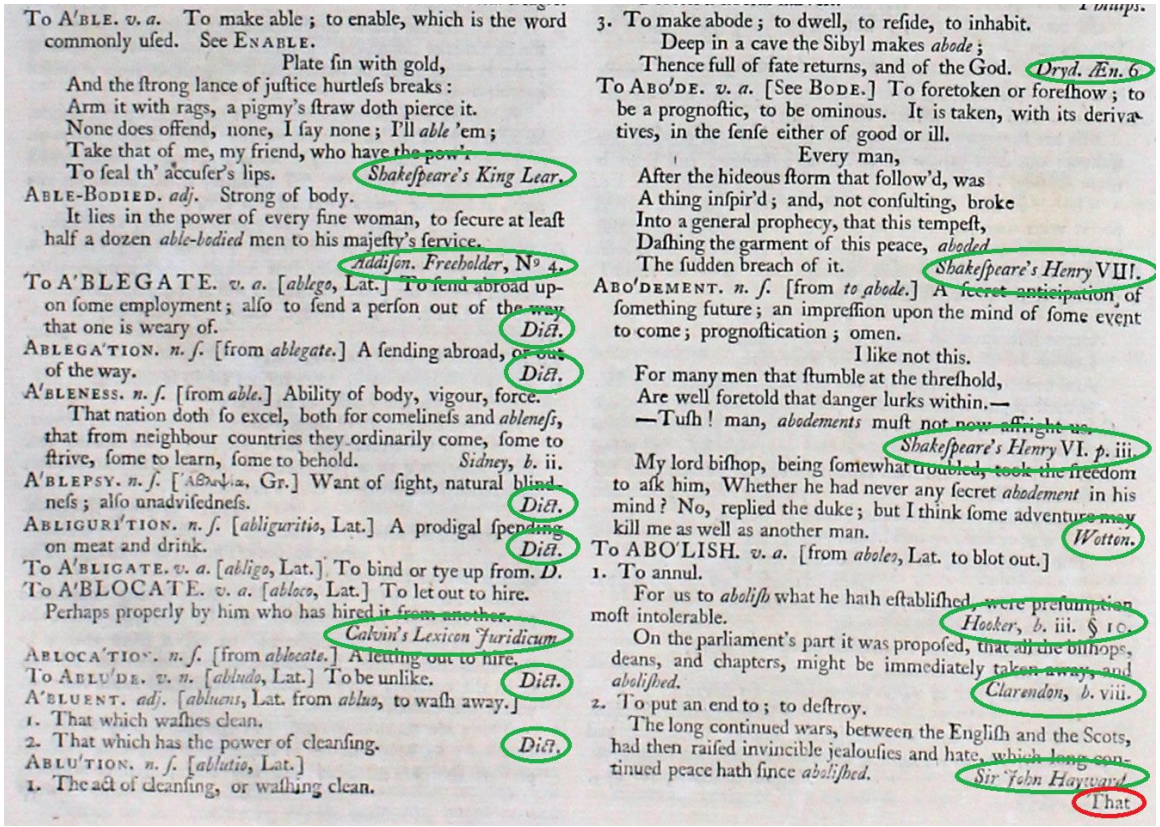
- The headwords are all the way to the left of the column, whereas the definition itself is indented.

The most apparent potential problems are:

- If a word has multiple definitions then the subsequent definitions stay on the same line as the headword. (Shown in figure 31 as the numbered definitions below “A’BLUENT.”)

- Sometimes the signature marks, press figures, and catchwords are up against the definitions on the bottom of the page. (Shown on the bottom right of figure 32 above.)

Now for the multiple definitions, the solution may actually be quite simple, as the numbers leading the definitions are relatively small and have a lot less of an impact than the headwords themselves. This can easily lead to the program being able to pick them out with little to no problems. As for the catchwords and other additional text being pushed up against the definitions on the bottom right of the page, the first thought that comes to mind is to check and see how far along to the right a word has been indented, and if there is a lot of empty space to the left of it then cut it out. This solution may work, except the references for the definitions are also indented all the way to the right, we just need to make sure our algorithm can pick and choose which is wanted, and which to get rid of.



To A'BLE. *v. a.* To make able; to enable, which is the word commonly uſed. See ENABLE.

Plate ſin with gold,
And the ſtrong lance of juſtice hurtleſs breaks:
Arm it with rags, a pigmy's ſtraw doth pierce it.
None does offend, none, I ſay none; I'll *able* 'em;
Take that of me, my friend, who have the pow'r
To ſeal th' accuſer's lips. *Shakeſpeare's King Lear.*

ABLE-BODIED. *adj.* Strong of body.

It lies in the power of every fine woman, to ſecure at leaſt half a dozen *able-bodied* men to his majeſty's ſervice.
Addiſon. Freeholder, N° 4.

To A'BLEGATE. *v. a.* [*ablego*, Lat.] To ſend abroad upon ſome employment; alſo to ſend a perſon out of the way that one is weary of. *Dict.*

ABLEGA'TION. *n. ſ.* [from *ablegate.*] A ſending abroad, or out of the way. *Dict.*

A'BLENESS. *n. ſ.* [from *able.*] Ability of body, vigour, force.

That nation doth ſo excel, both for comelineſs and *ableneſs*, that from neighbour countries they ordinarily come, ſome to ſtrive, ſome to learn, ſome to behold. *Sidney, b.* ii.

A'BLEPSY. *n. ſ.* [Ἀβλεψία, Gr.] Want of ſight, natural blindneſs; alſo unadviſedneſs. *Dict.*

ABLIGURI'TION. *n. ſ.* [*abliguritio*, Lat.] A prodigal ſpending on meat and drink. *Dict.*

To A'BLIGATE. *v. a.* [*abligo*, Lat.] To bind or tye up from. *D.*

To A'BLOCATE. *v. a.* [*abloco*, Lat.] To let out to hire.

Perhaps properly by him who has hired it from another.
Calvin's Lexicon Juridicum.

ABLOCA'TION. *n. ſ.* [from *ablocate.*] A letting out to hire.

To ABLU'DE. *v. n.* [*abludo*, Lat.] To be unlike. *Dict.*

A'BLUENT. *adj.* [*abluens*, Lat. from *ablue*, to waſh away.]

1. That which waſhes clean.
2. That which has the power of cleanſing. *Dict.*

ABLU'TION. *n. ſ.* [*ablutio*, Lat.]

1. The act of cleanſing, or waſhing clean.

3. To make abode; to dwell, to reſide, to inhabit.

Deep in a cave the Sibyl makes *abode*;
Thence full of fate returns, and of the God. *Dryd. Æn. 6.*

To ABO'DE. *v. a.* [See BODE.] To foretoken or foreſhow; to be a prognoſtic, to be ominous. It is taken, with its derivatives, in the ſenſe either of good or ill.

Every man,
After the hideous ſtorm that follow'd, was
A thing inſpir'd; and, not conſulting, broke
Into a general prophecy, that this tempeſt,
Daſhing the garment of this peace, *aboded*
The ſudden breach of it. *Shakeſpeare's Henry VIII.*

ABO'DEMENT. *n. ſ.* [from *to abode.*] A ſecret anticipation of ſomething future; an impreſſion upon the mind of ſome event to come; prognoſtication; omen.

I like not this.
For many men that ſtumble at the threſhold,
Are well foretold that danger lurks within.—
—Tuſh! man, *abodements* muſt not now affright us.
Shakeſpeare's Henry VI. p. iii.

My lord biſhop, being ſomewhat troubled, took the freedom to aſk him, Whether he had never any ſecret *abodement* in his mind? No, replied the duke; but I think ſome adventure may kill me as well as another man. *Wotton.*

To ABO'LISH. *v. a.* [from *aboleo*, Lat. to blot out.]

1. To annul.

For us to *aboliſh* what he hath eſtabliſhed, were preſumption moſt intolerable. *Hooker, b.* iii. § 10.

On the parliament's part it was propoſed, that all the biſhops, deans, and chapters, might be immediately taken away, and *aboliſhed.* *Clarendon, b.* viii.

2. To put an end to; to deſtroy.

The long continued wars, between the Engliſh and the Scots, had then raiſed invincible jealouſies and hate, which long continued peace hath ſince *aboliſhed.* *Sir John Hayward.*

That

Figure 33: References and Catchword

Examples of these can be shown in figure 33 above, where the catchword and signature mark are shown circled in red, but the references (which we want to keep in the definitions) are circled in green. In this case, the reference for the bottom right most definition may appear like a catchword to the algorithm and be removed. Due to the fact that there are so many pages in the dictionary, it is also

difficult to ensure that our algorithm is working perfectly on every single page, and for an issue like this one, it would be very easy to miss the fact that a reference was cut out of the definition.

The best way to capture the body of the word being defined is to first segregate both columns. The best way to separate both columns is to consider the space between the columns.

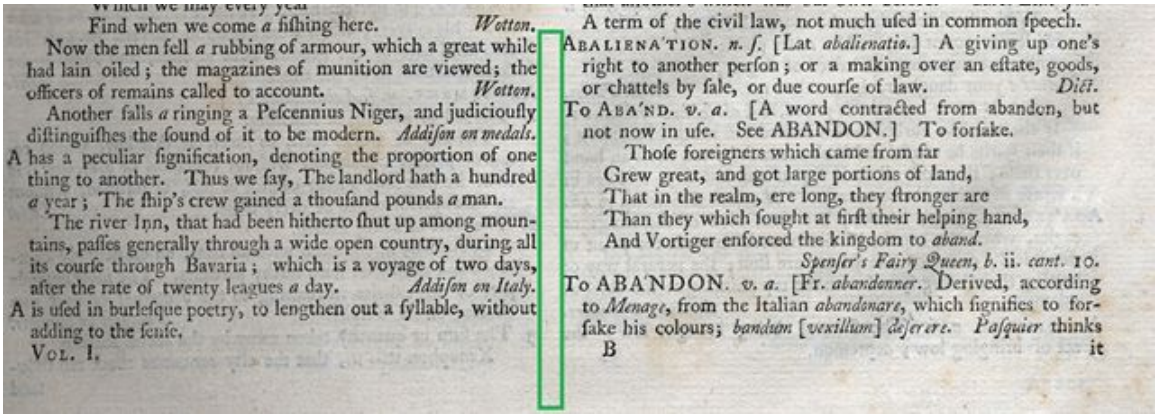

Find when we come *a* fishing here. *Wotton.*
Now the men fell *a* rubbing of armour, which a great while had lain oiled; the magazines of munition are viewed; the officers of remains called to account. *Wotton.*
Another falls *a* ringing a Pescennius Niger, and judiciously distinguishes the sound of it to be modern. *Addison on medals.*
A has a peculiar signification, denoting the proportion of one thing to another. Thus we say, The landlord hath a hundred *a* year; The ship's crew gained a thousand pounds *a* man.
The river Inn, that had been hitherto shut up among mountains, passes generally through a wide open country, during all its course through Bavaria; which is a voyage of two days, after the rate of twenty leagues *a* day. *Addison on Italy.*
A is used in burlesque poetry, to lengthen out a syllable, without adding to the sense.
VOL. I.
A term of the civil law, not much used in common speech.
ABALIENA'TION. *n. s.* [Lat. *abalienatio.*] A giving up one's right to another person; or a making over an estate, goods, or chattels by sale, or due course of law. *Diet.*
To ABA'ND. *v. a.* [A word contracted from abandon, but not now in use. See ABANDON.] To forsake.
Those foreigners which came from far
Grew great, and got large portions of land,
That in the realm, ere long, they stronger are
Than they which fought at first their helping hand,
And Vortiger enforced the kingdom to *aband.*
Spenser's Fairy Queen, b. ii. *cant.* 10.
To ABA'NDON. *v. a.* [Fr. *abandonner.* Derived, according to *Menage*, from the Italian *abandonare*, which signifies to forsake his colours; *bandon* [*vexillum*] *deserere. Pasquier* thinks
B
it

Figure 34: Column Space Detection

The field that is highlighted green is the spacing that is between the columns. This space will be the location where we would separate the two columns so that they could be individually processed. During the processing phase, we will separate both columns into individual files and from that point, they will be individually processed by the program. This way, we would be able to parse the data that we need from the body of text with further ease. And it will also improve the performance of the algorithm, since it would have to process less information and focus on the data that we want to extract.

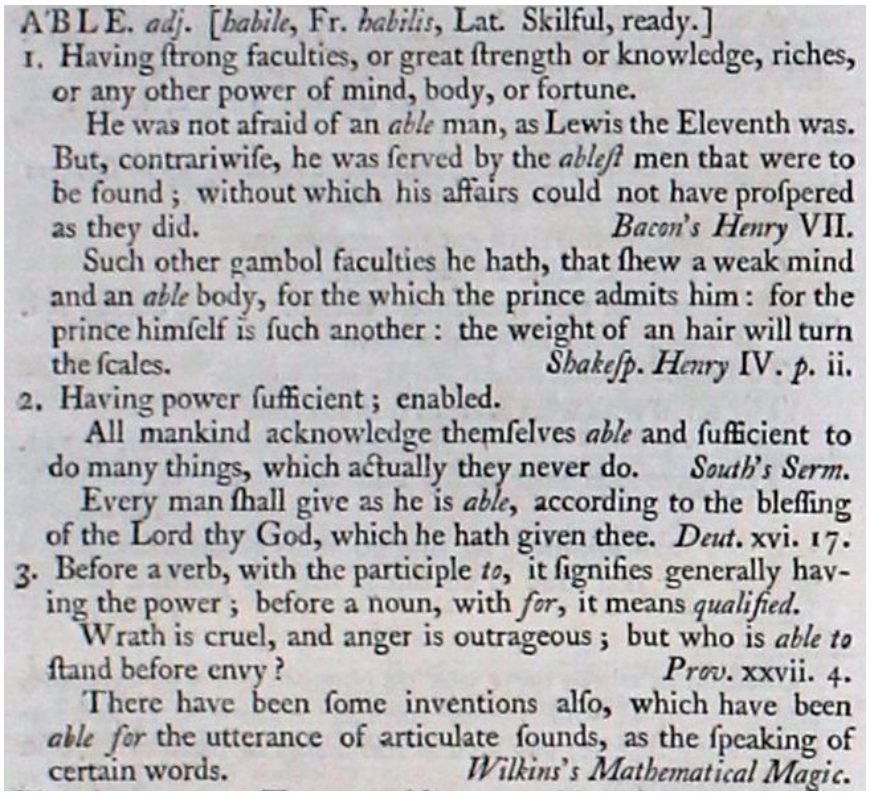
The program will then separate the columns by the space between the columns, highlighted by the green bounding box. After this, the program will split the columns into two individual files. Separating the initial files' word columns will increase the processing steps of the program, but we believe that the overall efficiency and accuracy of the parsing of the headword and its definition will increase. We believe that it will increase its efficiency and accuracy because there will be less information that the program would need to parse.

We do not want our algorithm confusing any of these for the headwords or definitions, so if possible, removing them during this phase is ideal. The pages are split into two columns that are full of text. The first step in our algorithm is to separate and crop these columns out of the page. In order to achieve this we

begin by thresholding the entire image into black and white. These will leave the text as black and turn the entire background white. This type of image is much easier to work with as it gets rid of most of the unnecessary “noise” in the image, and will only leave the text and sometimes a few blemishes to deal with. Currently, two different algorithms for thresholding are being considered, the two being Otsu’s Binarization and Canny Edge Detection.

5.3.1 Desired Output

The desired output for our program is shown below in Figure 35. This definition is properly cropped and it contains the entire definition for the word “A’BLE”.



A'BLE. *adj.* [*habile*, Fr. *habilis*, Lat. Skilful, ready.]

1. Having ſtrong faculties, or great ſtrength or knowledge, riches, or any other power of mind, body, or fortune.

He was not afraid of an *able* man, as Lewis the Eleventh was. But, contrariwiſe, he was ſerved by the *ableſt* men that were to be found; without which his affairs could not have proſpered as they did. *Bacon's Henry* VII.

Such other gambol faculties he hath, that ſhew a weak mind and an *able* body, for the which the prince admits him: for the prince himſelf is ſuch another: the weight of an hair will turn the ſcales. *Shakeſp. Henry* IV. *p.* ii.

2. Having power ſufficient; enabled.

All mankind acknowledge themſelves *able* and ſufficient to do many things, which actually they never do. *South's Serm.*

Every man ſhall give as he is *able*, according to the bleſſing of the Lord thy God, which he hath given thee. *Deut.* xvi. 17.

3. Before a verb, with the participle *to*, it ſignifies generally having the power; before a noun, with *for*, it means *qualified*.

Wrath is cruel, and anger is outrageous; but who is *able to* ſtand before envy? *Prov.* xxvii. 4.

There have been ſome inventions alſo, which have been *able for* the utterance of articulate ſounds, as the ſpeaking of certain words. *Wilkins's Mathematical Magic.*

Figure 35: Cropped Definition Example

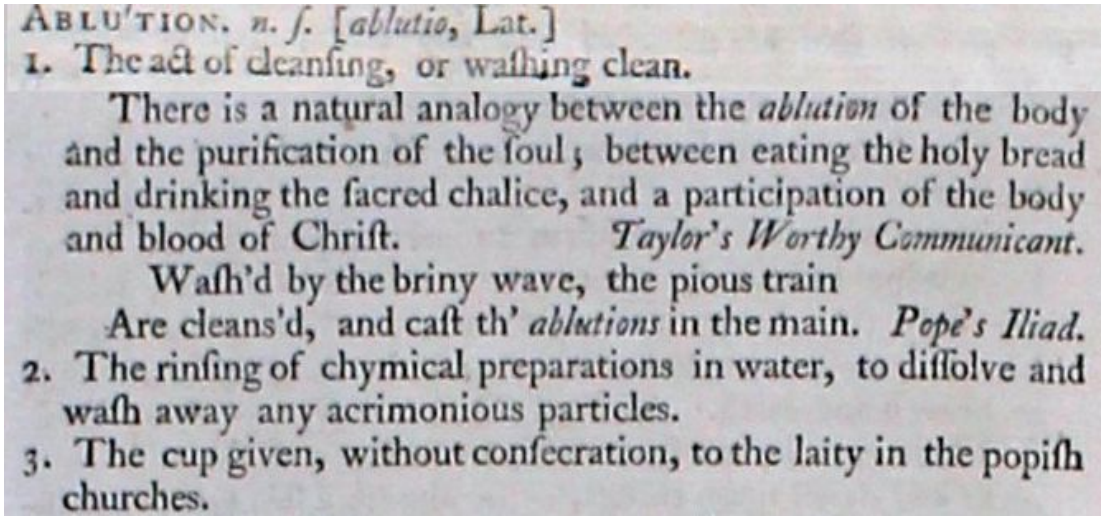

ABLU'TION. *n. ſ.* [*ablutio*, Lat.]
1. The act of cleanſing, or waſhing clean.
There is a natural analogy between the *ablution* of the body and the purification of the ſoul; between eating the holy bread and drinking the ſacred chalice, and a participation of the body and blood of Chriſt. *Taylor's Worthy Communicant.*
Waſh'd by the briny wave, the pious train
Are cleans'd, and caſt th' *ablutions* in the main. *Pope's Iliad.*
2. The rinſing of chymical preparations in water, to diſſolve and waſh away any acrimonious particles.
3. The cup given, without conſecration, to the laity in the popiſh churches.

Figure 36: Multiple Page Spanning Output

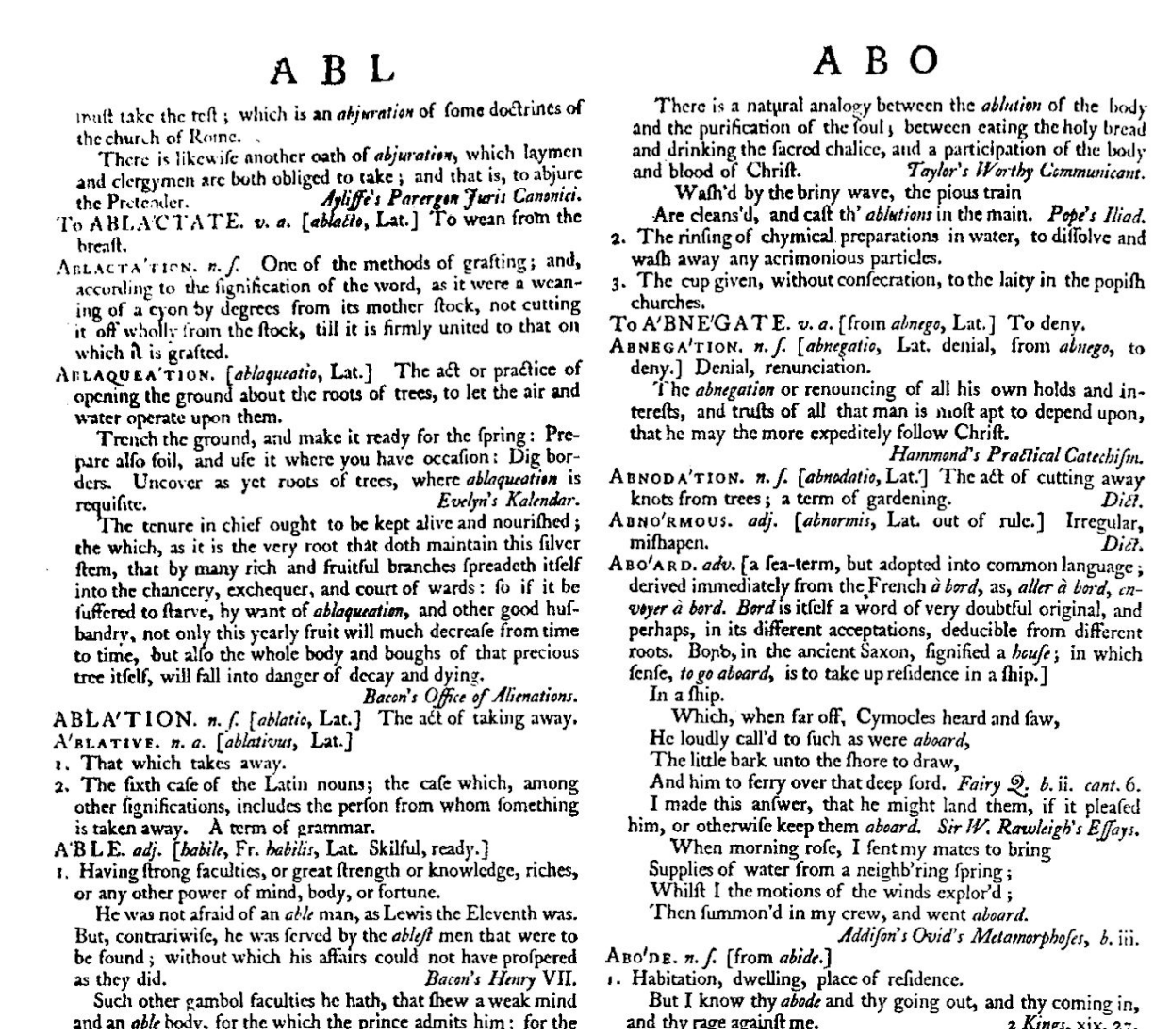
Figure 36 is an example of the proper output of a definition that spans multiple pages or columns. Every piece of the definition will be cropped separately and then stitched back together afterwards. For this particular example, color correction for the discoloration between pages is not accounted for, therefore the output is not entirely complete, but the color correction algorithm is not yet complete. After the algorithm is complete the output should look similar to the figure directly above it, and appear as one seamless definition as though it were on the same page or column.

This output also must be named properly. For example, the name of the first example's file would be as follows:

f1755-able.tif

This begins with the format, which is "f" for folio, which will essentially be the same for all of our files. Second is the year that the book was made, which is 1755. This will be one of two years, 1755 and 1775, which are the different years for the two dictionaries. After that, a hyphen is followed by the headword itself.

5.4 Otsu's Binarization



A B L

must take the teſt; which is an *abjuration* of ſome doctrines of the church of Rome.

There is likewiſe another oath of *abjuration*, which laymen and clergymen are both obliged to take; and that is, to abjure the Pretender. *Ayliffe's Parergon Juris Canonici.*

To ABLA'CTATE. *v. a.* [*ablacto*, Lat.] To wean from the breaſt.

ABLACTA'TION. *n. ſ.* One of the methods of grafting; and, according to the ſignification of the word, as it were a weaning of a cyon by degrees from its mother ſtock, not cutting it off wholly from the ſtock, till it is firmly united to that on which it is grafted.

ABLAQUEA'TION. [*ablaqueatio*, Lat.] The act or practice of opening the ground about the roots of trees, to let the air and water operate upon them.

Trench the ground, and make it ready for the ſpring: Prepare alſo ſoil, and uſe it where you have occaſion: Dig borders. Uncover as yet roots of trees, where *ablaqueation* is requiſite. *Evelyn's Kalendar.*

The tenure in chief ought to be kept alive and nouriſhed; the which, as it is the very root that doth maintain this ſilver ſtem, that by many rich and fruitful branches ſpreadeth itſelf into the chancery, exchequer, and court of wards: ſo if it be ſuffered to ſtarve, by want of *ablaqueation*, and other good huſbandry, not only this yearly fruit will much decreaſe from time to time, but alſo the whole body and boughs of that precious tree itſelf, will fall into danger of decay and dying. *Bacon's Office of Alienations.*

ABLA'TION. *n. ſ.* [*ablatio*, Lat.] The act of taking away.

A'BLATIVE. *n. a.* [*ablativus*, Lat.]

1. That which takes away.
2. The ſixth caſe of the Latin nouns; the caſe which, among other ſignifications, includes the perſon from whom ſomething is taken away. A term of grammar.

A'BLE. *adj.* [*habile*, Fr. *habilis*, Lat. Skilful, ready.]

1. Having ſtrong faculties, or great ſtrength or knowledge, riches, or any other power of mind, body, or fortune.

He was not afraid of an *able* man, as Lewis the Eleventh was. But, contrariwiſe, he was ſerved by the *ableſt* men that were to be found; without which his affairs could not have proſpered as they did. *Bacon's Henry VII.*

Such other gambol faculties he hath, that ſhew a weak mind and an *able* body, for the which the prince admits him: for the

A B O

There is a natural analogy between the *ablution* of the body and the purification of the ſoul; between eating the holy bread and drinking the ſacred chalice, and a participation of the body and blood of Chriſt. *Taylor's Worthy Communicant.*

Waſh'd by the briny wave, the pious train
Are cleans'd, and caſt th' *ablutions* in the main. *Pope's Iliad.*

2. The rinſing of chymical preparations in water, to diſſolve and waſh away any acrimonious particles.
3. The cup given, without conſecration, to the laity in the popiſh churches.

To A'BNE'GATE. *v. a.* [from *abnego*, Lat.] To deny.

ABNEGA'TION. *n. ſ.* [*abnegatio*, Lat. denial, from *abnego*, to deny.] Denial, renunciation.

The *abnegation* or renouncing of all his own holds and intereſts, and truſts of all that man is moſt apt to depend upon, that he may the more expeditely follow Chriſt. *Hammond's Practical Catechiſm.*

ABNODA'TION. *n. ſ.* [*abnodatio*, Lat.] The act of cutting away knots from trees; a term of gardening. *Dict.*

ABNO'RMOUS. *adj.* [*abnormis*, Lat. out of rule.] Irregular, miſhapen. *Dict.*

ABO'ARD. *adv.* [a ſea-term, but adopted into common language; derived immediately from the French *à bord*, as, *aller à bord*, *envoyer à bord*. *Bord* is itſelf a word of very doubtful original, and perhaps, in its different acceptations, deducible from different roots. Bord, in the ancient Saxon, ſignified a *houſe*; in which ſenſe, *to go aboard*, is to take up reſidence in a ſhip.]

In a ſhip.

Which, when far off, Cymocles heard and ſaw,
He loudly call'd to ſuch as were *aboard*,
The little bark unto the ſhore to draw,
And him to ferry over that deep ford. *Fairy Q. b.* ii. *cant.* 6.

I made this anſwer, that he might land them, if it pleaſed him, or otherwiſe keep them *aboard*. *Sir W. Rawleigh's Eſſays.*

When morning roſe, I ſent my mates to bring
Supplies of water from a neighb'ring ſpring;
Whilſt I the motions of the winds explor'd;
Then ſummon'd in my crew, and went *aboard*.
Addiſon's Ovid's Metamorphoſes, b. iii.

ABO'DE. *n. ſ.* [from *abide*.]

1. Habitation, dwelling, place of reſidence.

But I know thy *abode* and thy going out, and thy coming in, and thy rage againſt me. *2 Kings*, xix. 27.

Figure 37: Thresholding or Otsu's Binarization

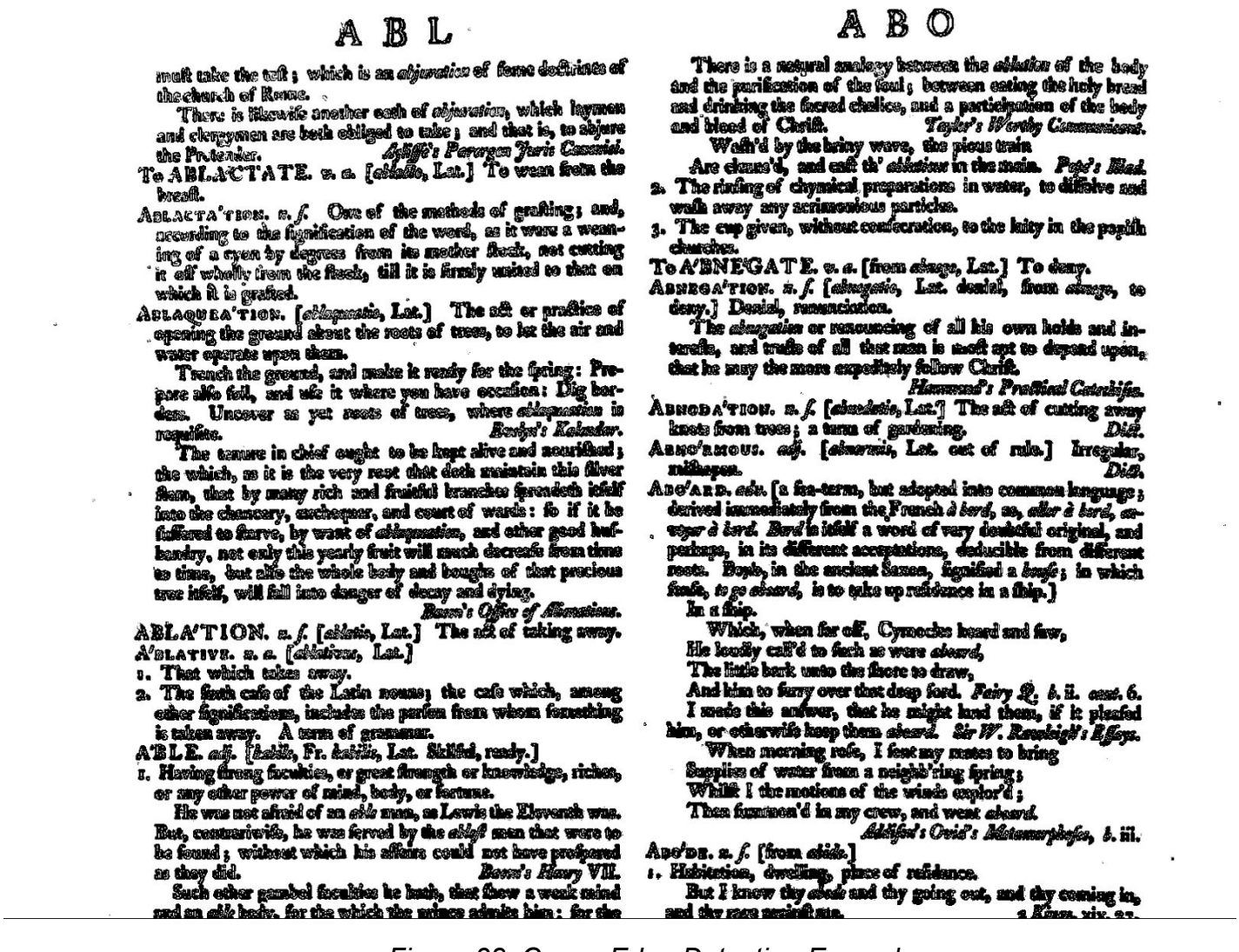
To perform the thresholding, we used a python algorithm called Otsu Thresholding, or Otsu's Binarization. This makes the thresholding easy, as for with normal thresholding you have to find the value that gets thresholded out of the picture and turned into the background. However, if there is a shadow on the image then this value will be more difficult to find and you may end up with an image that is half black due to the shadow being recognized as darker than the background, and therefore recognized as text to your algorithm. Otsu's

Binarization finds the optimal threshold value for your image, and gives it back to you for you to use in your thresholding functions. The algorithm works by attempting to find a threshold value that reduces the amount of weighted within-class variance to an absolute minimum. This all but ensures that every page will have a proper thresholded version, and we will not have any issues with pages being entirely black or entirely white, or any text being cut out and ruining our crop. For the most part, if some word was removed in the middle of the definition it would not matter. This is because we are only using the thresholded image to calculate exactly what we need to perform on the normal image. The thresholded image is much more convenient for figuring out where exactly the columns start and stop, and this allows us to figure this out on the thresholded image and perform the exact same operations on the normal image. The only issue with this would be any blemishes that may be on the page, such as a coffee stain, or any sort or black speck that was on the page when the scan was performed. This would be taken as a piece of text, and if it is dark enough then Otsu's Binarization would not get rid of it. We will use a Gaussian filter to remove any additional noise or specks that are on the image, this will blur the image slightly and the darker specks will get blended in with the lighter background around it. If the darker specks are on the text it does not matter, as the text will already come through as black after applying Otsu's Binarization.

To summarize Otsu's Binarization:

1. The algorithm works by creating a histogram of the input image and finding the optimal thresholding value from the pixels. Which can range from 0-255. The input that Otsu's algorithm is expecting is in black-and-white. If the incoming image is not so, then it will be converted into a grayscale before it is subsequently given as an input.
2. Once the thresholds are established, they will be separated into those categories. This is where the foreground and background are determined and classified accordingly.
3. The variance between the two categories is determined and from this data, the text can be isolated.

5.5 Canny Edge Detection



A B L

must take the oath; which is an *abjuration* of some doctrines of the church of Rome.

There is likewise another oath of *abjuration*, which laymen and clergymen are both obliged to take; and that is, to abjure the Pretender. *Ayliffe's Parergon Juris Canonici.*

To ABLACTATE. *v. a.* [*ablacto*, Lat.] To wean from the breast.

ABLACTA'TION. *n. f.* One of the methods of grafting; and, according to the signification of the word, as it were a weaning of a cyon by degrees from its mother stock, not cutting it off wholly from the stock, till it is firmly united to that on which it is grafted.

ABLAQUEA'TION. [*ablaqueatio*, Lat.] The act or practice of opening the ground about the roots of trees, to let the air and water operate upon them.

Trench the ground, and make it ready for the spring: Prepare also soil, and stir it where you have occasion: Dig borders. Uncover as yet roots of trees, where *ablaquation* is requisite. *Evelyn's Kalendar.*

The tenure in chief ought to be kept alive and nourished; the which, as it is the very root that doth maintain this silver stem, that by many rich and fruitful branches spreadeth itself into the chancery, exchequer, and court of wards: so if it be suffered to starve, by want of *ablaquation*, and other good husbandry, not only this yearly fruit will much decrease from time to time, but also the whole body and boughs of that precious tree itself, will fall into danger of decay and dying. *Bacon's Office of Alienations.*

ABLA'TION. *n. f.* [*ablatio*, Lat.] The act of taking away.

A'BLATIVE. *n. s.* [*ablativus*, Lat.]

1. That which takes away.
2. The sixth case of the Latin nouns; the case which, among other significations, includes the person from whom something is taken away. A term of grammar.

A'BLE. *adj.* [*habile*, Fr. *habilis*, Lat. Skilful, ready.]

1. Having strong faculties, or great strength or knowledge, riches, or any other power of mind, body, or fortune.

He was not afraid of an *able* man, as Lewis the Eleventh was. But, contrariwise, he was served by the *ablest* men that were to be found; without which his affairs could not have prospered as they did. *Bacon's Henry VII.*

Such other gambol faculties he hath, that shew a weak mind and an *able* body, for the which the prince admits him: for the

A B O

There is a natural analogy between the *ablution* of the body and the purification of the soul; between eating the holy bread and drinking the sacred chalice, and a participation of the body and blood of Christ. *Taylor's Worthy Communicant.*

Wash'd by the briny wave, the pious train
Are cleans'd, and cast th' *ablution* in the main. *Pope's Iliad.*

2. The rinsing of chymical preparations in water, to dissolve and wash away any acrimonious particles.
3. The cup given, without consecration, to the laity in the popish churches.

To ABNE'GATE. *v. a.* [from *abnego*, Lat.] To deny.

ABNEGA'TION. *n. f.* [*abnegatio*, Lat. denial, from *abnego*, to deny.] Denial, renunciation.

The *abnegation* or renouncing of all his own holds and interests, and trusts of all that man is most apt to depend upon, that he may the more expeditely follow Christ. *Hammond's Practical Catechism.*

ABNODA'TION. *n. f.* [*abnodatio*, Lat.] The act of cutting away knots from trees; a term of gardening. *Dict.*

ABNO'RMOUS. *adj.* [*abnormis*, Lat. out of rule.] Irregular, misshapen. *Dict.*

ABO'ARD. *adv.* [a sea-term, but adopted into common language; derived immediately from the French *à bord*, as, *aller à bord, envoyer à bord. Bord* is itself a word of very doubtful original, and perhaps, in its different acceptations, deducible from different roots. Boyle, in the ancient Saxon, signified a *boyle*; in which sense, *to go aboard*, is to take up residence in a ship.]

In a ship.

Which, when far off, Cymocles heard and saw,
He loudly call'd to such as were *aboard*,
The little bark unto the shore to draw,
And him to ferry over that deep ford. *Fairy Q.* b. ii. cant. 6.

I made this answer, that he might hand them, if it pleased him, or otherwise keep them *aboard*. *Sir W. Raleigh's Essays.*

When morning rose, I sent my mates to bring
Supplies of water from a neighb'ring spring;
Whilst I the motions of the winds explor'd;
Then summon'd in my crew, and went *aboard*.
Addison's Ovid's Metamorphosis, b. iii.

ABO'DE. *n. f.* [from *abide*.]

1. Habitation, dwelling, place of residence.

But I know thy *abode* and thy going out, and thy coming in, and thy rage against me. 2 *Kings*, xix. 27.

Figure 38: Canny Edge Detection Example

Additionally, we may also use Canny Edge Detection as an alternative for thresholding, in case we run into any problems on a significant amount of pages through the thresholding process. Canny Edge Detection, as the name implies, is an algorithm to detect and emphasize edges of objects or text in an image. This will isolate the text from the background, just like Otsu's Binarization, except it will emphasize only the outline of the text and not the text itself. This is not an issue because we only need to know the areas that the text is encapsulating within the image, so any way of finding that area will work for our project. Similar

to before, we will use a Gaussian filter because Canny Edge Detection is very vulnerable when there is any sort of noise in the image.

The blurred or smoothened image after the Gaussian filter is then filtered again with a Sobel kernel. The Sobel kernel, or Sobel filter, is used to emphasize the edges within an image. A suppression scan is then performed on the entire image in order to remove any extra specks or noise that were not filtered out by the Gaussian filter. It does this by comparing every dark point with the edges that have been isolated by the Sobel filter. If the point is along an edge that was already isolated, it is kept. The point is first compared to the edge itself to determine whether it lies along it, and then it is compared to points normal (or parallel) to the edge. If the point forms a local maximum and is darker than the points along the normal that it is being compared to, then the algorithm determines that the point is along the edge. The points that are not along edges are then suppressed, and changed to be part of the background. Determining whether edges should stay considered as an edge is different. If an edge has passed the previous stage and the points along it are all considered to be a part of the edge, then it is compared to two different threshold values. These values are the maximum value and the minimum value. If an edge with an intensity gradient is above the maximum value, the edge is considered to be an edge. If the edge has an intensity gradient below the minimum value, the edge is no longer considered to be an edge, and is suppressed and changed to be a part of the background. However, there is one additional case, edges that are in between the maximum value and the minimum value. When an edge is above the maximum value, the algorithm is certain that it is an edge, therefore if an edge is partially above the maximum and has pieces that dip in between the maximum and minimum, it is still considered an edge. If every part of an edge is in between the maximum and minimum threshold values, then the edge is not considered an edge and is then suppressed.

To summarize the Canny Edge Detection algorithm, a step by step guide follows:

1. The first stage of the algorithm will reduce the amount of noise in the image by applying a 5 x 5 Gaussian filter.

2. The second stage of the algorithm will find the intensity gradient of the image. It will do this by attaining the Sobel Derivatives, an approximation of the derivative, then multiplying by a set of matrices, like the two illustrated below producing two separate images.

$$G_x = \begin{bmatrix} -3 & 0 & +3 \\ -10 & 0 & +10 \\ -3 & 0 & +3 \end{bmatrix}$$

$$G_y = \begin{bmatrix} -3 & -10 & -3 \\ 0 & 0 & 0 \\ +3 & +10 & +3 \end{bmatrix}$$

(Scharr Function will improve the approximation of the Sobel Function)

After the images are generated in the X and Y directions, the edge gradient is calculated as follows:

$$Edge\_Gradient\ (G) = \sqrt{G_x^2 + G_y^2}$$

$$Angle\ (\theta) = \tan^{-1}\frac{G_y^2}{G_x^2}$$

3. After the previous step is completed, the image is scanned, in its entirety, and any pixel that is an outlier in its immediate vicinity will be set to off. In a process called Non-maximum Suppression.

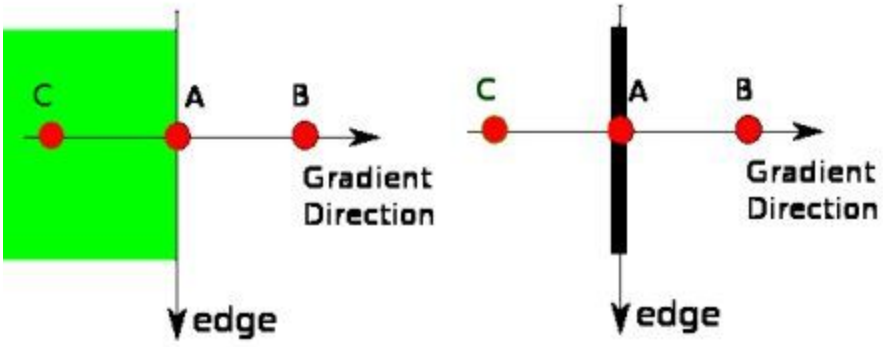


Figure 39: Pixel is A is edge and it is compared with gradient directions C and B

4. In the final step of the algorithm, two values will be generated. One of the values will be a maximum value and the other will be a minimum value. If the intensity gradient is lower than the minimum value, it will be discarded, as it is not an edge. If the value is between the maximum value and the minimum value, it will be further considered. Consequently, if the intensity gradient that was calculated in the previous step of the algorithm, is higher or at the maximum, it is sure to be an edge and will therefore be included.

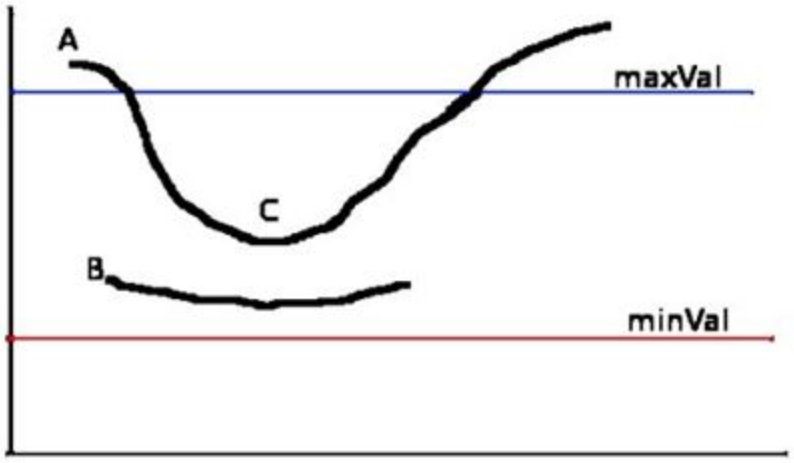


Figure 40: Maximum and minimum values

The good thing about Canny Edge Detection is that OpenCV contains every part of the algorithm within one function, cv2.Canny(). The arguments it takes are the image that is being processed with the algorithm, the minimum

threshold value for the final step, the maximum threshold value for the final step, the aperture size, which is the size of the Sobel kernel, and finally L2Gradient, which can change the equation that finds the gradient magnitude. This should always be set to true however, because the equation that it will use is more accurate. The resulting image will be white outlines on a black background, and the inside of the text will not be filled in, only the outlines of the text will be emphasized.

The following code written in Python is demonstrating the application of Canny Edge Detection on an “after1.tiff”. In addition, this Python code will produce two outputs, the first output will be the unmodified “after1.tiff” file and the modified “edge.tiff” file.

On Line 6, the function call cv.Canny() takes 3 parameters. The parameters required for this function are the following: the image file, the minimum value, and the maximum value.

```
1  import numpy as np
2  import cv2 as cv
3  from matplotlib import pyplot as plt
4
5  image = cv.imread('after1.tiff', 0)
6  edges = cv.Canny(image, 100, 200)
7
8
9  plt.subplot(121), plt.imshow(image , cmap = 'gray')
10 plt.title('Original Image'), plt.xticks([]), plt.yticks([])
11
12
13 plt.subplot(122), plt.imshow(edges, cmap = 'gray')
14 plt.title('Edge Image'), plt.xticks([]), plt.yticks([])
15
16
17 plt.savefig('edge.tiff')
```

Figure 41: Python code illustrating the application of Canny Edge Detection algorithm to an image file

After running the above Python code, the following output is obtained:

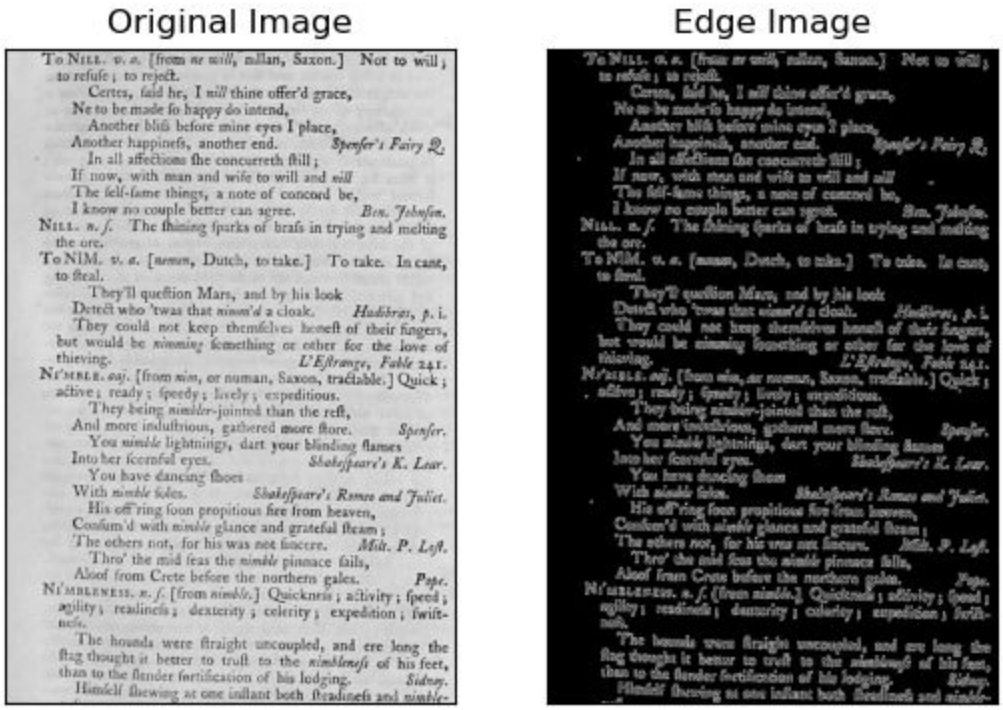


Figure 42: Output after applying Canny Edge Detection algorithm on a .tiff file

As you can see, the image on the left is the original .tiff file that was entered into the program and the resulting image is on the right. A benefit of applying the Canny Edge Detection algorithm is that it reduces the overall noise of the image, in addition to clearly differentiating the data with the background. Another benefit of applying this algorithm is that This will prepare the image file for the next phase of processing.

5.6 Dilation

In this next phase of processing, a morphological operation called Dilation will be performed on the image. This process will further remove noise from the image and cause the elements of interest to join, into a clump, and it will simplify the next phase of processing.

Four iterations of the Python code below yielded desirable results.

```
31  for x in range(4):
32      image = cv.dilate(edges, kernel, iterations = (x + 1))
33
34      plt.imshow(image, cmap = 'gray')
35
36      plt.savefig(str(x + 1) + "_edge.tiff", bbox_inches = 'tight', pad_inches = 0)
```

Figure 43: Python code illustrating the application of morphological dilation algorithm to an image file

On line 32, the dilate function is called with three parameters. The first parameter is called “edges”. This variable contains an image file that was processed by the Canny Edge Detection algorithm. The second parameter is called “kernel”. This parameter is storing a five by five matrix of ones. The last parameter of this function call is setting the amount of iterations that the dilation function will perform on the image. In this case, the contents of the variable “edges”.

On line 36, a variant of the final implementation of the naming convention can be observed. In this variant, the current iteration of the loop has been appended to the name of the output file in the following format: ITERATION + “_edge.tiff”. This is testing the mechanic that will likely be in the final implementation of the naming convention mechanics.

After observing several output images, I noticed that the best resolution is 1380 x 2000. resolutions that are bigger than this causes the prototype program to take longer to process the data. This could be because the program must process and produce larger files, and therefore, increases the programs resource usage and processing time. This, of course, is subject to change as the program is further refined and developed.

After running the Python code above, it yielded the following favorable results,

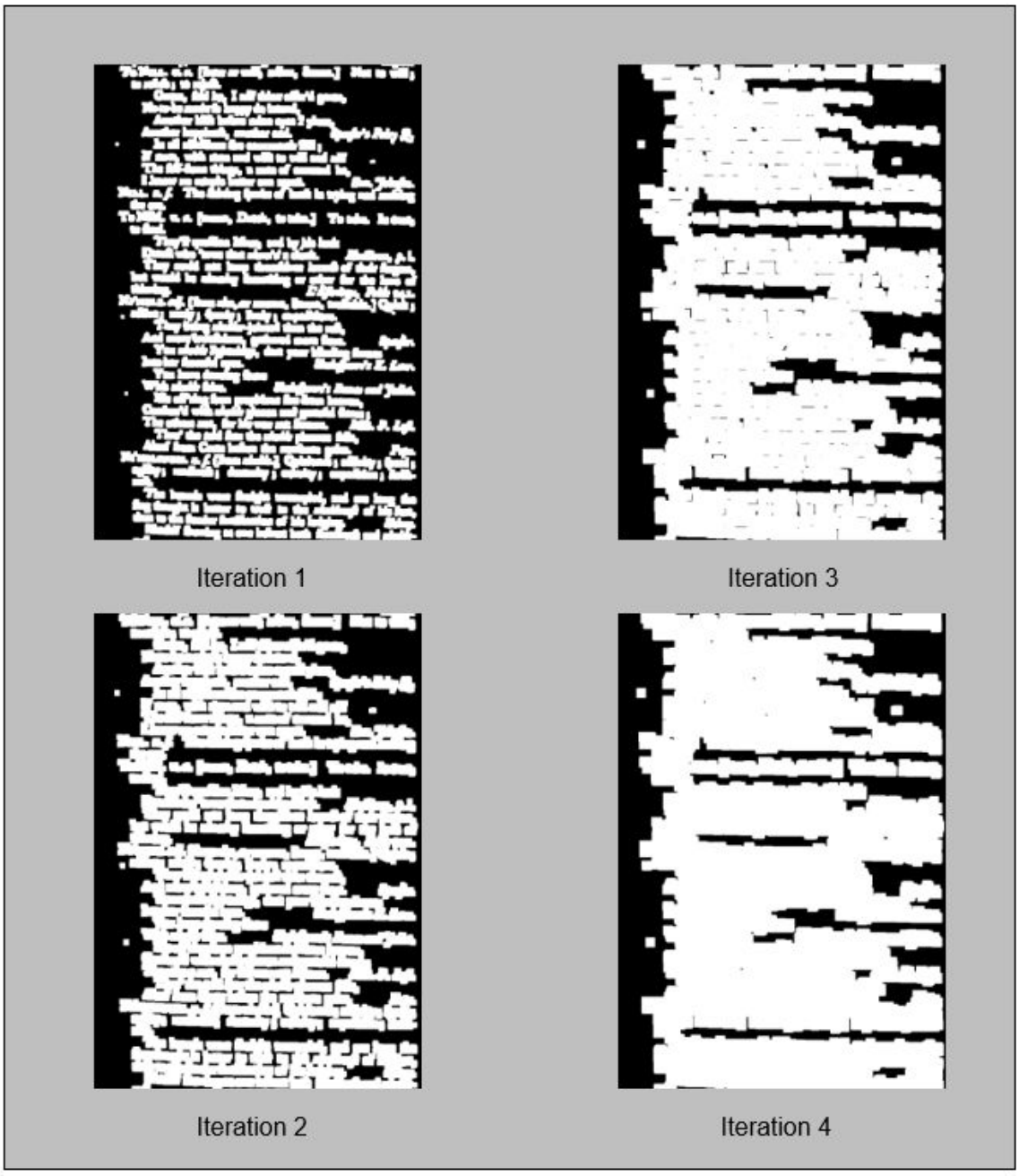


Figure 44: Four iterations of dilation applied to a dictionary entry

As you can see from the images above, the text has been dilated into bigger masses. As a result, the relevant data has been isolated in white. And the background, colored in black, has also been isolated and would later be removed from the image.

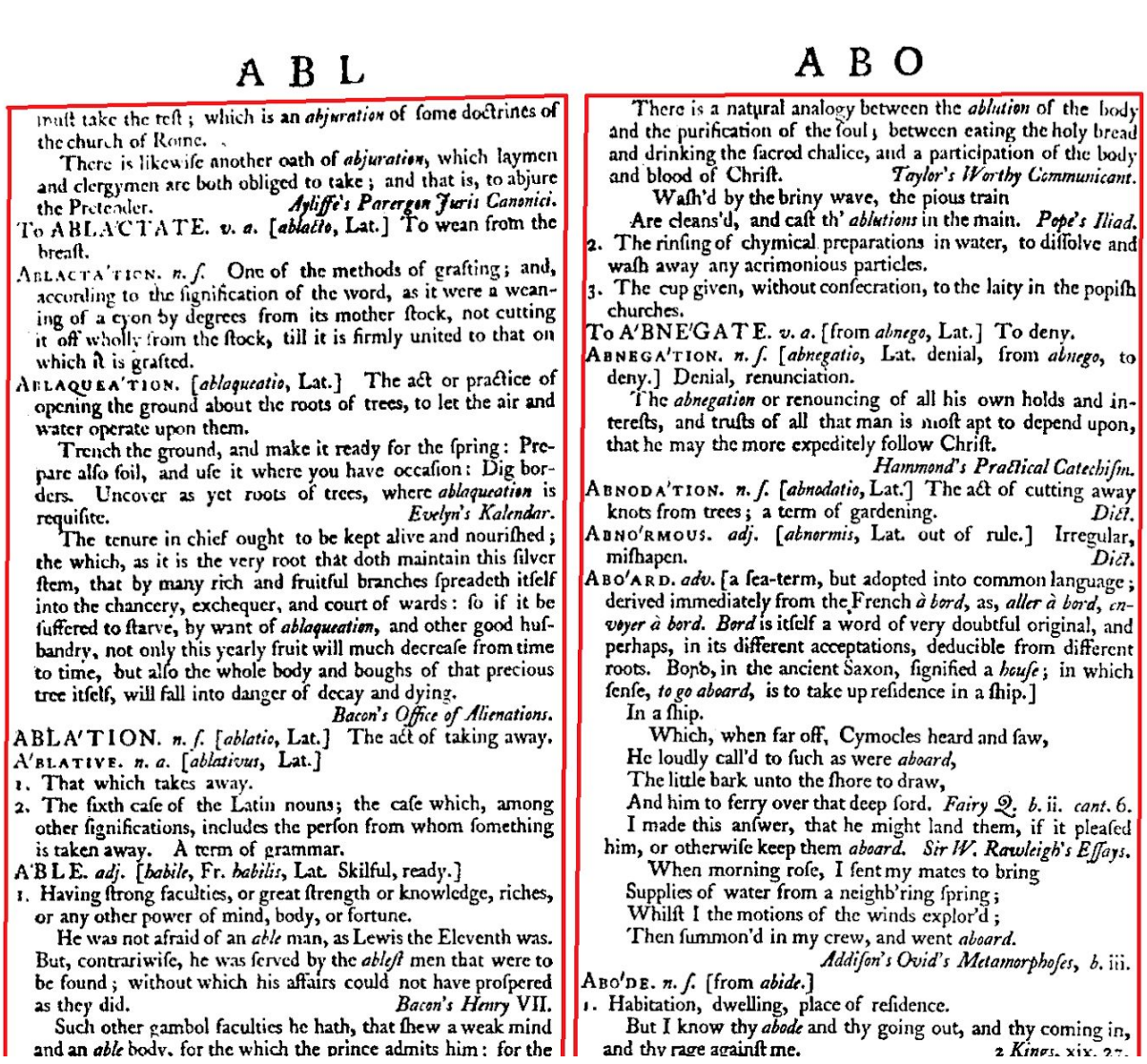
Note that there are still areas that have been isolated in white that are not areas of interest but were isolated in error. These regions are highlighted in Red below,

Figure 45: Regions of errant text detection to be discarded

Note that the areas highlighted in red are areas that were accentuated in error while being processed. These areas will later be removed in later stages of the processing pipeline.

5.7 Binarized Column Cropping

Once the image is entirely black and white, we will look at the columns of pixels across the image to find where there is a lot of empty whitespace and a lack of text. This can be done by averaging the pixels in each column and finding the columns with a higher (more white) average. The example image has not yet been dewarped and rotated, so the example columns are at an angle. For the actual column cropping process, the page will be dewarped and straightened to make processing the cropping much easier and significantly more efficient.



A B L

must take the teſt; which is an *abjuration* of ſome doctrines of the church of Rome.

There is likewiſe another oath of *abjuration*, which laymen and clergymen are both obliged to take; and that is, to abjure the Pretender. *Ayliffe's Parergon Juris Canonici.*

To ABLACTATE. *v. a.* [*ablacto*, Lat.] To wean from the breaſt.

ABLACTA'TION. *n. ſ.* One of the methods of grafting; and, according to the ſignification of the word, as it were a weaning of a cyon by degrees from its mother ſtock, not cutting it off wholly from the ſtock, till it is firmly united to that on which it is grafted.

ABLAQUEA'TION. [*ablaqueatio*, Lat.] The act or practice of opening the ground about the roots of trees, to let the air and water operate upon them.

Trench the ground, and make it ready for the ſpring: Prepare alſo ſoil, and uſe it where you have occaſion: Dig borders. Uncover as yet roots of trees, where *ablaqueation* is requiſite. *Evelyn's Kalendar.*

The tenure in chief ought to be kept alive and nouriſhed; the which, as it is the very root that doth maintain this ſilver ſtem, that by many rich and fruitful branches ſpreadeth itſelf into the chancery, exchequer, and court of wards: ſo if it be ſuffered to ſtarve, by want of *ablaqueation*, and other good huſbandry, not only this yearly fruit will much decreaſe from time to time, but alſo the whole body and boughs of that precious tree itſelf, will fall into danger of decay and dying. *Bacon's Office of Alienations.*

ABLA'TION. *n. ſ.* [*ablatio*, Lat.] The act of taking away.

A'BLATIVE. *n. a.* [*ablativus*, Lat.]

1. That which takes away.
2. The ſixth caſe of the Latin nouns; the caſe which, among other ſignifications, includes the perſon from whom ſomething is taken away. A term of grammar.

A'BLE. *adj.* [*habile*, Fr. *habilis*, Lat. Skilful, ready.]

1. Having ſtrong faculties, or great ſtrength or knowledge, riches, or any other power of mind, body, or fortune.

He was not afraid of an *able* man, as Lewis the Eleventh was. But, contrariwiſe, he was ſerved by the *ableſt* men that were to be found; without which his affairs could not have proſpered as they did. *Bacon's Henry VII.*

Such other gambol faculties he hath, that ſhew a weak mind and an *able* body, for the which the prince admits him: for the

A B O

There is a natural analogy between the *ablution* of the body and the purification of the ſoul; between eating the holy bread and drinking the ſacred chalice, and a participation of the body and blood of Chriſt. *Taylor's Worthy Communicant.*

Waſh'd by the briny wave, the pious train
Are cleans'd, and caſt th' *ablutions* in the main. *Pope's Iliad.*

2. The rinſing of chymical preparations in water, to diſſolve and waſh away any acrimonious particles.
3. The cup given, without conſecration, to the laity in the popiſh churches.

To A'BNE'GATE. *v. a.* [from *abnego*, Lat.] To deny.

ABNEGA'TION. *n. ſ.* [*abnegatio*, Lat. denial, from *abnego*, to deny.] Denial, renunciation.

The *abnegation* or renouncing of all his own holds and intereſts, and truſts of all that man is moſt apt to depend upon, that he may the more expeditely follow Chriſt. *Hammond's Practical Catechiſm.*

ABNODA'TION. *n. ſ.* [*abnodatio*, Lat.] The act of cutting away knots from trees; a term of gardening. *Dict.*

ABNO'RMOUS. *adj.* [*abnormis*, Lat. out of rule.] Irregular, miſhapen. *Dict.*

ABO'ARD. *adv.* [a ſea-term, but adopted into common language; derived immediately from the French *à bord*, as, *aller à bord*, *envoyer à bord*. *Bord* is itſelf a word of very doubtful original, and perhaps, in its different acceptations, deducible from different roots. Boꝛd, in the ancient Saxon, ſignified a *houſe*; in which ſenſe, *to go aboard*, is to take up reſidence in a ſhip.]

In a ſhip.

Which, when far off, Cymocles heard and ſaw,
He loudly call'd to ſuch as were *aboard*,
The little bark unto the ſhore to draw,
And him to ferry over that deep ford. *Fairy Q. b.* ii. *cant.* 6.

I made this anſwer, that he might land them, if it pleaſed him, or otherwiſe keep them *aboard*. *Sir W. Rawleigh's Eſſays.*

When morning roſe, I ſent my mates to bring
Supplies of water from a neighb'ring ſpring;
Whilſt I the motions of the winds explor'd;
Then ſummon'd in my crew, and went *aboard*.
Addiſon's Ovid's Metamorphoſes, b. iii.

ABO'DE. *n. ſ.* [from *abide*.]

1. Habitation, dwelling, place of reſidence.

But I know thy *abode* and thy going out, and thy coming in, and thy rage againſt me. 2 *Kings*, xix. 27.

Figure 46: Column Cropping Example

Due to the way the pages are laid out, it is a lot easier to surround the columns on the left and right side than the top and bottom. This is because the unneeded text that was mentioned earlier, (guide words, signature marks, press figures, and catchwords), is located on the top and bottom of the page, generally very close to the columns, and sometimes even attached to them. Luckily, the guide words at the top of the page are separated from the text and written in a much larger font, so those will be very easy to pull out. The problem lies in the catchwords sometimes directly touching the definitions. Seeing as we are not yet cropping out the definitions themselves, but simply the columns of text, it is better to ignore this fault for now and crop the columns with the catchwords and other additional text on the bottom of the page. Once the columns are cropped out, we want to make sure we remember which columns were on the left or right side of the page. This is both for our data and because the catchwords are only on the right column. We know that the only time we need to look for a catchword is if we have pulled out the right column, and if we try to remove the catchword on the left column we may remove text that is necessary for the definition. Since we have dewarped and rotated the pages already, the columns should be relatively straight, so we should not need to dewarp or rotate again. Depending on our findings however, we may want to dewarp and rotate during this stage, maybe multiple times, as the amount of text that the dewarped has to deal with is lower and there will be less room for errors and likely a greater output file. Once the columns are cropped out, we will move on to cropping out the definitions themselves.

5.8 Definition Detection and Cropping

So to summarize the previous sections, we have both made sure that the text and columns are straightened, and we have cropped out the columns to remove whitespace and the guide words. The next step is to figure out how to crop out the definitions and headwords. In order to detect where the headwords are and how long the definition goes on for, we must find how the text differs when the headwords begin and from headword to definition.

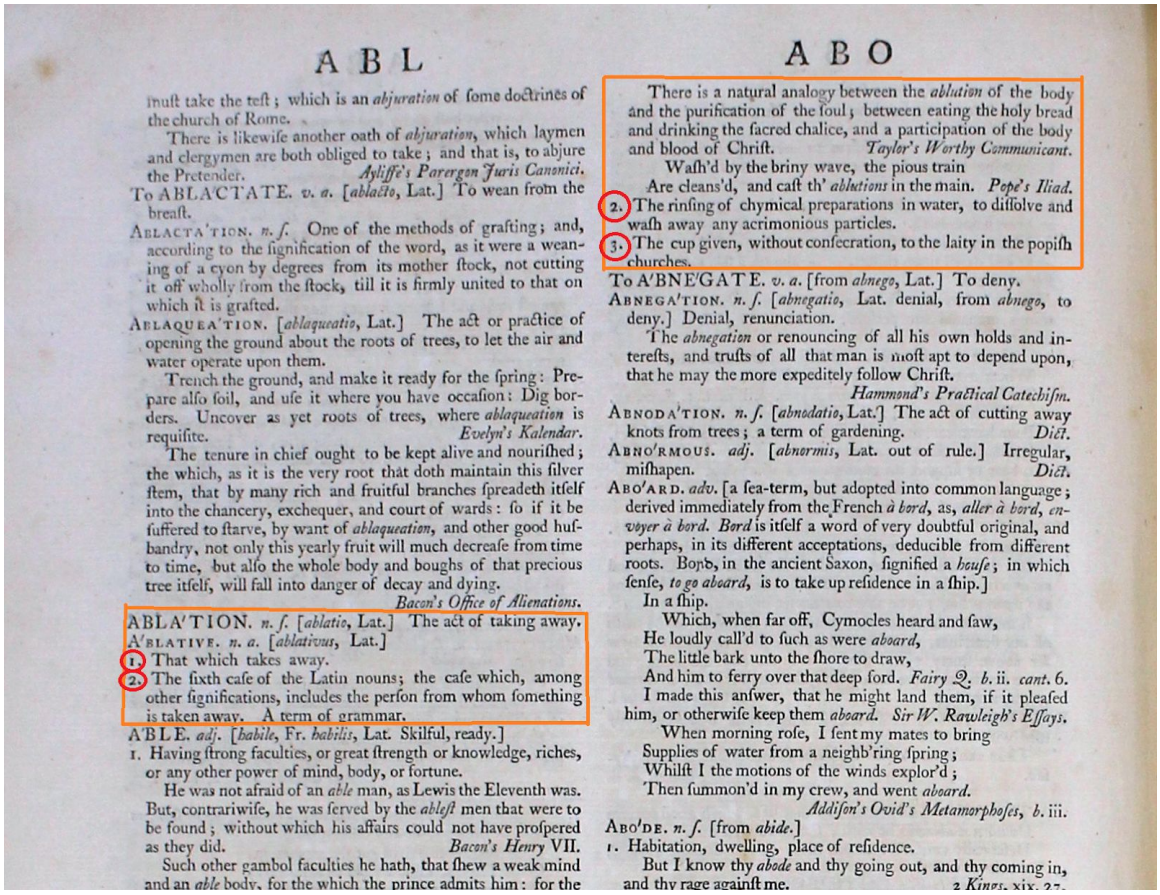
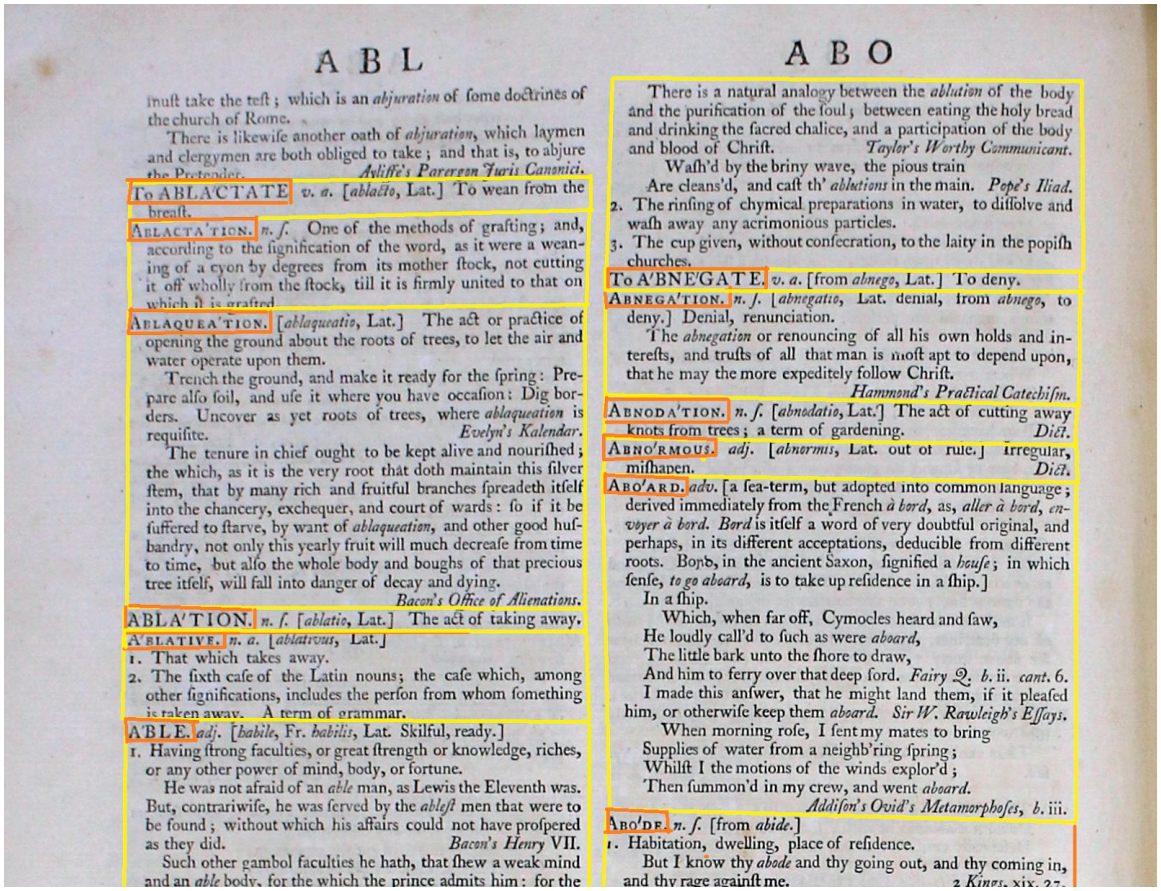


Figure 47: Multiple Definition Example

Figure 47 directly shows two separate definitions that have multiple numbered definitions. Note that the definitions on the left contain the headword and every subsequent definition for the word, while the definition on the right is an example of a definition that spans multiple pages. Only approximately half of the page is being shown, so the definition begins somewhere on the bottom left of this page, and continues through to the top right of the page. For some words, the definitions span multiple pages like this and can have over 30 different

numbered definitions. This shows how important both page stitching and our definition detection needs to be, as we do not want a headword to lose any part of the definition.

Now we need to know how exactly we want the definitions to be cropped, and this is shown in figure 48 below. The headwords are boxed in orange and the definitions are boxed in yellow. If you look at the word “A’BLATIVE.” on the left side of the page, you will see that headwords with multiple definitions will have them all boxed together and included in the crop for that word.



A B L

must take the teſt; which is an *abjuration* of ſome doctrines of the church of Rome.

There is likewiſe another oath of *abjuration*, which laymen and clergymen are both obliged to take; and that is, to abjure the Pretender. *Ayliffe's Parergon Juris Canonici.*

To ABLACTATE *v. a.* [*ablactо*, Lat.] To wean from the breaſt.

ABLACTA'TION. *n. ſ.* One of the methods of graſting; and, according to the ſignification of the word, as it were a weaning of a cyon by degrees from its mother ſtock, not cutting it off wholly from the ſtock, till it is firmly united to that on which it is grafted.

ABLAQUEA'TION. [*ablaqueatio*, Lat.] The act or practice of opening the ground about the roots of trees, to let the air and water operate upon them.

Trench the ground, and make it ready for the ſpring: Prepare alſo ſoil, and uſe it where you have occaſion: Dig borders. Uncover as yet roots of trees, where *ablaqueation* is requiſite. *Evelyn's Kalendar.*

The tenure in chief ought to be kept alive and nouriſhed; the which, as it is the very root that doth maintain this ſilver ſtem, that by many rich and fruitful branches ſpreadeth itſelf into the chancery, exchequer, and court of wards: ſo if it be ſuffered to ſtarve, by want of *ablaqueation*, and other good huſbandry, not only this yearly fruit will much decreaſe from time to time, but alſo the whole body and boughs of that precious tree itſelf, will fall into danger of decay and dying. *Bacon's Office of Alienations.*

ABLA'TION. *n. ſ.* [*ablatio*, Lat.] The act of taking away.

A'BLATIVE. *n. a.* [*ablativus*, Lat.]

1. That which takes away.
2. The ſixth caſe of the Latin nouns; the caſe which, among other ſignifications, includes the perſon from whom ſomething is taken away. A term of grammar.

A'BLE. *adj.* [*habile*, Fr. *habilis*, Lat. Skilful, ready.]

1. Having ſtrong faculties, or great ſtrength or knowledge, riches, or any other power of mind, body, or fortune.

He was not afraid of an *able* man, as Lewis the Eleventh was. But, contrariwiſe, he was ſerved by the *ableſt* men that were to be found; without which his affairs could not have proſpered as they did. *Bacon's Henry* VII.

Such other gambol faculties he hath, that ſhew a weak mind and an *able* body, for the which the prince admits him: for the

A B O

There is a natural analogy between the *ablution* of the body and the purification of the ſoul; between eating the holy bread and drinking the ſacred chalice, and a participation of the body and blood of Chriſt. *Taylor's Worthy Communicant.*

Waſh'd by the briny wave, the pious train
Are cleans'd, and caſt th' *ablutions* in the main. *Pope's Iliad.*

2. The rinſing of chymical preparations in water, to diſſolve and waſh away any acrimonious particles.
3. The cup given, without conſecration, to the laity in the popiſh churches.

To A'BNE'GATE. *v. a.* [from *abnego*, Lat.] To deny.

ABNEGA'TION. *n. ſ.* [*abnegatio*, Lat. denial, from *abnego*, to deny.] Denial, renunciation.

The *abnegation* or renouncing of all his own holds and intereſts, and truſts of all that man is moſt apt to depend upon, that he may the more expeditely follow Chriſt. *Hammond's Practical Catechiſm.*

ABNODA'TION. *n. ſ.* [*abnodatio*, Lat.] The act of cutting away knots from trees; a term of gardening. *Dict.*

ABNO'RMOUS. *adj.* [*abnormis*, Lat. out of rule.] Irregular, miſhapen. *Dict.*

ABO'ARD. *adv.* [a ſea-term, but adopted into common language; derived immediately from the French *à bord*, as, *aller à bord*, *envoyer à bord*. *Bord* is itſelf a word of very doubtful original, and perhaps, in its different acceptations, deducible from different roots. Bopd, in the ancient Saxon, ſignified a *houſe*; in which ſenſe, *to go aboard*, is to take up reſidence in a ſhip.]

In a ſhip.

Which, when far off, Cymocles heard and ſaw,
He loudly call'd to ſuch as were *aboard*,
The little bark unto the ſhore to draw,
And him to ferry over that deep ford. *Fairy Q. b.* ii. *cant.* 6.

I made this anſwer, that he might land them, if it pleaſed him, or otherwiſe keep them *aboard*. *Sir W. Rawleigh's Eſſays.*

When morning roſe, I ſent my mates to bring
Supplies of water from a neighb'ring ſpring;
Whilſt I the motions of the winds explor'd;
Then ſummon'd in my crew, and went *aboard*.
Addiſon's Ovid's Metamorphoſes, b. iii.

ABO'DE. *n. ſ.* [from *abide*.]

1. Habitation, dwelling, place of reſidence.

But I know thy *abode* and thy going out, and thy coming in, and thy rage againſt me. 2 *Kings*, xix. 27.

Figure 48: Boxed Headword and Definitions

Words with definitions that span either multiple pages or multiple columns must be stitched together. Some words have definitions that span multiple pages entirely, and this is where the page dewarping and rotation can have the greatest effect on the cropping. We want the definitions, no matter how long they are, to be as straight as we can possibly get them. It is also very important to our sponsor that we retain the feeling of an actual book, so we both cannot let the

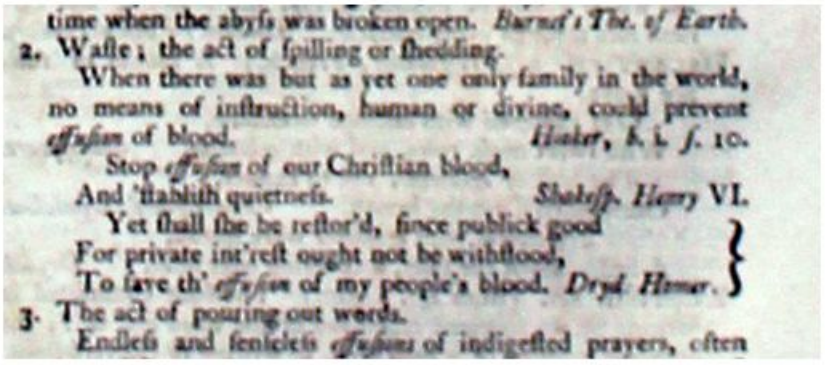
dewarping drop the quality down too much, and we cannot keep the image thresholded to black and white, the thresholding is only for column and definition detection. This is where the image stitching and brightness manipulation come into play, because the color across multiple pages can change, where some pages may appear to have a blue tint, and some may look more normal, like the above page shown in Figure 48. In order for the final crops to appear uniform and be appealing to the eyes, the colors should be adjusted so that the crops look as though they came from the same page. This was originally a stretch goal as stated by our sponsor, but we put it upon ourselves to make sure that this aspect of the program was implemented, and made it a requirement for the project to be considered successful.

5.9 Image Brightness Manipulation

Image brightness manipulation works separating an image into their individual pixels and assigning values to those pixels based on their color with values generally ranging from 0 to 255. It is important to note that all grids in this section are examples of how image manipulation software stores this information. They are not literal representations of the data present in the accompanying images. Such a literal representation of the images used would take up most of this document's length as each image is made up of well over hundreds of pixel values that must be stored in such an array.

2	123	64
195	71	39
0	77	214
20	64	178
6	87	164
11	38	13
137	46	231
159	194	250
185	147	201
190	255	130
180	200	89
38	24	43

Table 1: An example of a grid of pixels from a picture reduced to their pixel values


time when the abyſs was broken open. *Burnet's The. of Earth.*

2. Waſte; the act of ſpilling or ſhedding.

When there was but as yet one only family in the world, no means of inſtruction, human or divine, could prevent *effuſion* of blood. *Hooker*, b. i. ſ. 10.

Stop *effuſion* of our Chriſtian blood,
And ſtabliſh quietneſs. *Shakeſp. Henry* VI.

Yet ſhall ſhe be reſtor'd, ſince publick good
For private int'reſt ought not be withſtood,
To ſave th' *effuſion* of my people's blood. *Dryd. Homer.*

3. The act of pouring out words.

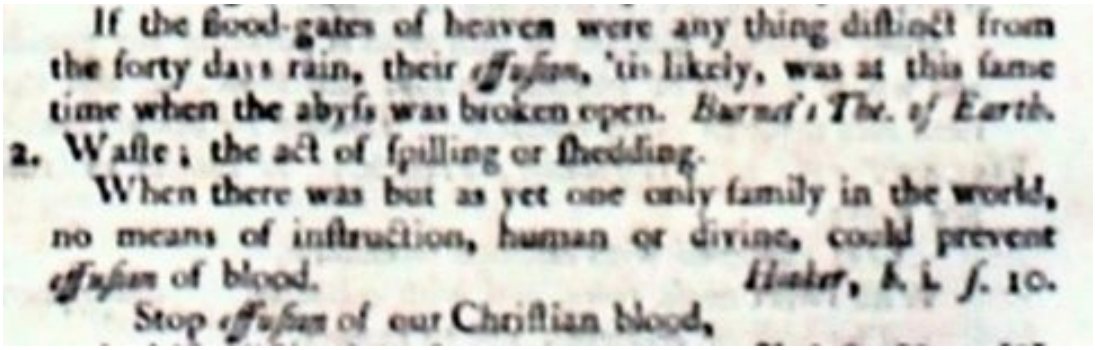
Endleſs and fenſeleſs *effuſions* of indigeſted prayers, often

Figure 49: A cropped shot of a page from the text

A very simple brightness contrast program would go through all these values and multiply each of them by a constant greater than 1. Any values that increase to a value greater than the max value of 255 are set to 255 instead. A value at 255, for example, remains unchanged.

3	184	96
255	106	58
0	115	255
30	96	255
9	130	253
16	57	19
205	69	255
238	255	255
255	220	255
255	255	195
1	0	133
57	36	64

Table 2: A table of contrasted pixel values


If the flood-gates of heaven were any thing diſtinct from the forty days rain, their *effuſion*, 'tis likely, was at this ſame time when the abyſs was broken open. *Burnet's The. of Earth.*

2. Waſte; the act of ſpilling or ſhedding.

When there was but as yet one only family in the world, no means of inſtruction, human or divine, could prevent *effuſion* of blood. *Hooker*, b. i. *ſ.* 10.

Stop *effuſion* of our Chriſtian blood,

Figure 50: An example of a contrast increased picture

A side effect of this is that faded background elements can be lost. For the sake of showing this we shall use a brightening program, as it results in the same change for the same reasons but is far more visible.

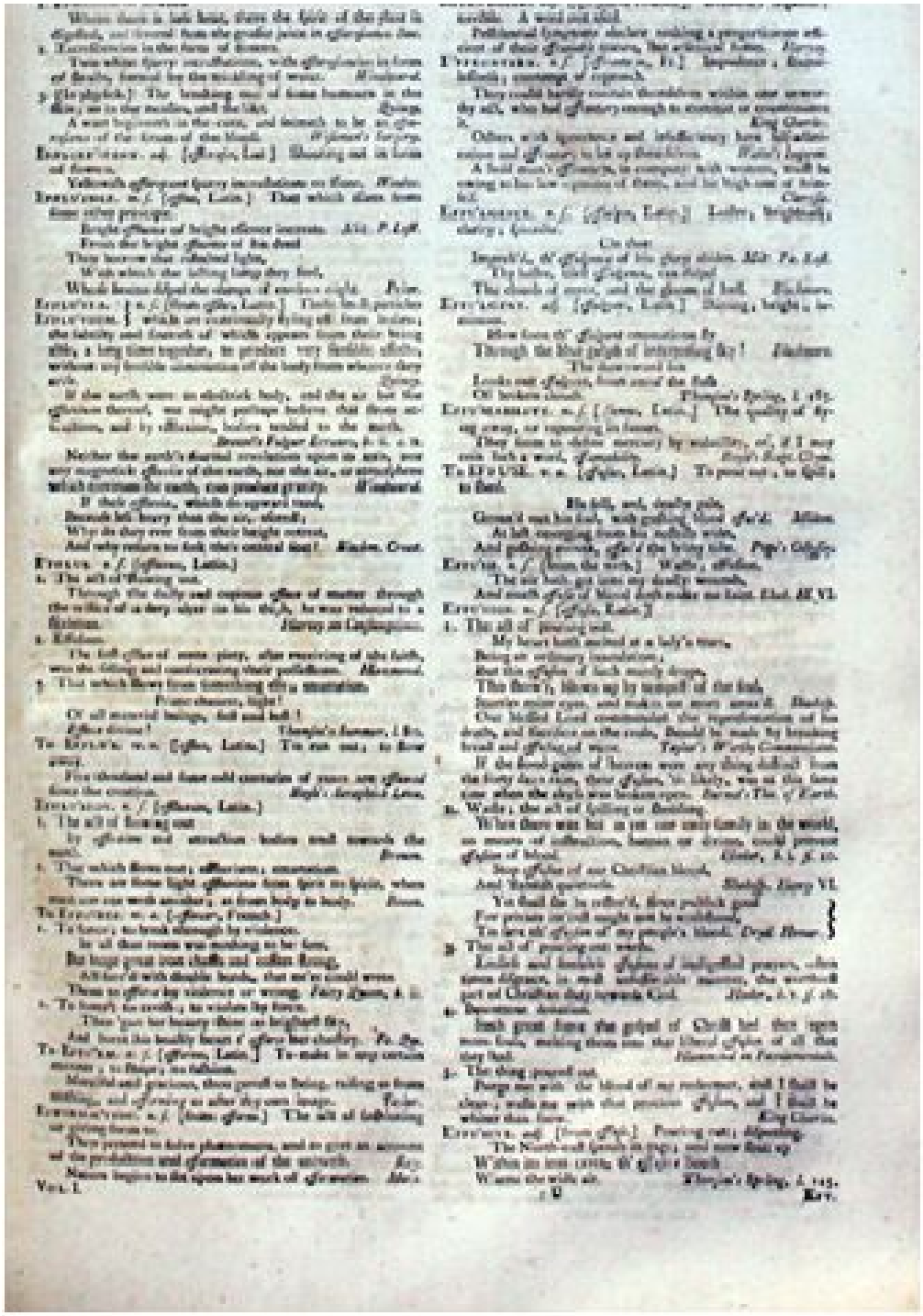


Figure 51: An example page before brightening

Figure 52: Page from Figure 51 after brightening

Note how the background page texture is largely removed by this operation. In some cases this improves readability, as the background pages nonuniform coloring naturally draws the eye away from the text. However, our project's goal is to archive an important historical text for easy accessibility online. Removing the page background fully would run counter to this goal. If all our project was expected to do was preserve the words on the paper themselves in an easy to read fashion, then we could create a program to detect the words and make a text document full of them. This would remove all blur and background noise, while also removing the unique qualities of a book entirely.

At that point the result would not be an easier to read alteration of the famous dictionary, it would be a simple dictation of its contents to a plain word file.

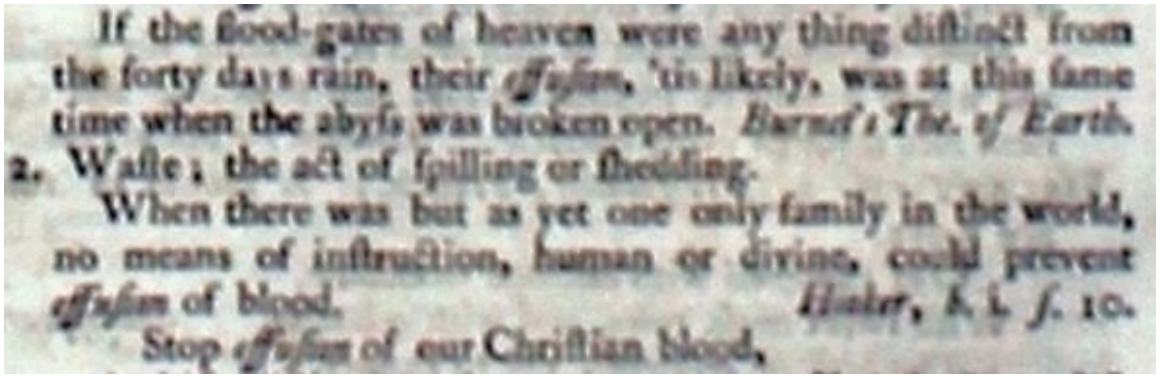
Contrast is decreased by using a constant between 0 and 1. The principle behind both is quite simple, multiplying by a value greater than 1 amplifies the difference between different pixels. A value between 0 and 1 reduces it.

$$P_o = C * P_i$$

Where p_i is the input pixel value, p_o is the output pixel value, and C is the constant.

1	61	32
97	35	19
0	38	107
10	32	89
3	43	84
5	19	6
68	23	115
79	97	125
92	73	100
95	127	65
0	0	44
19	12	21

Table 3: A contrast reduced grid


If the flood-gates of heaven were any thing diſtinct from the forty days rain, their *effuſion*, 'tis likely, was at this ſame time when the abyſs was broken open. *Burnet's The. of Earth.*

2. Waſte; the act of ſpilling or ſhedding.

When there was but as yet one only family in the world, no means of inſtruction, human or divine, could prevent *effuſion* of blood. *Hooker*, b. i. ſ. 10.

Stop *effuſion* of our Chriſtian blood,

Figure 52: A cropped shot of a page from the text

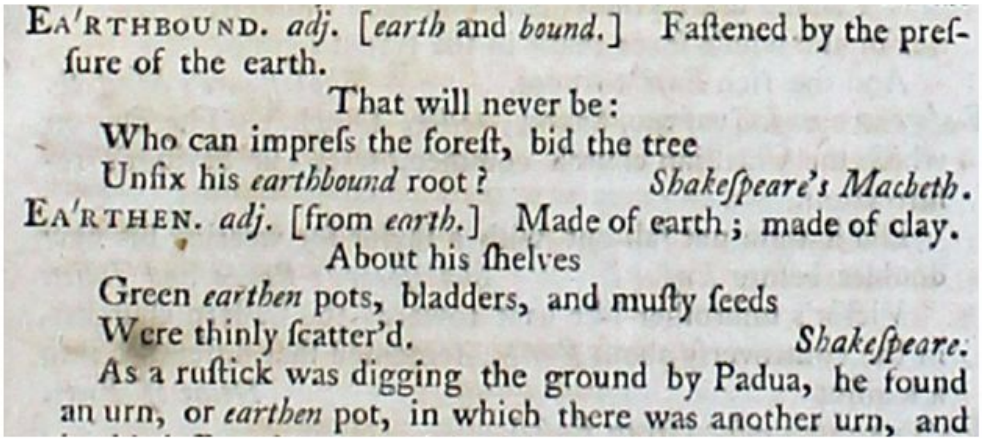
In the case of our program however, this program would lead to any text in a blurred image becoming faded as well. To combat this, we will be setting a threshold onto the brightness program to differentiate text pixels from background pixels. By doing this, we can use reduced contrast on the background that has been darkened and blurred without affecting the text's brightness. To keep the undimmed pages background at a brightness comparable to its original level we will use a normalization constant to affect how much the contrast constant changes a pixel's value. Pixels that fit within the range of a normal page's background value will be largely unaffected. This change can be demonstrated in a simple formula.

$$P_O = P_i * \hat{C}(\sigma)$$

In which sigma is represented with:

$$\sigma = \left| \frac{P_i - N}{N} \right|$$

Where N is the normalizing constant, set at a value close to that of the page background.


EA'RTHBOUND. *adj.* [*earth* and *bound.*] Faſtened by the preſ-
ſure of the earth.

That will never be:
Who can impreſs the foreſt, bid the tree
Unfix his *earthbound* root? *Shakeſpeare's Macbeth.*

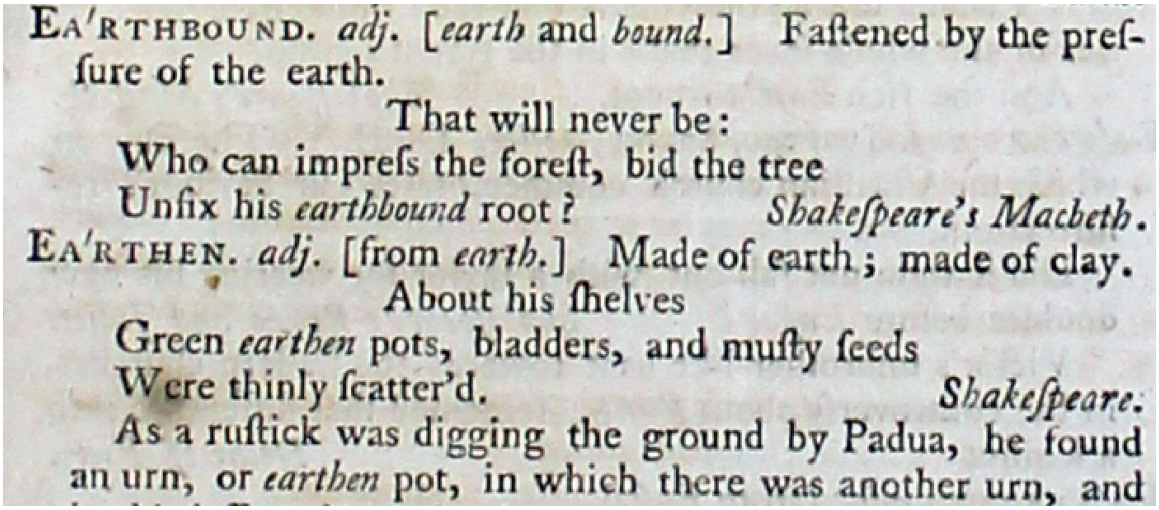
EA'RTHEN. *adj.* [from *earth.*] Made of earth; made of clay.

About his ſhelves
Green *earthen* pots, bladders, and muſty ſeeds
Were thinly ſcatter'd. *Shakeſpeare.*

As a ruſtick was digging the ground by Padua, he found
an urn, or *earthen* pot, in which there was another urn, and

Figure 53: Example page

In the case of this page the discoloration is over a single word rather than several lines. The completed program will have to remove this discoloration while retaining the same texture of the page around it.


EA'RTHBOUND. *adj.* [*earth* and *bound.*] Faſtened by the preſ-
ſure of the earth.

That will never be:
Who can impreſs the foreſt, bid the tree
Unfix his *earthbound* root? *Shakeſpeare's Macbeth.*

EA'RTHEN. *adj.* [from *earth.*] Made of earth; made of clay.

About his ſhelves
Green *earthen* pots, bladders, and muſty ſeeds
Were thinly ſcatter'd. *Shakeſpeare.*

As a ruſtick was digging the ground by Padua, he found
an urn, or *earthen* pot, in which there was another urn, and

Figure 54: Zoomed image on the blurred passage

5.9.1 Automatic Gamma Correction

The way the human eye detects light is not the same as the way all visual technology does. Where cameras pick up light at a linear rate, double the light is double the value, human eyes automatically adjust the light they receive nonlinearly.

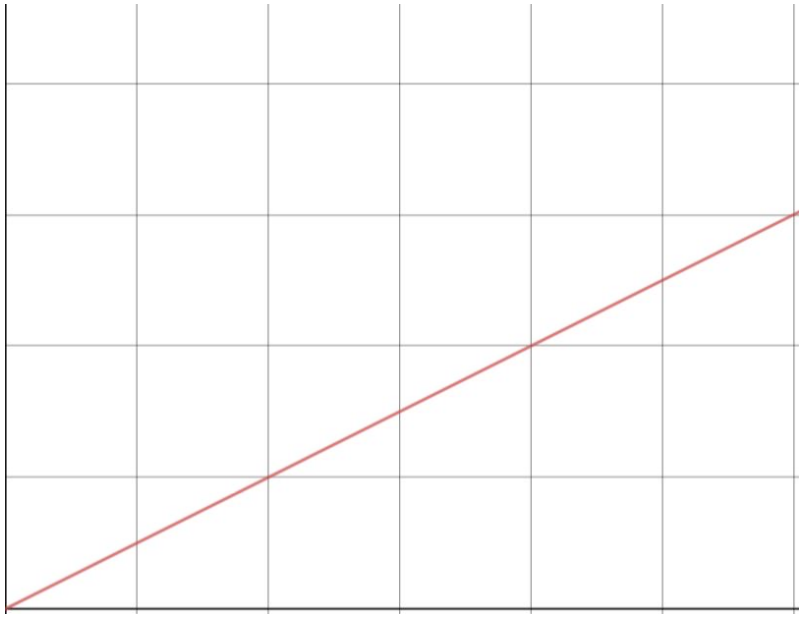


Figure 55: Linear pixel value contrast rate

To account for this when adjusting contrast and ensure the enhanced image does not gain unnatural changes to the human eyes that are not present to digital technology, automatic gamma correction is used to adjust it nonlinearly. The way human eyes perceive light can be approximated by power law function[1].

Automatic gamma correction is more specifically for camera images. Our project works on scanned images and because of this it may not have as much use. However, automatic gamma correction is also meant to allow contrast changes to images without resulting in them looking unnatural to the human eye. These changes are more present in naturally high contrast images, which should be nearly every page in our text. Even in a book the pages are still going to be

dark words on a bright background, the exact case where the principles used in automatic gamma correction apply most. By basing our contrast code off this formula it is possible that the resulting changes will be more fitting with the product we want our images to look like. Automatic gamma correction works very similar to our normalization constant In the contrast program now, both use power functions.

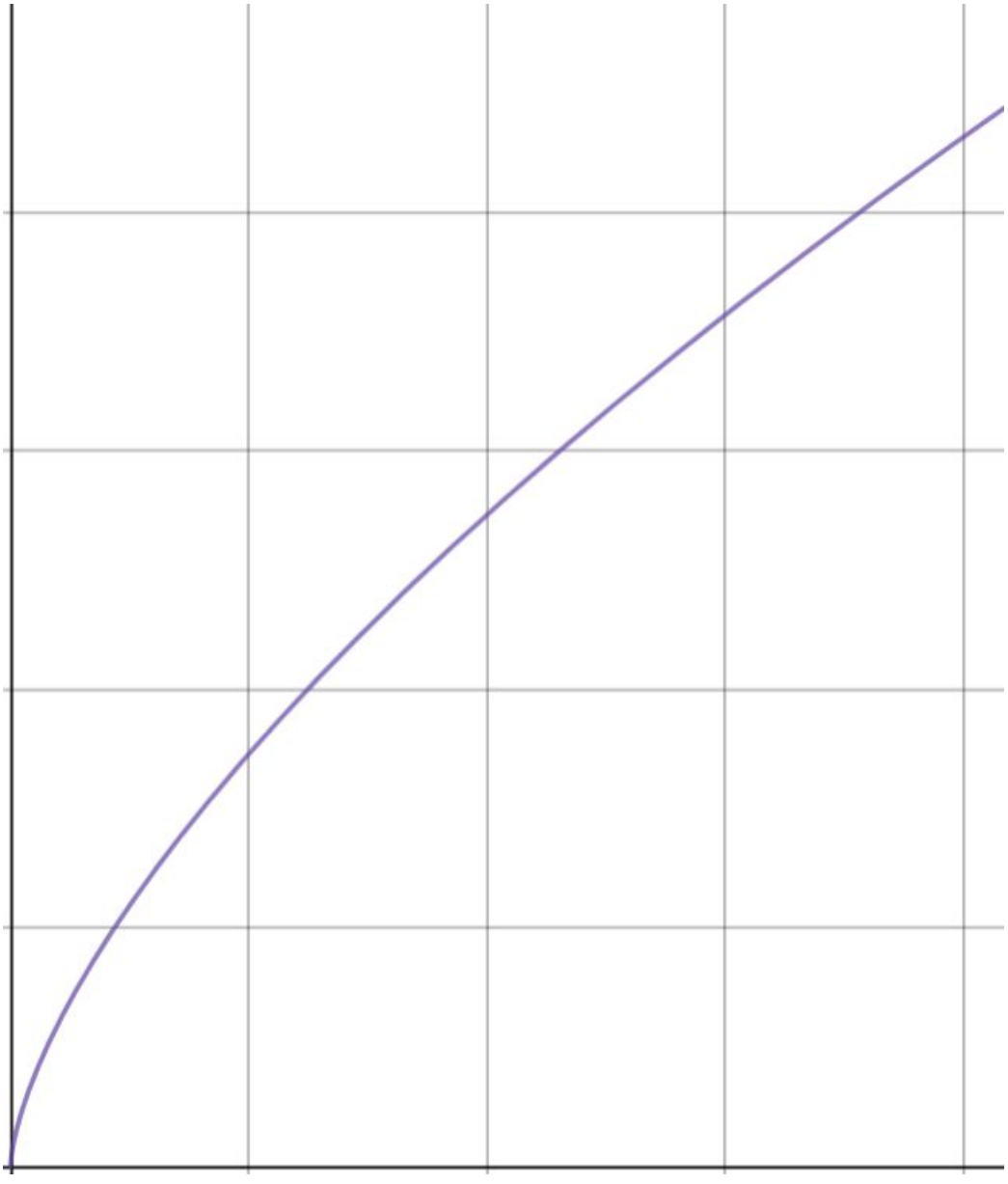


Figure 56: Power law pixel value contrast rate

5.9.2 Zeta Gamma Correction

Zeta gamma correction works somewhat like automatic gamma correction. The key difference between them is that zeta gamma correction is designed to ensure the constant for manipulation passes through the origin. This is useful for adjusting for digital information gleaned from image processing devices such as cameras and scanners. Given the nature of our project however, it would be excessive to use such a complicated formula for the image manipulation. This is especially true as zeta gamma correction is designed for pictures, it is more specifically a component useful for HDR video [3].

5.9.3 Image Sharpness Manipulation

While Image contrast is adjusted by manipulating the difference between pixels, image sharpness is adjusted removing borderline values.

20	14	32	9
99	102	67	2
223	200	84	92
190	220	160	173

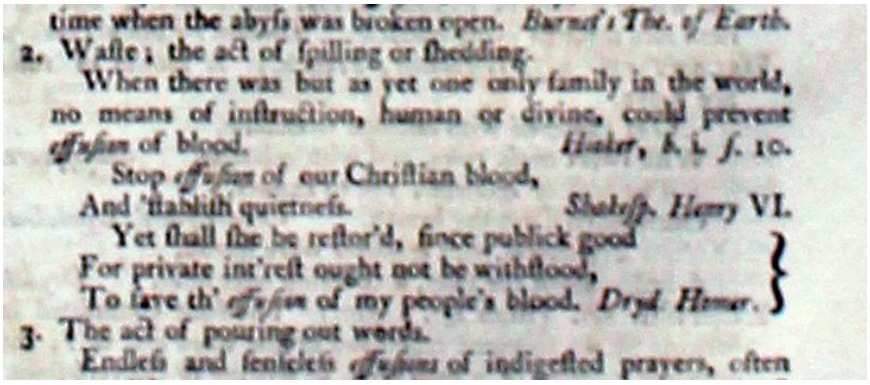
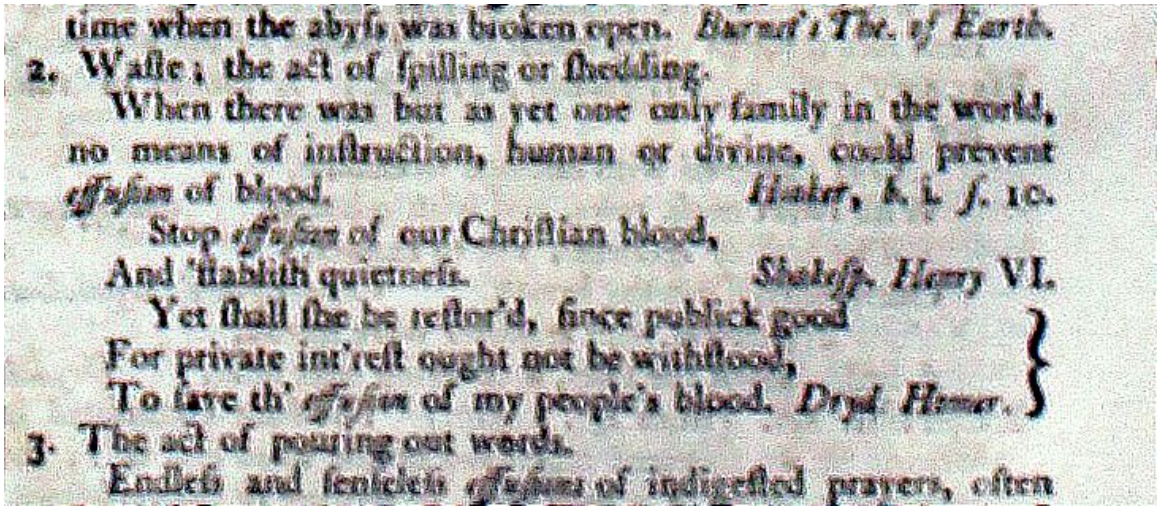

time when the abyſs was broken open. *Burnet's The. of Earth.*
2. Waſte; the act of ſpilling or ſhedding.
When there was but as yet one only family in the world, no means of inſtruction, human or divine, could prevent *effuſion* of blood. *Hooker*, b. i. ſ. 10.
Stop *effuſion* of our Chriſtian blood,
And 'ſtabliſh quietneſs. *Shakeſp. Henry VI.*
Yet ſhall ſhe be reſtor'd, ſince publick good
For private int'reſt ought not be withſtood,
To ſave th' *effuſion* of my people's blood. *Dryd. Homer.*
3. The act of pouring out words.
Endleſs and fenſeleſs *effuſions* of indigeſted prayers, often

Figure 57: An example of a grid of pixel values near a blurred edge.

A program to manipulate an image to be sharper would find values like those above placed adjacent to high threshold values and remove them. This method works best on pictures and will require much finer settings on threshold

removal to remove blurring on text without risking removing any portion of the letters.

20	14	32	9
0	0	0	2
223	200	0	0
190	220	160	173


time when the abyſs was broken open. *Burnet's Th. of Earth.*

2. Waſte; the act of ſpilling or ſhedding.

When there was but as yet one only family in the world, no means of inſtruction, human or divine, could prevent *effuſion* of blood. *Hooker*, b. i. *ſ.* 10.

Stop *effuſion* of our Chriſtian blood,
And 'ſtablith quietneſs. *Shakſp. Henry* VI.

Yet ſhall ſhe be reſtor'd, ſince publick good
For private int'reſt ought not be withſtood,
To ſave th' *effuſion* of my people's blood. *Dryd. Hind.*

3. The act of pouring out words.

Endleſs and fenſeleſs *effuſions* of indigeſted prayers, often

Figure 58: The grid and image after being sharpened

In the case of our program, because of the background value not being 0, we will have to use our own constant value. This value will ensure any blurring removed does not lead to white spots inconsistent with the coloring of the page background present elsewhere. As mentioned before, in the case of our program the sharpening program will need to make very subtle adjustments, overuse of sharpening such as this on an image can easily lead to a result like this.

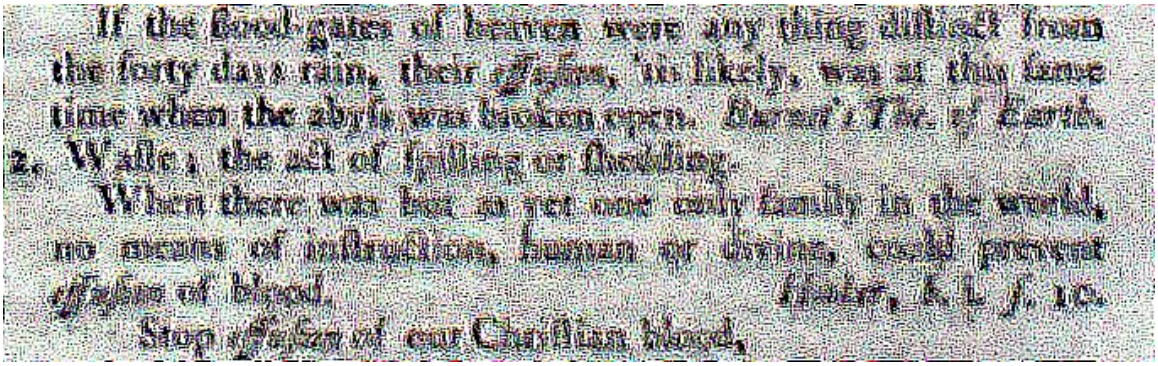


Figure 59: Over-sharpened image.

The blur is fairly well removed from the word, but due to the thin shapes of text small gaps appear in the words on the page. Additionally, the background of the page gains an almost distorted appearance. A limit on the values affected should mitigate this effect on the background, as it is far brighter than the text in the image. It is also possible to fix this to a degree with a blurring program.

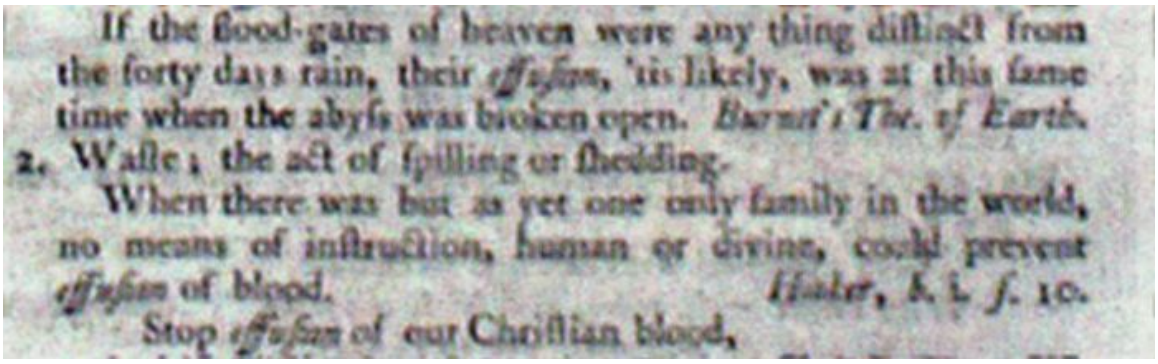
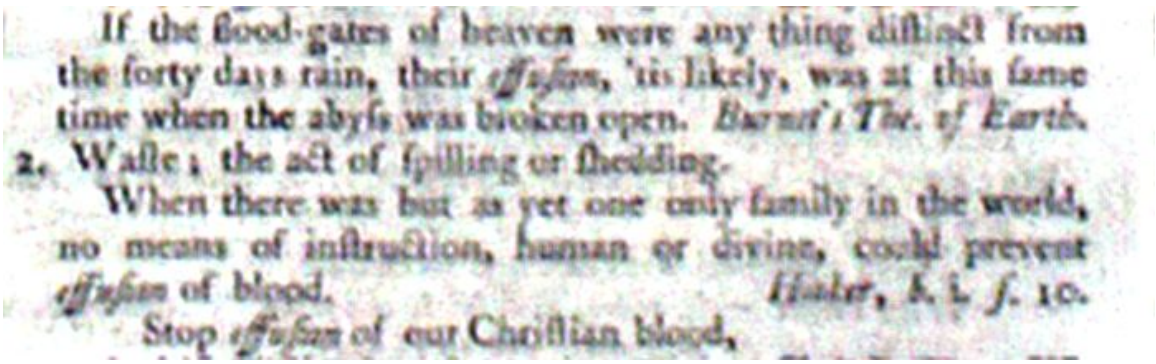


Figure 60: Over sharpened image blurred lightly.

The dotted effect on the background and the text is removed quite well but the page takes on a more discolored effect, this can also be fixed however with brightening or contrast programs.


If the flood-gates of heaven were any thing diſtinct from the forty days rain, their *effuſion*, 'tis likely, was at this ſame time when the abyſs was broken open. *Burnet's The. of Earth.*

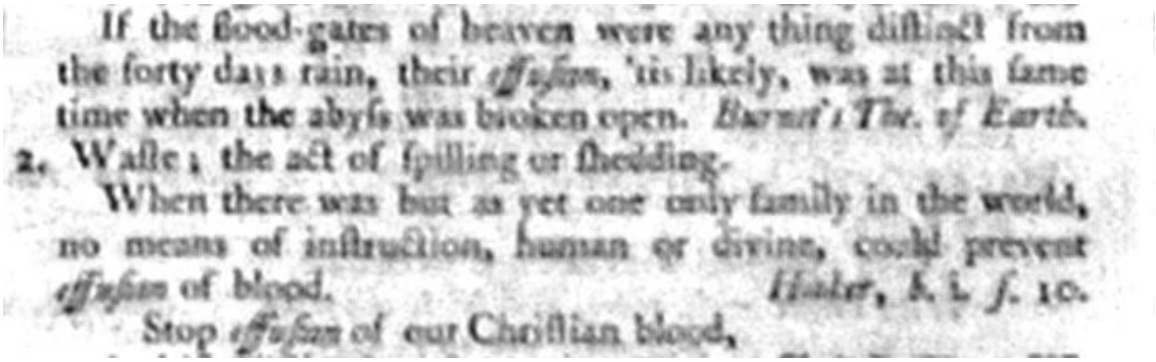
2. Waſte; the act of ſpilling or ſhedding.

When there was but as yet one only family in the world, no means of inſtruction, human or divine, could prevent *effuſion* of blood. *Hooker*, b. i. *ſ.* 10.

Stop *effuſion* of our Chriſtian blood,

Figure 61: Image brightened to reduce background noise.

There is still a discoloration to the page however along the edges of words both on this page and from the faded words on the other side of the page. A slight black and white effect removes this.


If the flood-gates of heaven were any thing diſtinct from the forty days rain, their *effuſion*, 'tis likely, was at this ſame time when the abyſs was broken open. *Burnet's The. of Earth.*

2. Waſte; the act of ſpilling or ſhedding.

When there was but as yet one only family in the world, no means of inſtruction, human or divine, could prevent *effuſion* of blood. *Hooker*, b. i. *ſ.* 10.

Stop *effuſion* of our Chriſtian blood,

Figure 62: Image with the discoloration removed.

A more complex program should be able to remove this without affecting the yellowed color naturally present in an aged page. With this the initially present is largely removed, however the quality of the text is reduced. The background is now more spotted, even with the blurring effect it still stands out more than the typical page background did. The words themselves might have lost the wider blurring that was present, but the edges of the text are less defined. In this case, the image is focused on the worst of the blurring on the page, these unwanted side effects are far less present on the rest of the page.

EFF

Your answering in the final crude, makes one believe you are at a loss for the *efficient*. *Collier on Thought.*

To EFFIGIATE. *v. a.* [*effigio*, Latin.] To form in semblance; to image.

EFFIGIA'TION. *n. f.* [from *effigiate.*] The act of imaging; or forming the resemblance of things or persons. *Dict.*

EFFIGIES. } *n. f.* [*effigies*, Latin.] Resemblance; image in
EFFI'GY. } painting or sculpture; representation; idea.

We behold the species of eloquence in our minds, the *effigies* or allied image of which we seek in the organs of our hearing. *Dryden's Dufresnoy, Preface.*

EFFLORE'SCENCE. } *n. f.* [*efflorescо*, Latin.]
EFFLORE'SCENCY.

1. Production of flowers.

Where there is less heat, there the spirit of the plant is digested, and severed from the grosser juice in *efflorescence.* *Bac.*

2. Excrescencies in the form of flowers.

Two white spurry incrustations, with *efflorescencies* in form of shrubs, formed by the trickling of water. *Woodward.*

3. [In physick.] The breaking out of some humours in the skin; as in the measles, and the like. *Quincy.*

A wart beginneth in the cutis, and ferments to be an *efflorescence* of the serum of the blood. *Wiseman's Surgery.*

EFFLORE'SCENT. *adj.* [*efflorescо*, Lat.] Shooting out in form of flowers.

Yellowish *efflorescent* spurry incrustations on stone. *Woodw.*

EFFLU'ENCE. *n. f.* [*effluo*, Latin.] That which flows from some other principle.

Bright *effluence* of bright essence increate. *Milt. P. Lost.*

From the bright *effluence* of his deed
They borrow that reflected light,
With which the lasting lamp they feed,
Whose beams dispel the damps of envious night. *Prior.*

EFFLU'VIA. } *n. f.* [from *effluo*, Latin.] Those small particles
EFFLU'VIUM. } which are continually flying off from bodies; the subtilty and fineness of which appears from their being able, a long time together, to produce very sensible effects, without any sensible diminution of the body from whence they arise. *Quincy.*

If the earth were an electrick body, and the air but the *effluvium* thereof, we might perhaps believe that from attraction, and by effusion, bodies tended to the earth. *Brown's Vulgar Errours, b. ii. c. 2.*

Neither the earth's diurnal revolution upon its axis, nor any magnetick *effluvia* of the earth, nor the air, or atmosphere which environs the earth, can produce gravity. *Woodward.*

If these *effluvia*, which do upward tend,
Because less heavy than the air, ascend;
Why do they ever from their height retreat,
And why return to seek their central seat? *Blackm. Creat.*

EFFLUX. *n. f.* [*effluxus*, Latin.]

1. The act of flowing out.

Through the daily and copious *efflux* of matter through the orifice of a deep ulcer in his thigh, he was reduced to a skeleton. *Harvey on Consumptions.*

2. Effusion.

The first *efflux* of mens piety, after receiving of the faith, was the selling and consecrating their possessions. *Hammond.*

3. That which flows from something else; emanation.

Prime cheerer, light!
Of all material beings, first and best!
Efflux divine! *Thomson's Summer, l. 80.*

To EFFLU'X. *v. n.* [*effluo*, Latin.] To run out; to flow away.

Five thousand and some odd centuries of years are *efflowed* since the creation. *Boyle's Seraphick Love.*

EFFLU'XION. *n. f.* [*effluxus*, Latin.]

1. The act of flowing out.

By *effluxion* and attraction bodies tend towards the earth. *Brown.*

2. That which flows out; effluvium; emanation.

There are some light *effluxions* from spirit to spirit, when men are one with another; as from body to body. *Bacon.*

To EFFO'RCE. *v. a.* [*efforcer*, French.]

1. To force; to break through by violence.

In all that room was nothing to be seen,
But huge great iron chests and coffers strong,
All barr'd with double bands, that none could ween
Them to *efforce* by violence or wrong. *Fairy Queen, b. ii.*

2. To force; to ravish; to violate by force.

Then 'gan her beauty shine as brightest fire,
And burst his headily heart t' *efforce* her chastity. *Fa. Qu.*

To EFFO'RM. *n. f.* [*efformo*, Latin.] To make in any certain manner; to shape; to fashion.

Merciful and gracious, thou gavest us being, raising us from nothing, and *efforming* us after thy own image. *Taylor.*

EFFORMA'TION. *n. f.* [from *efform.*] The act of fashioning or giving form to.

They pretend to solve phænomena, and to give an account of the production and *efformation* of the universe. *Ray.*

Nature begins to sit upon her work of *efformation.* *Mere.*

EFF

EFFO'RT. *n. f.* [*effort*, French.] Struggle; laborious endeavour.

If, after having gained victories, we find made the same *efforts* as if we had lost them, France could not have wished us. *Addison on the State of the War.*

Though the same sun, with all diffusive rays,
Blush in the rose, and in the diamond blaze,
We prize the stronger *effort* of his pow'r,
And always fix the great above the low'r. *Pope, Epist. i.*

EFFO'SION. *n. f.* [*effossus*, Latin.] The act of digging up from the ground; detestation.

He set apart annual sums for the recovery of manuscripts, the *effossion* of coins, and the procuring of mummies. *Arbuthnot.*

EFFRA'CTABLE. *adj.* [*effractus*, Latin.] Destructible; frangible; terrible. A word not used.

Pestilential symptoms declare nothing a proportionate efficient of their *effractable* nature, but arsenical fumes. *Harvey.*

EFFRONTERY. *n. f.* [*effronterie*, Fr.] Impudence; shamelessness; contempt of reproach.

They could hardly contain themselves within one unworthy act, who had *effrontery* enough to commit or countenance it. *King Charles.*

Others with ignorance and insufficiency have self-admiration and *effrontery* to set up themselves. *Watts's Improv.*

A bold man's *effrontery*, in company with women, must be owing to his low opinion of them, and his high one of himself. *Clarissa.*

EFFU'LGENCE. *n. f.* [*effulgeo*, Latin.] Lustre; brightness; clarity; splendour.

On thee
Impress'd, th' *effulgence* of his glory abides. *Milt. Pa. Lost.*

Thy lustre, thick *effulgence*, can dispel
The clouds of error, and the gloom of hell. *Blackmore.*

EFFU'LGENT. *adj.* [*effulgens*, Latin.] Shining; bright; luminous.

How soon th' *effulgent* emanations fly
Through the blue gulph of interposing sky! *Blackmore.*

The downward sun
Looks out *effulgent*, from amid the flush
Of broken clouds. *Thomson's Spring, l. 185.*

EFFU'MABILITY. *n. f.* [*fumus*, Latin.] The quality of flying away, or vapouring in fumes.

They seem to define mercury by volatility, or, if I may coin such a word, *effumability.* *Boyle's Sceptical Chym.*

To EFFU'SE. *v. a.* [*effusus*, Latin.] To pour out; to spill; to shed.

He fell, and deadly pale,
Groan'd out his soul, with gushing blood *effus'd.* *Milton.*

At last emerging from his suffrils wide,
And gushing streams, *effus'd* the briny tide. *Pope's Odyssey.*

EFFU'SE. *n. f.* [from the verb.] Waste; effusion.

The air hath got into my deadly wounds,
And much *effuse* of blood doth make me faint. *Shak. H. VI.*

EFFU'SION. *n. f.* [*effusio*, Latin.]

1. The act of pouring out.

My heart hath melted at a lady's tears,
Being an ordinary inundation;
But this *effusion* of such manly drops,
This show'r, blown up by tempest of the soul,
Startles mine eyes, and makes me more amaz'd. *Shakesp.*

Our blessed Lord commanded the representation of his death, and sacrifice on the cruse, should be made by breaking bread and *effusion* of wine. *Taylor's Worthy Communicant.*

If the flood-gates of heaven were any thing distinct from the forty days rain, their *effusion*, 'tis likely, was at this same time when the abyss was broken open. *Burnet's The. of Earth.*

2. Waste; the act of spilling or shedding.

When there was but as yet one only family in the world, no means of satisfaction, human or divine, could prevent *effusion* of blood. *Hooker, b. v. f. 10.*

Stop *effusion* of our Christian blood,
And stablish quietness. *Shakesp. Henry VI.*

Yet shall the be softer'd, force publick good
For private int'rest ought not be withstood,
To save th' *effusion* of my people's blood. *Dryd. Hynn.*

3. The act of pouring out words.

Loathsome and senseless *effusions* of indigested prayers, often times disgrace, in most unbeseemable manner, the worthiest part of Christian duty towards God. *Hooker, b. v. f. 28.*

4. Bounteous donation.

Such great force the gospel of Christ had then upon mens souls, melting them into that liberal *effusion* of all that they had. *Hammond on Fundamentals.*

5. The thing poured out.

Purge me with the blood of my redeemer, and I shall be clean; wash me with that precious *effusion*, and I shall be whiter than snow. *King Charles.*

EFFU'SIVE. *adj.* [from *effuse.*] Pouring out; dispersing.

The North-east spends its rage; and now shut up
Within its iron caves, th' *effusive* South
Warms the wide air. *Thomson's Spring, l. 145.*

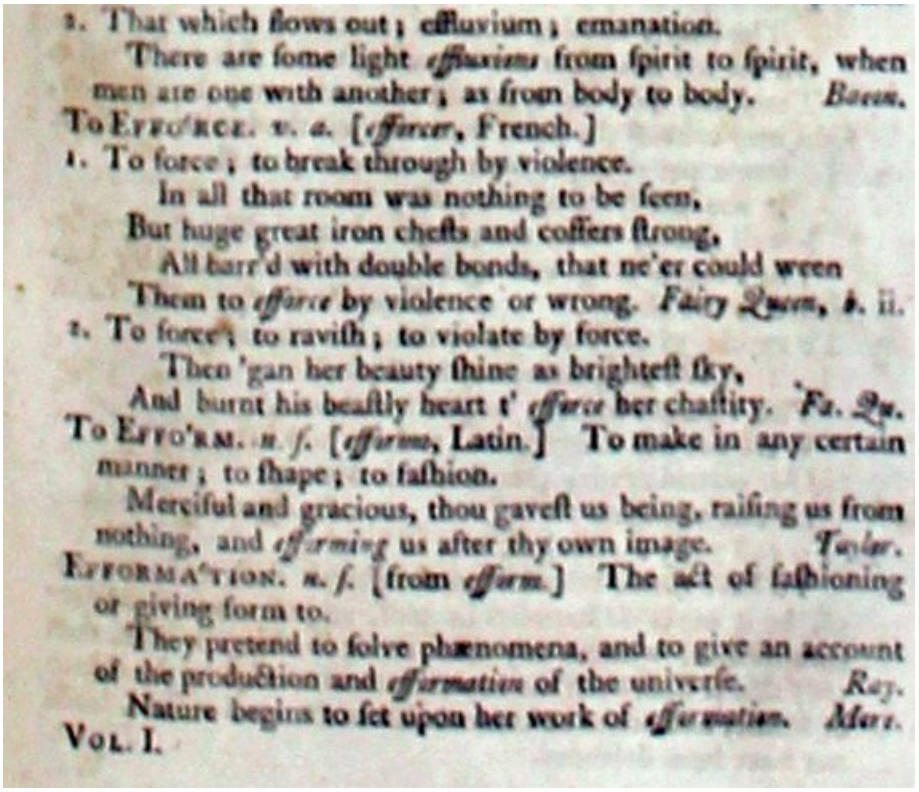
Figure 63: Full page after the process

In the previous example of a darkened page, we used a specific discoloration present on the page. This next example is a page which has both blurred and discolored text due to scanner error.

Figure 64: Page affected by scanner error.

At a full view this page shows less uniform background color enhanced from the normal pattern of the pages color. The focus of our code is to fix the smaller scale blurring and discoloration.

Another possible method for removing blur without affecting text is with edge detection. While blur and sharpening programs look for threshold cases to adjust, they are ultimately designed for more gradual pictures. A shot of landscape or a city skyline, the varying details present in such pictures is where blur and sharpening programs are most used. Edge detection is used for pictures with less background noise. Several shapes on a white background or, in this case, words on a page. On its own, edge detection can not improve the quality of the page, but it can possibly serve to detect words and separate them from the effects of any image manipulation done on the background. Doing so may allow the program to remove the blur present with the methods above without fading out the text in the image.



2. That which flows out; effluvium; emanation.
There are ſome light *effluxions* from ſpirit to ſpirit, when men are one with another; as from body to body. *Bacon.*

To EFFO'RCE. *v. a.* [*efforcer*, French.]

1. To force; to break through by violence.
In all that room was nothing to be ſeen,
But huge great iron cheſts and coffers ſtrong,
All barr'd with double bonds, that ne'er could ween
Them to *efforce* by violence or wrong. *Fairy Queen*, b. ii.

2. To force; to raviſh; to violate by force.
Then 'gan her beauty ſhine as brighteſt ſky,
And burnt his beaſtly heart t' *efforce* her chaſtity. *Fa. Qu.*

To EFFO'RM. *v. a.* [*efformo*, Latin.] To make in any certain manner; to ſhape; to faſhion.
Merciful and gracious, thou gaveſt us being, raiſing us from nothing, and *efforming* us after thy own image. *Taylor.*

EFFORMA'TION. *n. ſ.* [from *efform.*] The act of faſhioning or giving form to.
They pretend to ſolve phænomena, and to give an account of the production and *efformation* of the univerſe. *Ray.*
Nature begins to ſet upon her work of *efformation*. *Mort.*

VOL. I.

Figure 65: Zoomed in example

In this passage there are several words that are clearly blurred, to the point of hurting readability. The page as a whole also suffers from a change in color from the normal coloring on other pages, both of these issues must be

addressed by our programs image brightness and image sharpness manipulation without altering any of the text.

Given the previous examples shown it should be clear that the order our program applies changes to the text is as important as the changes themselves. Each process has a clear impact on the changes caused by the processes run after it. Contrast applied before sharpening will result in a very different picture then contrast after sharpening. It is also clear that minor adjustments counter to these main changes will have a large impact. Our final program will likely apply smoothing or even blur as a part of creating the most ideal image possible. Even if the initial problem with the pages was a blurred word.

This also showcases how a combination of the concepts between each of these methods might help our program to enhance the image. A brightness program might remove the background texture normally. But if we apply the principles present in sharpening programs, applying thresholds at which there is an effect. The program could brighten just the background around the text with very little impact to the page itself.

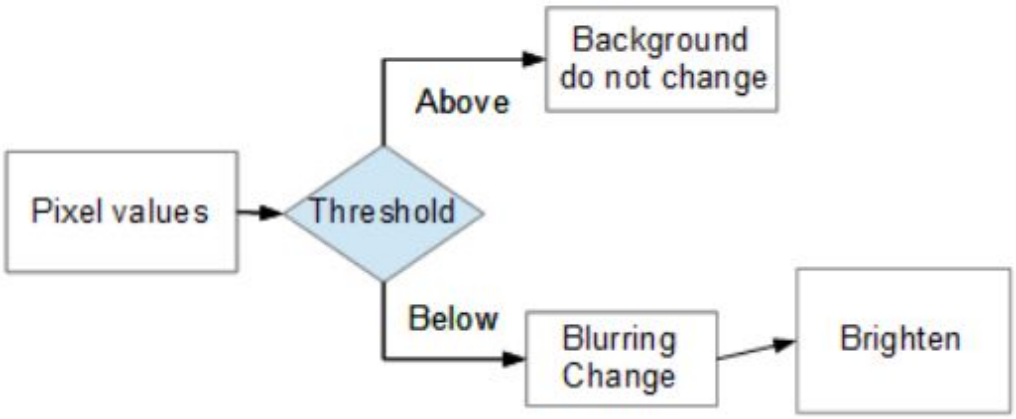


Figure 66: Flowchart for applying brightening

Then we apply a secondary threshold to separate text from blurring. This allows us to brighten the blurring without fading the text on the page as we would normally with any brighten program.

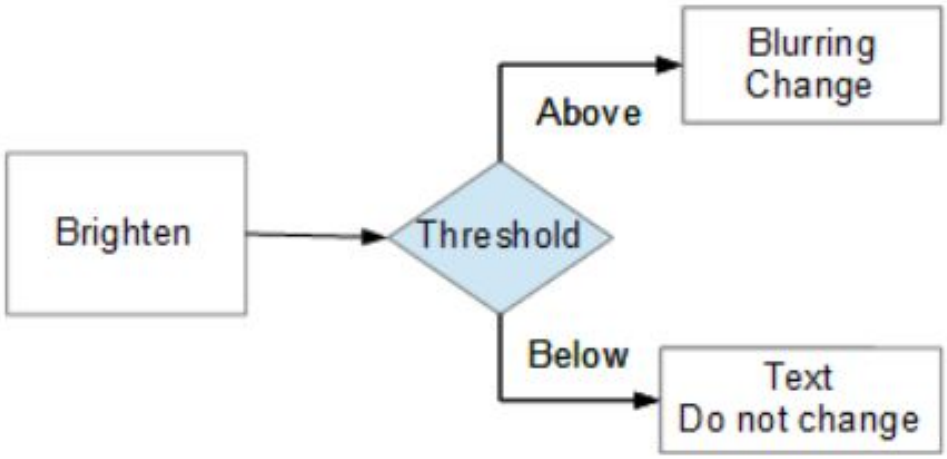


Figure 67: Flowchart for separating text

Applying a normalizing constant to the blurring should ensure this threshold does not lead to a large gap in color compared to the normal page background. This does however risk removing text fade in from the other side of the page. In addition, if the text is sufficiently blurred, setting a threshold too low would risk the program detecting the text as blurring. If the blurring in question is as dark as the text then this hybrid program will fail to remove it. Setting the code to ensure the text remains means setting it so that the blur would as well. For such a case we will need to rely on image sharpening to remove the blur rather than a hybrid brightness manipulation program. For any other case however, this method removes much of the risk of leaving holes in text present in the previously shown sharpened images. Applying the hybrid program in addition to a more lenient sharpening program to remove any remaining traces should serve for the vast majority of text discoloration.

5.9.4 Convolutional Neural networks and Image brightness manipulation

Convolutional Neural networks are an ever-growing technique in the robot vision field for teaching programs how to identify various patterns in a large set of images [6]. In our project specifically this is most obviously useful for our cropping program. We have a massive database of images to pull from with a very specific pattern present in all of them for entries. What is less obvious however is how this could be applied to our image enhancement code. By the same logic that governs its use in finding entries. Convolutional neural networks could be applied to enhancing the image. We have a large database of images to work on, and there is a specific pattern present to each of them for the program to learn.

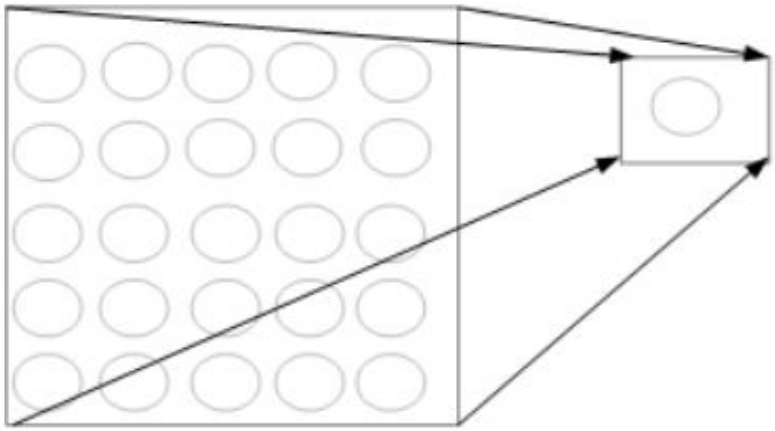


Figure 68: Pattern for convolutional neural network

Unlike our entry program which would work based on following the pattern, a machine learning program for image enhancement would work on finding where the pattern falls through. Text has a very clear standard throughout the document, this could be used to find places that are discolored as they would not fit this pattern. Rather than teaching our program a pattern to follow, we could teach it a pattern to spot where the pattern fails. For any image cropping program using machine learning, this case would be a problem. If the dictionary does not use a clear uniform standard for entries throughout it, then the program cannot detect such entries for all the text reliably. It will always miss some cases and pouring through the text to find these cases would create a large amount of work. Admittedly this work would still be far less then is currently posed by attempting

to crop the entire text by hand, and doing so would almost assuredly create far more errors then our program will have even if a small number of entries fail to follow this pattern.

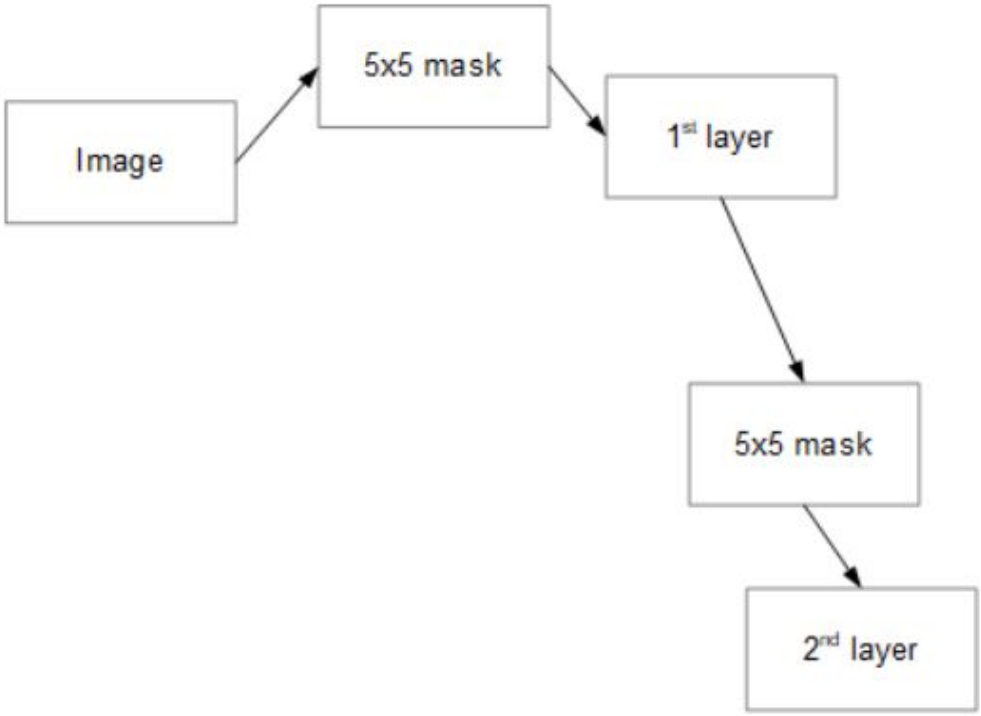


Figure 69: Pattern for convolutional neural network

5.9.5 Unsharp masking

Unsharp masking is a unique technique for image enhancement in that it is fairly simple and has a low computational cost [11]. It works based on a relatively easy to understand formula in its basic form.

$$y(n,\ m) = x(n,\ m) + \lambda * z(n,\ m)$$

In which y is the output sharpened image, x is the input blurred image, λ is the chosen gain factor, and z is the details to be removed. The unsharp masking method works by creating a negative version of the image made up of the blurred components. When these are added to the base image, they allow the program to remove the blur, leaving behind only the original text. This is an interesting contrast to the most basic sharpening program. In the basic sharpening program covered previously, the borderline pixels in any shape in the image are removed from the image to create sharp edges. The unsharp masking method effectively achieves this with a more gradual effect on the whole image at once. This will allow us to get around the problem previously discussed with a basic sharpening program where incredibly specific thresholds must be chosen to avoid removing text due to its thin nature. Because of how unsharp masking targets blur it is far less likely to cause this problem. The gain factor in the unsharp masking method works like the threshold does in a basic sharpening program, using too great of a gain factor can lead to over sharpening in an image just like incorrect use of a threshold can. In the case of our project, the blurring in our images are largely inconsistent; most pages have limited blurring sometimes to the point that only a single word or letter is blurred on occasion.

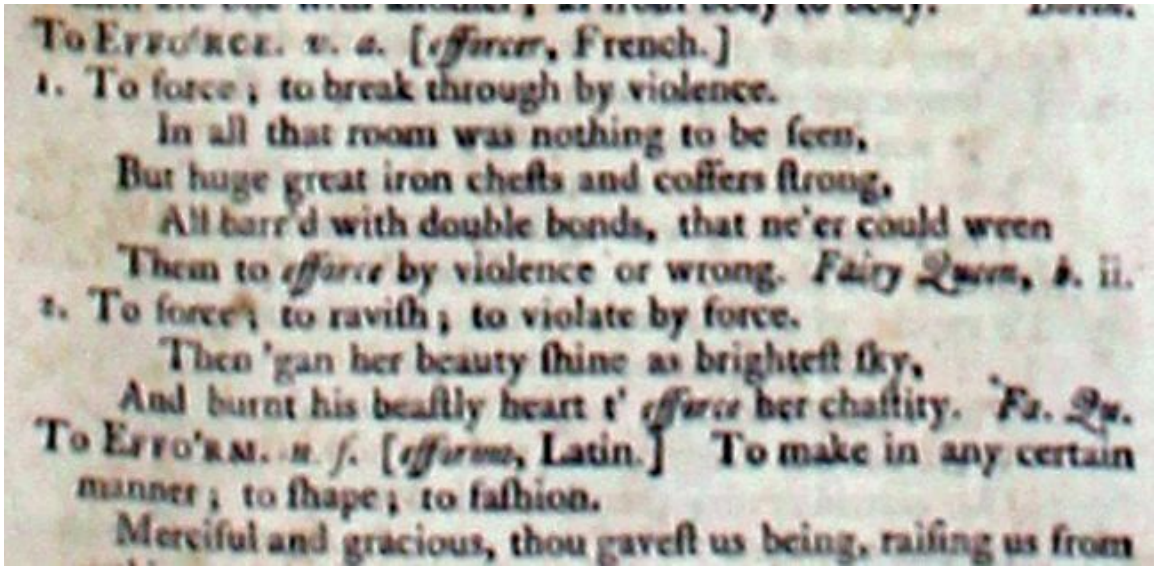

To Effo'rce. v. a. [*efforcer*, French.]
1. To force; to break through by violence.
In all that room was nothing to be seen,
But huge great iron chefts and coffers ſtrong,
All barr'd with double bonds, that ne'er could ween
Them to *efforce* by violence or wrong. *Fairy Queen*, b. ii.
2. To force; to raviſh; to violate by force.
Then 'gan her beauty ſhine as brighteſt ſky,
And burnt his beaſtly heart t' *efforce* her chaſtity. *Fa. Qu.*
To Effo'rm. v. ſ. [*efformo*, Latin.] To make in any certain manner; to ſhape; to faſhion.
Merciful and gracious, thou graveſt us being, raiſing us from

Figure 70: Example of single letter blur

In such a case where the blur is extremely limited and varied, the basic unsharp program will prove less effective. Splitting up the program to run the same effect on separate portions of the image however will mitigate this. By running the unsharp masking code on portions of the image rather than the whole image at once, it is possible to remove some of the effects from nonuniform blur. Normally an adaptive unsharp masking program would be used on such a document. In an adaptive unsharp masking program, the gain value used to create the detail map is adjusted based on the gradient of the given block of the image. This is a high computing power process and usually involves far more complex code then the average unsharp masking program. In our case while the blur is inconsistent, the text in the book is not. Using an adaptive unsharp masking program would risk over sharpening and losing text. The high computing consumption for such a program is less of a concern for this project, as the amount of time available for the program to run is quite large and must only be done once for the final product. While our cropping program will need several runs for testing, the image manipulation program is applied to a much more generic problem. It will not need a sample size anywhere near the size of the image cropping system we plan to implement. Additionally, it is possible to mitigate some of the danger from an adaptive unsharp masking program by limiting the range of possible gain factors.

5.9.6 Kernel estimation

Kernel estimation is a technique used to remove blur in an image by predicting the conditions that caused the blur and reversing them. This is most used in motion blur, as the blurring caused to an image by motion blur is quite predictable. It is unclear in our case how effective such a method would be as most of the blur in our images comes from the age of the pages and the placement of the text on the scanner at the time of scanning. However, a portion of this blur could have come from slight motion in the pages as they were being scanned. There are hundreds of pages of this dictionary in .tif files, scanning each one without a single instance of motion blur is highly unlikely. These few cases of motion blur are not likely worth investing any more time in applying Kernel estimation to our program especially with the possible risk to parts of the text with non-motion blur on them. It is unpredictable how such a program would change blur not caused by motion.

5.10 Image Stitching

Our program is required to crop entire entries into a single image. This is true even for entries which span multiple pages. As such, our program must be able to stitch two different images together for any entry on more then one page. Doing this will require a buffer of some sort, some place to store the portions of an image cropped so far to ensure that it can switch to a new image file for the next page and still keep the cropped entry present on the last page. This image file will add on any other cropped sections of the text into the same file directly below the first crop. The loop provided below should allow for any manner of entry, even if the entire dictionary were to be one single entry. This would however create a single excessively large file. If any entries are found to be exceptionally long, which can be done simply by searching for the largest of the image files, it would be a simple matter to implement a threshold on length if it was not necessary to split entries to fit within certain lengths.

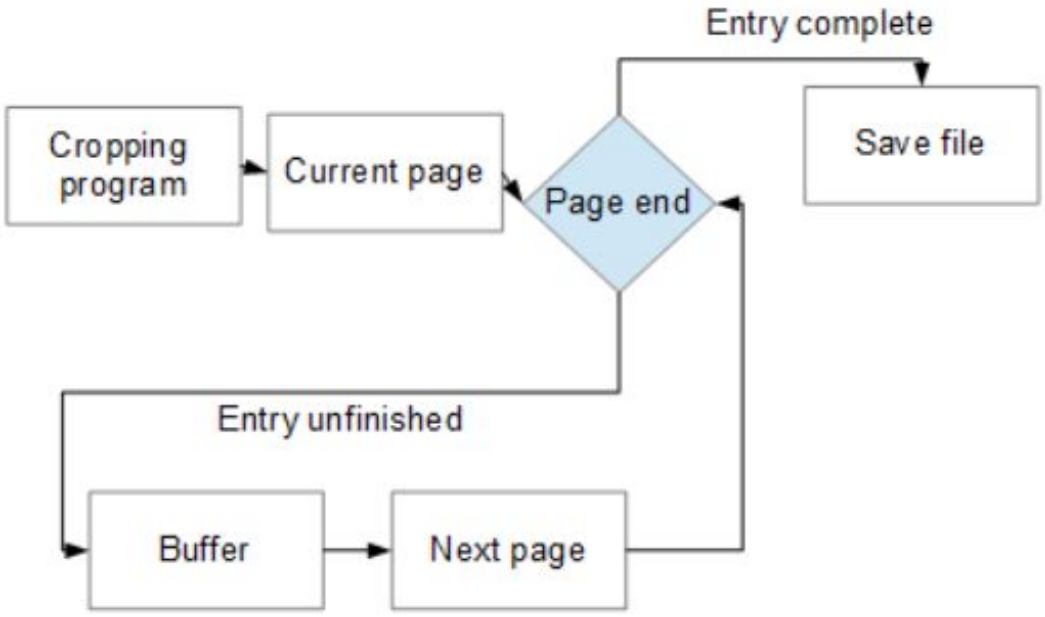
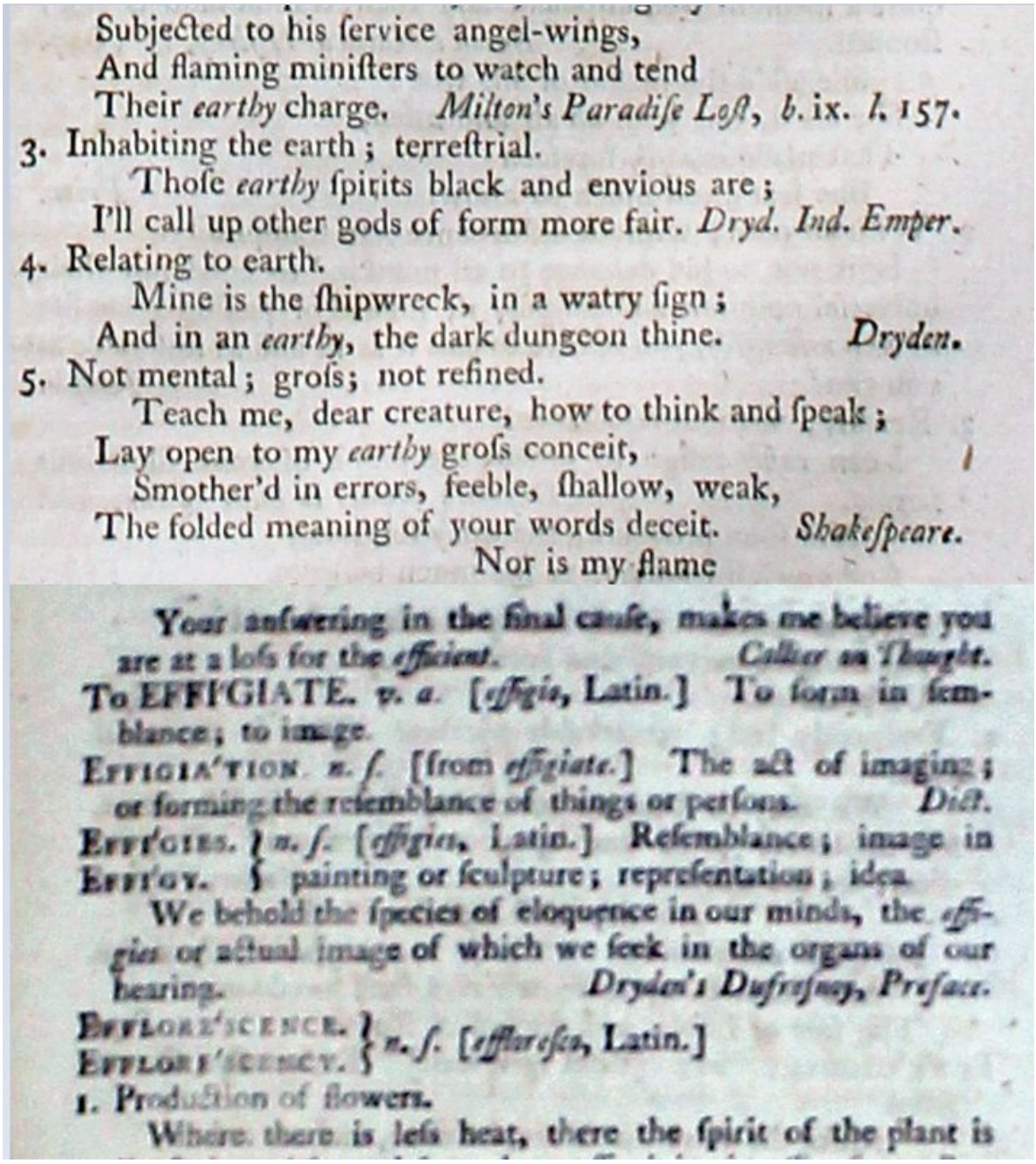


Figure 71: Image stitching flowchart

The actual difficulty in stitching two images together is dependent on how natural the stitch must appear. Simply placing one image on top of another is quite easy. Removing all white space so that the two images appear as one, like a new line rather than a new page, is not so simple. The pages are also not all uniform, one page will look darker or brighter or even blurrier than another page. Indeed, some pages look so different from the page preceding them that they

may be mistaken as being from an entirely different book, were it not for the distinct format used in the dictionary to denote entries and examples.

This can be easily shown:



Subjected to his ſervice angel-wings,
And flaming miniſters to watch and tend
Their *earthy* charge. *Milton's Paradiſe Loſt*, *b.* ix. *l.* 157.

3. Inhabiting the earth; terreſtrial.
Thoſe *earthy* ſpirits black and envious are;
I'll call up other gods of form more fair. *Dryd. Ind. Emper.*

4. Relating to earth.
Mine is the ſhipwreck, in a watry ſign;
And in an *earthy*, the dark dungeon thine. *Dryden.*

5. Not mental; groſs; not refined.
Teach me, dear creature, how to think and ſpeak;
Lay open to my *earthy* groſs conceit,
Smother'd in errors, feeble, ſhallow, weak,
The folded meaning of your words deceit. *Shakeſpeare.*
Nor is my flame

Your anſwering in the final cauſe, makes me believe you are at a loſs for the *efficient*. *Collier on Thought.*

To EFFIGIATE. *v. a.* [*effigio*, Latin.] To form in ſemblance; to image.

EFFIGIA'TION. *n. ſ.* [from *effigiate*.] The act of imaging; or forming the reſemblance of things or perſons. *Dict.*

EFFI'GIES. } *n. ſ.* [*effigies*, Latin.] Reſemblance; image in
EFFI'GY. } painting or ſculpture; repreſentation; idea.
We behold the ſpecies of eloquence in our minds, the *effigies* or actual image of which we ſeek in the organs of our hearing. *Dryden's Dufreſnoy, Preface.*

EFFLORE'SCENCE. }
EFFLORE'SCENCY. } *n. ſ.* [*effloreſco*, Latin.]

1. Production of flowers.
Where there is leſs heat, there the ſpirit of the plant is

Figure 72: Two entries stitched together

These two pages are drastically different and there is a stark contrast at the point they are stitched together. However, this difference is a natural result of the text's age and does not actually do much to hurt readability, as such removing it is not a high priority.

They are also all skewed, something our program will also adjust. But if two pages that skew in different directions are stitched together, even that might not be enough to keep the stitching unnoticed. Our image brightness and sharpening programs will take care of differences between page shades to a good degree, but it is unlikely to remove all signs. This may be mitigated by the fact that while pages are not uniform between them, they are also not wholly uniform in themselves. The varying nature of a single page means that the change in background between two pages may blend in if gradual enough. It is vital that no image stitching is applied before our program has rotated the image to level out any skew from scanner error. In the worst case, image stitching done before may cut out text. This is fairly unlikely even with the excessive skew and warping present on several of the pages in the text however. To help remove the stitching it makes the most sense for all image manipulation except for skew adjustment to be done after stitching. This ensures that the two images are adjusted to similar degrees

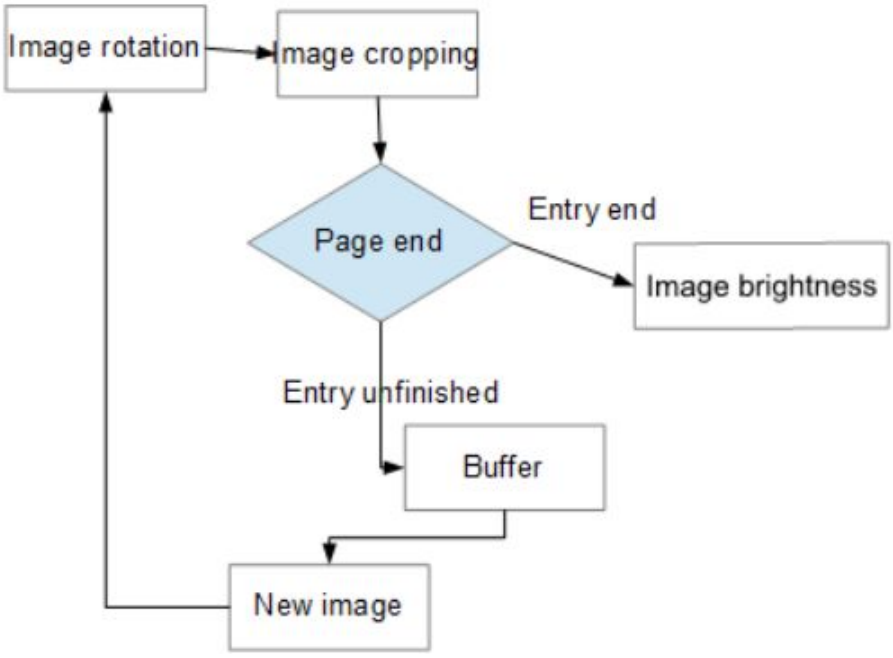


Figure 73: Stitching and image brightness flowchart

5.11 Graphical User Interface

The Graphical User Interface's job will be to make using the program easier to use and more intuitive, while still maintaining all program capabilities and functions. The first version of the Graphical User Interface is shown below.

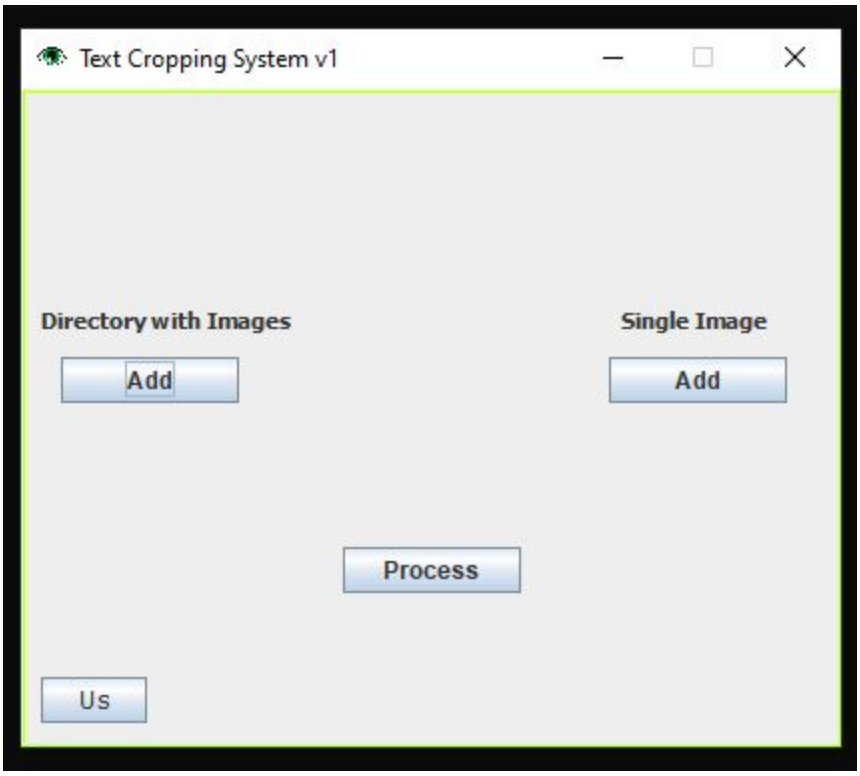


Figure 74: Main GUI panel

There will be a total of four buttons for the first version of the Graphical User Interface. The features currently are:

- Directory with Images
- Single Image
- Process
- Us

As the project progresses more features will be added to enhance the users' experience and reflect the new features. The following will be a demonstration of the Graphical User Interface along with its features.

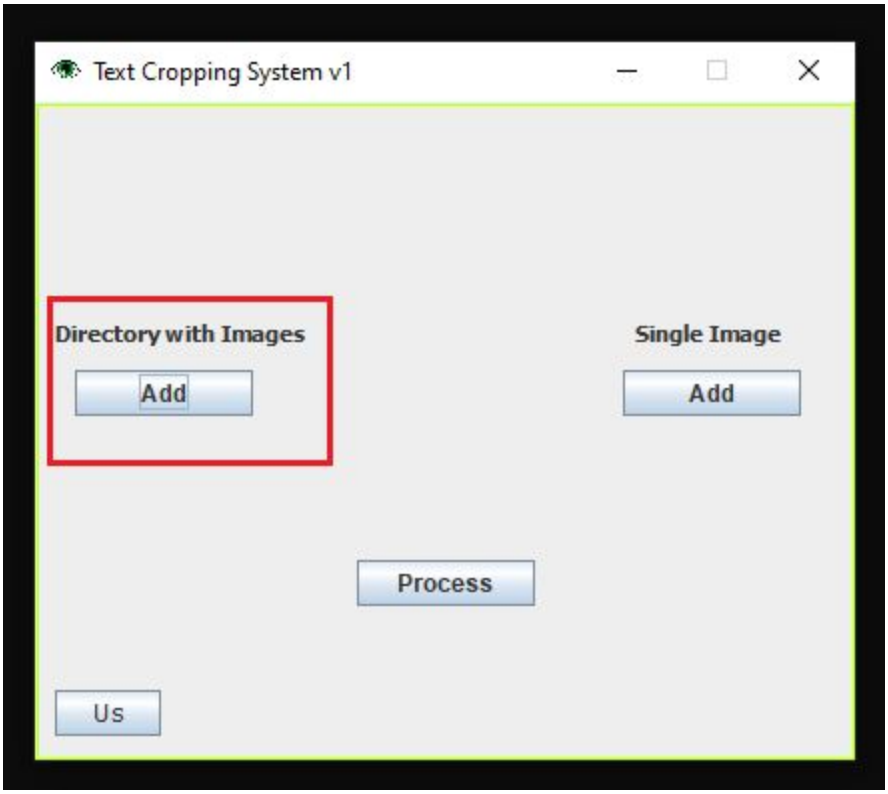


Figure 75: Selecting an image directory

This button will allow users to select a directory and subsequently enter a directory will the images that would need to get processed. We are still in the process of expanding the accepted file formats, from exclusively .tiff, to other file formats. Once the directory is selected, the user would then press the button labeled “Process” in order to start processing all the files that are currently within the previously chosen directory.

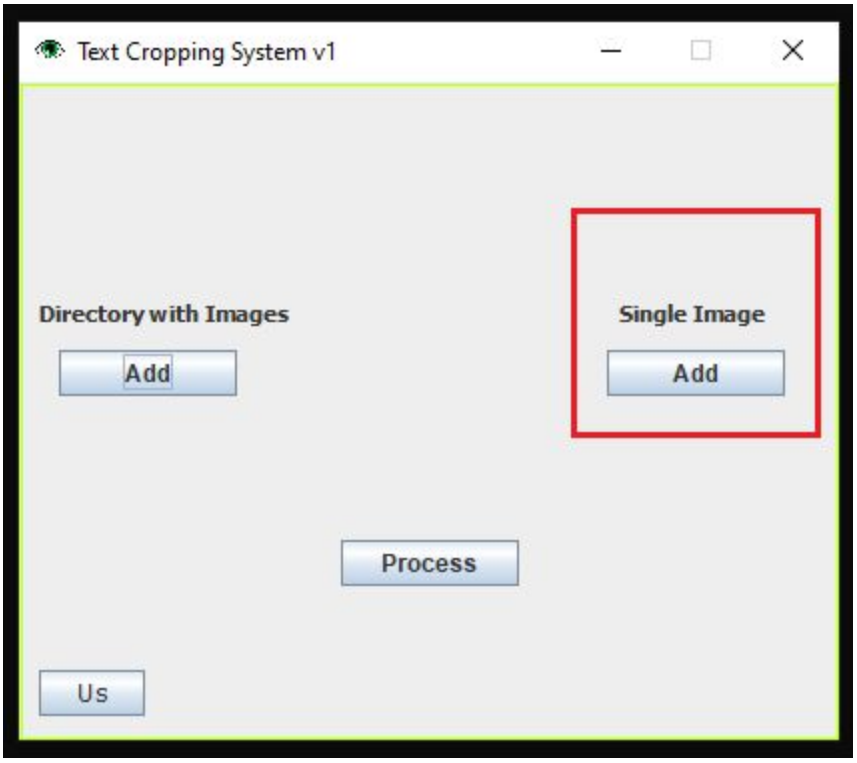


Figure 76: Adding a single image

The next chosen field, encased in a red box, will allow the user to select a single image file for later processing. The Single Image add button will only support a single image file. The user will not be able to upload more than one file through this medium and if a user selects multiple files, the program will only choose the first choice selected prior to the second choice. The program will also disregard any other additions that the user might try to make to upload more files through the “Single Image” medium. They will instead be prompted to load those images using the “Directory with Images” button. This will keep users from accidentally uploading more images than the selected medium was intended to handle.

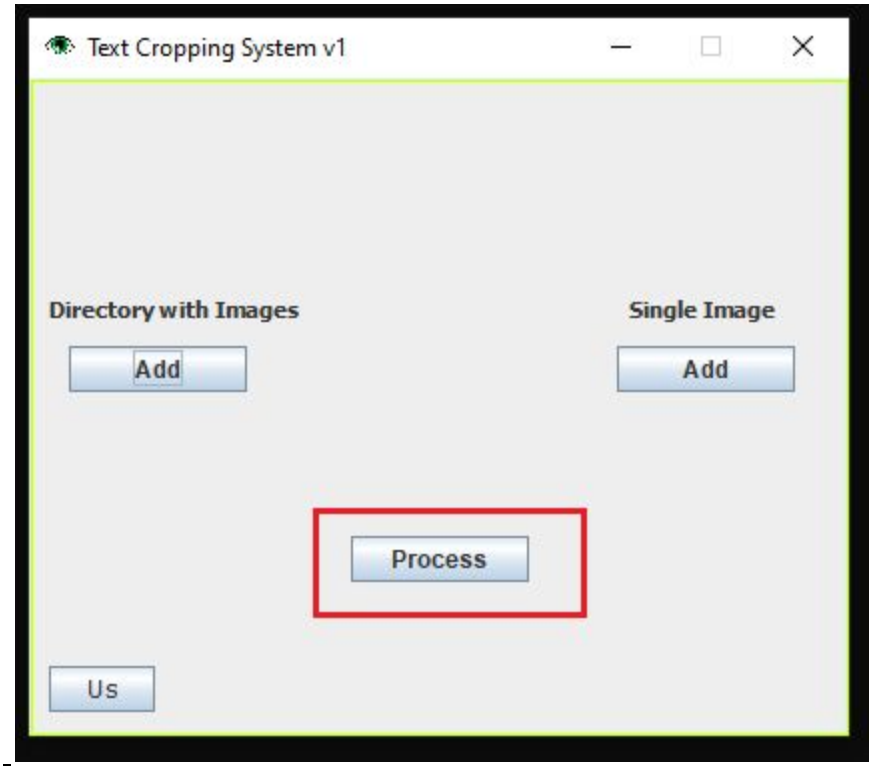


Figure 77: Processing the input

Once the “Process” button is pressed, the user will be prompted with a confirmation box showing if they would like to continue with the processing of the file(s). This confirmation box will contain an estimate of how long it would take to process the data that has been appended to it. And will subsequently ask the user if they would like to continue. Depending on whether the user chooses to proceed with the processing or deciding not to proceed, the user will either be redirected towards the main screen, if declining, or the program will start to process the images and a progress bar of the completed work will be shown.

Once the data is fully processed then the user will again be prompted indicating that the data has been successfully processed and can be viewed. If there was an issue while processing the data, the user will be prompted with the error and the reason the program was not able to successfully process the information.

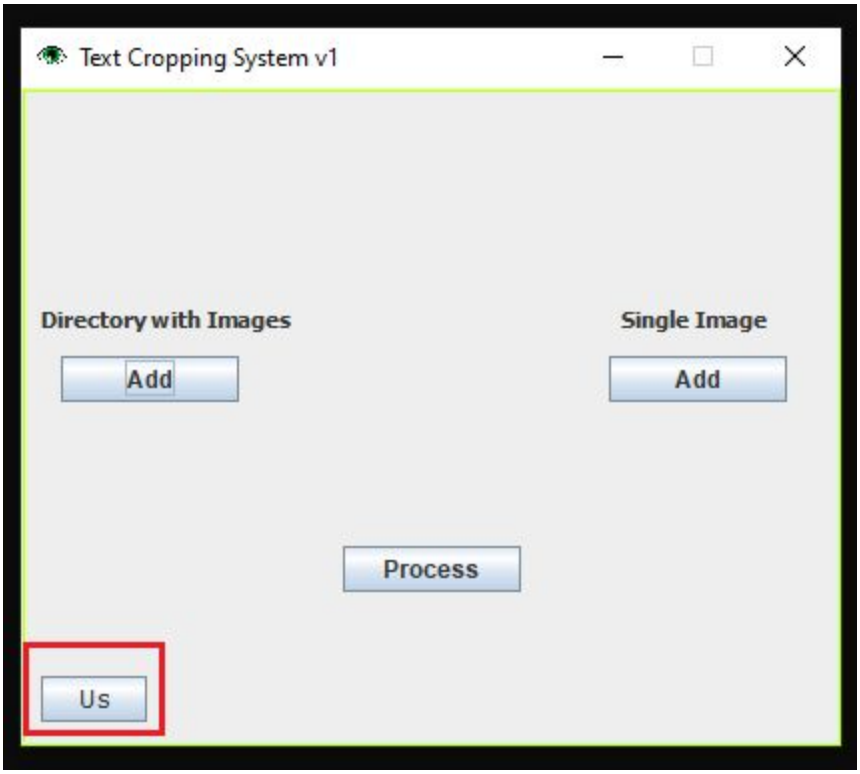


Figure 78: Displaying the group members

The final button on the Graphical User Interface is the “Us” button. When this button is pressed, it will show the team members of this project and their respective contact information. As illustrated below,

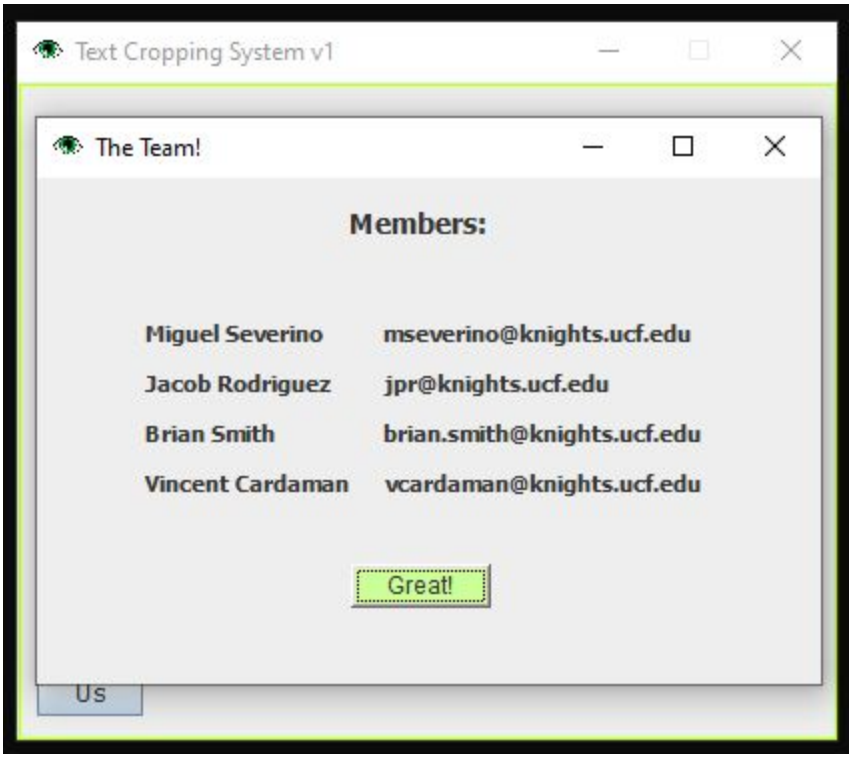


Figure 79: Group member information

This would conclude the first version of the Graphical User Interface.

5.12 Docker Implementation

To ensure portability and reusability of our solution, we are implementing the use of Docker in lieu of a virtual machine. Docker is a virtualization that contains the necessary software, dependencies, libraries, and configurations that allow for our program to be used anywhere. The implementation is very simple, we will create a Dockerfile that contains the necessary dependencies and actions to get the solution running on any machine with some simple command line commands. In addition to our Docker implementation, we will have a Github/Gitlab repository with our code available.

There was not much time spent deciding between whether to use a virtual machine or to use a docker container. Since we knew our solution would not need to emulate an entire system, we knew that a virtual machine would be unnecessary. We wanted the solution to be portable and since a virtual machine needs to emulate an entire operating system and its underlying hardware, it would be overkill. A container on the other hand simply virtualizes the operating system it was created in and runs on top of your system's host operating system. It depends on the applications, binaries, and libraries already installed on the host operating system. This allows for the container to be only a few megabytes in size and allows for it to run in seconds, whereas a virtual machine can be multiple gigabytes and take minutes to start. Figure 80 below displays the differences between a virtual machine and a Docker container.

The process for running the container consists of a few steps, first when the Dockerfile is run using the `docker build` [7] command, an image is created. A docker image is simply a file that is used to execute the code within the Docker container and to be a go-between for other Docker functions. It is essentially a record of a Docker container at a certain point in time, in our case, the finished product. One helpful characteristic of the Docker image is that it is immutable, so if in the future our group or someone else decided to continue making changes, our first final project would always be there to fall back on. It allows for intermittent backups of our work at different points in time. Once this image is created, it can be run using the `docker run` [7] command which then creates a new container and starts the container. This container being created from our image is just an instance of our image.

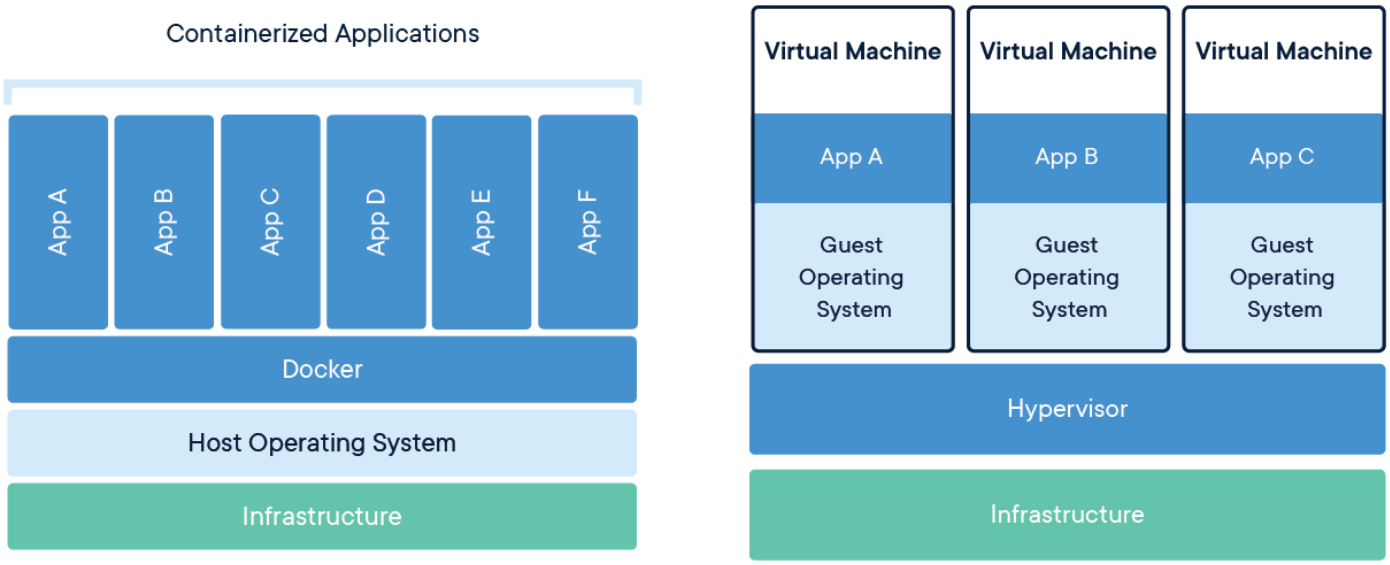


Figure 80: Differences between Docker container and VM

For our solution, our Dockerfile will only have a few steps, which are as follows:

- Pull from repo
    - Codebase
    - Documentation
- Install Python (if not already installed)
- Download and install dependencies
    - Numpy
    - Scipy
    - OpenCV
    - Matplotlib

6. Administrative Content

This section contains various administrative content regarding the project. It outlines our budget and financing for the project as well as our milestones for both semesters.

6.1 Budget and Financing

Open-source software will be used in the implementation of this software. By the very nature of open source software this means no costs will be incurred with our current plans for software development. The site on which this information will be placed is already made and maintained by the sponsor, its costs are outside the purview of the project. All the needed databases of information that our team will need for training our program is also readily available from the work done on the dictionaries manually. Once our software has been trained it’s processing requirements will be minimal, during training we will request use of the hardware in the Senior Design lab to complete training in a timely manner. Overall, we will likely not need to spend any amount of money throughout the completion of this project.

6.2 Project Milestones

This section describes the milestones and deadlines throughout Senior Design 1 (Spring) and Senior Design 2 (Summer).

Number	Task	Start Date	End Date
Senior Design 1			
1	Project Selection	1/8/2020	1/22/2020
2	Meet with Sponsor	2/3/2020	2/3/2020
3	Assign roles	2/3/2020	2/3/2020
4	Research languages	2/4/2020	2/19/2020
5	Research libraries	2/4/2020	2/19/2020
6	Research Docker	2/4/2020	2/19/2020
7	Setup Github	2/5/2020	2/5/2020
8	TA Check In 1	2/7/2020	2/7/2020
9	Initial Project Proposal	2/3/2020	2/12/2020
10	Draft Final Document	2/3/2020	4/20/2020
11	Setup skeleton code for first prototype	2/10/2020	2/12/2020
12	Project Status/Doc Review	2/24/2020	3/24/2020
13	Team progress report	3/8/2020	3/8/2020
14	Spring Break	3/9/2020	3/14/2020
15	TA Check In 2	3/30/2020	3/30/2020
16	Start second prototype	2/28/2020	3/24/2020

17	Finish and test second prototype	3/24/2020	4/2/2020
18	Finish document and proofread	4/20/2020	4/21/2020
19	Submit Final Document	2/3/2020	4/21/2020
Senior Design 2			
1	First prototype	2/12/2020	2/24/2020
2	Second prototype	2/28/2020	4/2/2020
3	Start final iteration	4/2/2020	6/20/2020
3	Break between Spring and Summer	4/24/2020	5/11/2020
4	Initial SD2 group meeting	5/11/2020	5/11/2020
5	CDR	3/30/2020	When Due
6	Project Completion	2/12/2020	6/20/2020
7	Full test of finished project	6/20/2020	7/28/2020
8	Final Presentation	4/21/2020	7/29/2020

7. Project Summary and Conclusions

This section provides an overall summary from each group member as well as conclusions about the overall project, its scope, and how we accomplished it all.

7.1 Personal Difficulties

In this section each member of our group will detail their own personal difficulties that were prevalent throughout the course of this project. We will talk about our difficulties going into the project, difficulties while working on the project, and difficulties for the group as a whole. Our group certainly was not without any difficulties, but with each problem faced we conquered together as a whole.

7.1.1 Jacob Rodriguez

Beginning this project was difficult for me because I did not have any experience with both machine learning and computer vision. I regret not taking a machine learning or computer vision class, as the topic is very interesting to me and it would have made this project significantly easier. The most difficult part of every project for me is finding out exactly where to start, this was especially difficult for this project due to my lack of overall experience in the area. Luckily, this is where my team helped me out a lot. They did not have much experience either, but their advice and guidance early on was extremely useful for me.

Partner programming also helped a lot early on in the project, as having another set of eyes going over your work can be vital when it comes to checking for errors. The most difficult aspect of the project was having to learn everything about computer vision and machine learning for the first time. I went into the project knowing absolutely nothing about computer vision or machine learning, and they are difficult topics to have to study and learn without a teacher.

Another difficulty that I faced with this project was the sheer amount of material and knowledge required for the design document. The document proved to be a fairly large issue for me, and I was certainly faced with difficulties throughout it. Having a lot of material for the document required an extensive amount of knowledge on the topic, which was gained over time throughout performing the project. Deciding on charts and images to include in the document is a difficult process as well. The images must convey a lot of information for

each, and that requires carefully choosing what exactly to include on each image, and whether or not to write on the image itself. For example, for many of the images of the pages of the dictionaries the examples that are being referred to in the respective sections are circled or highlighted in the image of the page itself. This helps convey exactly what must be shown on the page, and allows the use of pointing out specific details for multiple different topics all on the same dictionary page. Careful discretion when choosing the images to display and how to mark them was certainly implemented.

Another difficulty I faced with this project is working with a team. My team was amazing, everyone put out a consistent amount of effort and work, but communication is a difficult aspect of group work that needs constant attention. As the semester continued our communication dwindled a bit and dealing with that is par for the course for any group project. Discord helped greatly with our communication but balancing this class with all of my other classes has made me put the class off a bit. This is also due to the fact that we have another semester to work on the project. I am certain this must lead to many people throughout the entire class putting off a lot of the project for work during the next semester through Senior Design 2. My group members seemed fairly good about doing a large amount of work during Senior Design 1 but this was absolutely a huge difficulty for me.

7.1.2 Brian Smith

The most difficult part of the project for me was assembling the document. Since it is impossible to write a technical document without any technical knowledge, I had to perform extensive research of not only the ideas and concepts behind the text dewarping process, but also computer vision in general. Having to learn the specifics of OpenCV and Python and the technical and mathematical aspects of the text dewarping algorithm itself while also taking other classes was very intensive and mentally draining. We all came into this project with essentially no computer vision experience, everything we have learned and implemented has been through trial and error and hours of research. Some of these struggles were alleviated through communicating with my team and sharing ideas and solutions. By having our regular meetings and discussing how we think we should be solving each issue and working together step by step, we were able to finally come to a conclusion, but it took a lot of work.

My portion of the project specifically was very math intensive and while I do have a strong background in mathematics, when dealing with computer vision, the levels of mathematics are outside of my realm of understanding. Due to this I struggled immensely when starting off and was lost on where to begin. When dealing with image manipulations of this nature there are a lot of remapping and transformation operations. While these methods are implemented within OpenCV, I would have to first have the knowledge on how to use them, which I did not. This led me to start researching to see if any open source solutions already existed and luckily I found several. Once I found these however the struggles were not over. Many of the proposed methods were research based which meant they were not implemented and the ones that were implemented only worked on certain input sets. After many hours of searching I finally found our current solution and was able to modify it to fit our needs.

Outside of the project, I had a big issue managing my time properly as the semester progressed. I am taking four other classes as well as interning part time so I am constantly busy and have many things to keep track of. We were having frequent meetings and these helped immensely, I was able to be aware of everyone else’s progress and it motivated me to keep making progress on my own despite the fact that I had so much going on. My problems with time management only got worse due to the COVID19 pandemic. Once school was closed we were no longer able to meet and everything transitioned to online. This rough transition to online classes affected all of my classes and caused my

academic life to decay. Without the need to show up or have a professor stand in front of me and lecture, my motivation had been entirely depleted. Since we were unable to meet up, I felt like I no longer had that pressure or drive to continue making progress. We continued to communicate on Discord and we have been very successful in updating each other and making sure that everyone is making the progress that they need to be making, but it was not the same.

Group work is always a difficult thing to overcome and our group was not without any issues. In the beginning, we were all motivated and excited to work on this project. We all shared an interest in computer vision and machine learning so we wanted to hit the ground running. The momentum started off very strong but started to dwindle as time went on. We are all busy and are taking multiple classes so distributing time between all of our classes was hard for all of us.

7.1.3 Miguel Severino

There were many challenges during the research phase of this project. One of the challenges was understanding how to perform proper research on advanced topics. These topics ranged from Machine Learning algorithms to specialized techniques in Computer Vision. Another challenge was that we needed to document all our steps along the way and finally combine all our individual research into a single document. I feel that the biggest challenge that we faced was all the turmoil that COVID-19 caused. We were used to meetings in person and having resources on campus that we frequently used, but that changed as COVID-19 began to spread globally and forced our learning to an online platform. That alone presented many challenges since now we needed to perform all our functions from a distance without the many resources that are on campus.

7.1.4 Vincent Cardaman

My first great difficulty with this project was in deciding which of the myriad of options for image manipulation to use. I was assigned to the stretch goal of image enhancement fairly early on. I spent a large amount of time researching the various options available, a fact that is quite obvious in this document. There are several different kinds of image manipulation that I researched enough to

justify putting them in simply to explain why they would not be the best choice for our project. I don't think this was an inherently negative decision, researching our options builds a good foundation for the project to work on and this first semester is meant to be focused more on the design than the coding. But the vast majority of my time spent this semester was on this. I definitely think I should have started this semester off learning the ropes of our chosen software more before moving onto trying to find the best method for the parts of the project that were my focus. The rest of my team has been quite helpful in clearing any confusion I had on these aspects as they focused more on learning the base software far more then I did.

None of this was helped by the COVID 19 pandemic, which started soon after there was a death in my family, when most of my close relatives were still in transit back from the funeral. Losing the meetings we used to have on campus to review our progress definitely hurt our groups' cohesion. But Discord proved invaluable in keeping us in contact throughout the semester, both before and after the college switched to remote classes. This switch was certainly an unusual change from what I was used to, the majority of my classes have been on campus and having all our classes online has been odd to adjust to. Overall I think I should have done more of the ground work early on, trying to do most of those things around spring break while juggling everything else that happened in the world and in my own personal life around that time was quite difficult. But it could have been solved simply by frontloading my work earlier on in the semester. I definitely have let too much of my focus this semester be on the design then I should have. Even for a senior design class, this is still computer science and I should have put more time on coding then I have for this project. Something I plan to fix both during the early semester and in the break between this semester and the next. While there have been many unexpected obstacles in this semester that I don't think there is any way we could have anticipated, I think both myself and my group have handled them admirably and made good work on our project.

7.2 Conclusion

After spending the entire semester working on this project and finally being able to step back and look at it in its entirety, we believe that this solution will be immensely helpful in the process of preserving history. We wanted to provide a simple and portable solution to the issue presented by Dr. Giroux that could easily be scaled to any medium. Researchers who are interested in document preservation will be able to easily document their findings with just the click of a button. This is not only a solution for researchers as stated in the Section 2.2, in addition to what is mentioned there, this can be used by any user that needs to digitize and preserve documents or records of any nature.

Since the beginning of the semester we have been working tirelessly to analyze each step in the process, brainstorming solutions, figuring out what we don't know and finding the resources necessary to obtain that knowledge, learning, and constantly tweaking our solution to be the best it can be. Despite the erratic nature of this semester and everyone having to transition to online and all of the hiccups and oddities that come from that, we are motivated to continue into Senior Design 2 even stronger than we were during this semester. We will not stop researching, designing, re-designing, and improving our solution. We still have much to learn about computer vision and machine learning and we will continue to learn and struggle until we have a product we are truly proud of.

While we did have many successes, there were many hardships along the way, as Dr. Heinrich and Dr. Leinecker mentioned, putting everything off until the last minute is not an option. This class is built around consistently making progress and moving forward at a steady pace. For the most part our group was good about making slow, steady progress, but on a few occasions there were some close calls which we would like to avoid in the coming semester. The communication with our sponsor is something else we would like to work on. We were not able to meet as often as we would have liked to and again with the pandemic being an unprecedented experience, it made having those meetings even harder. Senior Design 2 will be our only class during the summer semester and while it is again online, we have slowly adapted to this new medium of class presentation and know exactly how to succeed. Our weekly meetings will continue and possibly increase since we are going to be constantly implementing and tweaking, we will need more interaction and feedback from each other to ensure that we are going in the right direction. In addition to that, we will meet

with the sponsor weekly or bi-weekly and keep her updated with every advancement that we make.

8. References

[1] Ma, K., Shu, Z., Bai, X., Wang, J., & Samaras, D. (2018). DocUNet: Document Image Unwarping via a Stacked U-Net. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. doi: 10.1109/cvpr.2018.00494

[2] Document image dewarping. (n.d.). Retrieved from https://github.com/xellows1305/Document-Image-Dewarping

[3] Leptonica. (n.d.). Retrieved from http://www.leptonica.org/

[4] Hata, K., & Savarese, S. (n.d.). CS231A Course Notes 1: Camera Models. Retrieved from https://web.stanford.edu/class/cs231a/course_notes/01-camera-models.pdf

[5] Connected Components Labeling. (n.d.). Retrieved from https://homepages.inf.ed.ac.uk/rbf/HIPR2/label.htm

[6] Image Analysis: Morphological Operations. (n.d.). Retrieved from https://sites.ualberta.ca/~ccwj/teaching/image/morph/

[7] Weber, X. (2020, March 29). Implementing a Connected Component Labeling algorithm from scratch. Retrieved from https://towardsdatascience.com/implementing-a-connected-component-labeling-algorithm-from-scratch-94e1636554f

[8] Mildon, L. (2019, September 22). What the Dither? Dithering Used in Images and Graphics. Retrieved from https://www.lifewire.com/what-is-dithering-4686105

[9] Zucker, M. (2016, August 15). Page dewarping. Retrieved from https://mzucker.github.io/2016/08/15/page-dewarping.html

[10] Eroding and Dilating. (n.d.). Retrieved from https://docs.opencv.org/2.4/doc/tutorials/imgproc/erosion_dilatation/erosion_dilatation.html

[11] Camera Calibration and 3D Reconstruction. (n.d.). Retrieved from https://docs.opencv.org/2.4/modules/calib3d/doc/camera_calibration_and_3d_reconstruction.html#solvepnp

[12] Camera Calibration and 3D Reconstruction. (n.d.). Retrieved from https://docs.opencv.org/2.4/modules/calib3d/doc/camera_calibration_and_3d_reconstruction.html#projectpoints

[13] scipy.optimize.minimize. (n.d.). Retrieved from https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.minimize.html

[14] Baek, J. (2016, August 16). Fast Document Rectification and Enhancement. Retrieved from Fast Document Rectification and Enhancement

[15] Stamatopoulos, N., Gatos, B., & Kesidis, A. (n.d.). Automatic Borders Detection of Camera Document Images .

[16] Reference documentation (n.d.) Retrieved from https://docs.docker.com/engine/reference

[17] What is a Container? (n.d.). Retrieved from https://www.docker.com/resources/what-container

[18] J. Canny, "A Computational Approach to Edge Detection," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-8, no. 6, pp. 679-698, Nov. 1986.

[19] H. Dewen, W. Xichang, L. Jiang and C. Xia, "Text Line Detection Based on Mutation Analysis," 2013 International Conference on Computational and Information Sciences, Shiyang, 2013, pp. 678-681.

[20] Wenshuo Gao, Xiaoguang Zhang, Lei Yang and Huizhong Liu, "An improved Sobel edge detection," 2010 3rd International Conference on Computer Science and Information Technology, Chengdu, 2010, pp. 67-71.

[21] K. Terasawa, T. Nagasaki and T. Kawashima, "Eigenspace method for text retrieval in historical document images," Eighth International Conference on Document Analysis and Recognition (ICDAR'05), Seoul, South Korea, 2005, pp. 437-441 Vol. 1.

[22] Zheng Zhang and Chew Lim Tan, "Straightening warped text lines using polynomial regression," Proceedings. International Conference on Image Processing, Rochester, NY, USA, 2002, pp. 977-980 vol.3.

[23] N. Stamatopoulos, B. Gatos and I. Pratikakis, "A Methodology for Document Image Dewarping Techniques Performance Evaluation," 2009 10th International Conference on Document Analysis and Recognition, Barcelona, 2009, pp. 956-960.

[24] P. Bao, Lei Zhang and Xiaolin Wu, "Canny edge detection enhancement by scale multiplication," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 9, pp. 1485-1490, Sept. 2005.

[25] S. Marinai, M. Gori and G. Soda, "Artificial neural networks for document analysis and recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 1, pp. 23-35, Jan. 2005.

[26] B. Gatos, I. Pratikakis and K. Ntirogiannis, "Segmentation Based Recovery of Arbitrarily Warped Document Images," Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Parana, 2007, pp. 989-993.

[27] L. Zhang and C. Tan, "Warped Document Image Restoration Using Shape-from-Shading and Physically-Based Modeling," 2007 IEEE Workshop on Applications of Computer Vision (WACV '07), Austin, TX, 2007, pp. 29-29.

[28] F. Zirari, A. Ennaji, S. Nicolas, and D. Mammass, "A Document Image Segmentation System Using Analysis of Connected Components," 2013 12th International Conference on Document Analysis and Recognition, Washington, DC, 2013, pp. 753-757.

[29] Xuhong Li and P. A. Ng, "A document classification and extraction system with learning ability," Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR '99 (Cat. No.PR00318), Bangalore, India, 1999, pp. 197-200.

[30] A. C. Özgen, M. Fasounaki and H. K. Ekenel, "Text detection in natural and computer-generated images," 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, 2018, pp. 1-4.

[31] H. Hu, I. Misra and L. van der Maaten, "Evaluating Text-to-Image Matching using Binary Image Selection (BISON)," 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South), 2019, pp. 1887-1890.

[32] D. Kumar and R. Singh, "Deep Learning Approach: A New Trend in Text Detection in Natural Images," 2018 4th International Conference on Computing Sciences (ICCS), Jalandhar, 2018, pp. 126-131.

[33] H. Jiang, T. Gonnot, W. Yi and J. Saniie, "Computer vision and text recognition for assisting visually impaired people using Android smartphone," 2017 IEEE International Conference on Electro Information Technology (EIT), Lincoln, NE, 2017, pp. 350-353.

[34] A. Ganjoo and A. Dhyani, "Text recognition on PAN card using template matching method," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, 2017, pp. 1-5.

[35] M. Opitz, M. Diem, S. Fiel, F. Kleber and R. Sablatnig, "End-to-End Text Recognition Using Local Ternary Patterns, MSER and Deep Convolutional Nets," 2014 11th IAPR International Workshop on Document Analysis Systems, Tours, 2014, pp. 186-190.

[36] S. Chowdhury, S. Mandal, A. Das and B. Chanda, "Segmentation of Text and Graphics from Document Images," Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Parana, 2007, pp. 619-623.

[37] O. Oksuz, U. Gudukbay and E. Cetin, "Computer vision-based text and equation editor for LATEX," 2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763), Taipei, 2004, pp. 1255-1258 Vol.2.

[38] T. Lelore and F. Bouchara, "Document Image Binarisation Using Markov Field Model," 2009 10th International Conference on Document Analysis and Recognition, Barcelona, 2009, pp. 551-555.

[39] S. Sural and P. K. Das, "A two-step algorithm and its parallelization for the generation of minimum containing rectangles for document image segmentation," Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR '99 (Cat. No.PR00318), Bangalore, India, 1999, pp. 173-176.

[40] F. Zirari, D. Mammass and A. Ennaji, "Text/image separation in document images based on statistical analysis of texture and morphological operations," 2011 International Conference on Multimedia Computing and Systems, Ouarzazate, 2011, pp. 1-5.

[41] C. Yang, X. Yin, H. Yu, D. Karatzas and Y. Cao, "ICDAR2017 Robust Reading Challenge on Text Extraction from Biomedical Literature Figures (DeTEXT)," 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, 2017, pp. 1444-1447.

[42] P. Hu, W. Wang and K. Lu, "A novel binarization approach for text in images," 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, 2015, pp. 956-960.'-

[43] R. Zhu, X. Mao, Q. Zhu, N. Li and Y. Yang, "Text detection based on convolutional neural networks with spatial pyramid pooling," 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, 2016, pp. 1032-1036.

[44] S. Roy, G. Adhikari, T. Dasgupta and T. Pradhan, "An Adaptive Warp Correction Algorithm for Handwritten Text Images with Non-Linear Baselines," 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Bangalore, 2018, pp. 1-7.

[45] L. Zhang and C. L. Tan, "Warped image restoration with applications to digital libraries," Eighth International Conference on Document Analysis and Recognition (ICDAR'05), Seoul, South Korea, 2005, pp. 192-196 Vol. 1.

[46] A. R. Ghods, S. Mozaffari and F. Ahmadpanahi, "Document image dewarping using Kinect depth sensor," 2013 21st Iranian Conference on Electrical Engineering (ICEE), Mashhad, 2013, pp. 1-6.

[47] M. Shamqoli and H. Khosravi, "Warped document restoration by recovering shape of the surface," 2013 8th Iranian Conference on Machine Vision and Image Processing (MVIP), Zanjan, 2013, pp. 262-265.

[48] N. Stamatopoulos, B. Gatos, I. Pratikakis and S. J. Perantonis, "A Two-Step Dewarping of Camera Document Images," 2008 The Eighth IAPR International Workshop on Document Analysis Systems, Nara, 2008, pp. 209-216.

[49] H. C. Vinod and S. K. Niranjan, "De-warping of camera captured document images," 2017 IEEE International Symposium on Consumer Electronics (ISCE), Kuala Lumpur, 2017, pp. 13-18.

[50] D. Oliveira, R. Lins, G. Torreão, J. Fan and M. Thielo, "An Efficient Algorithm for Segmenting Warped Text-Lines in Document Images," 2013 12th International Conference on Document Analysis and Recognition, Washington, DC, 2013, pp. 250-254.

[51] D. Sutapirat, K. Sookhanaphibarn and C. Lursinsap, "Model-Based Book Dewarping Method for Content-Independent Document Images," 2009 International Conference on Digital Image Processing, Bangkok, 2009, pp. 190-194.

[52] H. Ezaki, S. Uchida, A. Asano and H. Sakoe, "Dewarping of document image by global optimization," Eighth International Conference on Document Analysis and Recognition (ICDAR'05), Seoul, South Korea, 2005, pp. 302-306 Vol. 1.

[53] F. Guo, Y. Li and P. Liu, "A Fast Page Outline Detection and Dewarping Method Based on Iterative Cut and Adaptive Coordinate Transform," 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), Sydney, Australia, 2019, pp. 1-6.

[54] M. Shamqoli and H. Khosravi, "Warped document restoration by recovering shape of the surface," 2013 8th Iranian Conference on Machine Vision and Image Processing (MVIP), Zanjan, 2013, pp. 262-265.

[55] S. Shetty, A. S. Devadiga, S. S. Chakkaravarthy and K. A. V. Kumar, "Ote-OCR based text recognition and extraction from video frames," 2014 IEEE 8th International Conference on Intelligent Systems and Control (ISCO), Coimbatore, 2014, pp. 229-232.

[56] A. Kumar, "An efficient text extraction algorithm in complex images," 2013 Sixth International Conference on Contemporary Computing (IC3), Noida, 2013, pp. 6-12.

[57] J. Zhang and R. Kasturi, "Text Detection Using Edge Gradient and Graph Spectrum," 2010 20th International Conference on Pattern Recognition, Istanbul, 2010, pp. 3979-3982.

[58] Z. Zhang, C. L. Tan, "Correcting document image warning based on regression of curved text lines," Proc. ICDAR, pp. 589-593, 2003.

[59] M. A. B. Siddique, R. B. Arif and M. M. R. Khan, "Digital Image Segmentation in Matlab: A Brief Study on OTSU's Image Thresholding," 2018 International Conference on Innovation in Engineering and Technology (ICIET), Dhaka, Bangladesh, 2018, pp. 1-5.

[60] S. Farid and F. Ahmed, "Application of Niblack's method on images," 2009 International Conference on Emerging Technologies, Islamabad, 2009, pp. 280-286.

[61] A. Sakila and S. Vijayarani, "A hybrid approach for document image binarization," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, 2017, pp. 645-650.

[62] Taekyung Kim and Joonki Paik, "Soft decision histogram-based image binarization for enhanced ID recognition," 2007 6th International Conference on Information, Communications & Signal Processing, Singapore, 2007, pp. 1-4.

[63] Bull, David R. "Digital Picture Formats and Representations.", *Communicating Pictures*, 2014, pages 99–132.

[64] Cao, Gang, et al. "Contrast Enhancement of Brightness-Distorted Images by Improved Adaptive Gamma Correction.", *Computers & Electrical Engineering*, vol. 66, 2018, pages 569–582.

[65] Dey, Soumyadeep, et al. "Margin Noise Removal from Printed Document Images.", *Proceeding of the Workshop on Document Analysis and Recognition - DAR '12*, 2012.

[66] Mann, S., and M.a. Ali. “The Fundamental Basis of HDR.”, *High Dynamic Range Video*, 2016, pages 1–59.

[67] Tyagi, Vipin. “Introduction to Digital Image Processing.”, *Understanding Digital Image Processing*, 2018, pages 1–12.

[68] Tuba, M., et al. “Improved Weighted Thresholded Histogram Equalization Algorithm for Digital Image Contrast Enhancement Using the Bat Algorithm.”, *Bio-Inspired Computation and Applications in Image Processing*, 2016, pages 61–86.,

[69] Srinivas, Suraj, et al. “An Introduction to Deep Convolutional Neural Nets for Computer Vision.”, *Deep Learning for Medical Image Analysis*, 2017, pages 25–52.

[70] Ebersole, Mark. “Scalable Parallel Execution.”, *Programming Massively Parallel Processors*, 2017, pages 43–69

[71] Nixon, Mark S., and Alberto S. Aguado. “Basic Image Processing Operations.”, *Feature Extraction and Image Processing*, 2002, pages 67–97.

[72] Jourlin, M. “Metrics Based on Logarithmic Laws.”, *Advances in Imaging and Electron Physics Logarithmic Image Processing: Theory and Applications*, 2016, pages 61–113.

[73] Mahmoudpour, Saeed, and Manbae Kim. “A Study on the Relationship between Depth Map Quality and Stereoscopic Image Quality Using Upsampled Depth Maps☆.”, *Emerging Trends in Image Processing, Computer Vision and Pattern Recognition*, 2015, pages 149–160.

[74] Mortezaie, Z., Hassanpour, H., & S, A. A. (2019). An adaptive block based un-sharp masking for image quality enhancement., *Multimedia Tools and Applications*, pages 1-14.

[75] Whyte, Oliver, et al. “Non-Uniform Deblurring for Shaken Images.”, *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 168-186 2010.

[76] Mei, Jianhan, et al. “DeepDeblur: Text Image Recovery from Blur to Sharp.”, *Multimedia Tools and Applications*, vol. 78, no. 13, 2019, pages. 18869–18885.

[77] Ali, Usman, and Muhammad Mahmood. “Analysis of Blur Measure Operators for Single Image Blur Segmentation.”, *Applied Sciences*, vol. 8, no. 5, 2018, page 807.

[78] Zaharescu, M., Boiangiu, C., & Bucur, I. (2017). Blind Image Deblurring Using A Single-Image Source: The State Of The Art., *Journal of Information Systems & Operations Management*, pages 17-28.

[79] Wang, Jing, et al. “Kernel Optimization for Blind Motion Deblurring with Image Edge Prior.”, *Mathematical Problems in Engineering*, vol. 2012, 2012, pages 1–10.

[80] Tsai, Hsin-Che, and Jiunn-Lin Wu. “An Improved Adaptive Deconvolution Algorithm for Single Image Deblurring.”, *Mathematical Problems in Engineering*, vol. 2014, 2014, pages 1–11.

[81] Chang, Yen-Ching. “An Almost Automatic Image Fusion Scheme for Balancing Clarity and Visual Effects.”, *Multimedia Tools and Applications*, vol. 76, no. 23, 2017, pages. 25455–25476.

[82] Al-Ameen, Z., Al-Ameen, S., & Al-Othman, A. (2018). Improving the sharpness of digital images using a modified laplacian sharpening technique., *Iptek*, 29(2), pages 44-48.

[83] Ma, Chengcheng, et al. “Accurate Blind Deblurring Using Salientpatch-Based Prior for Large-Size Images.”, *Multimedia Tools and Applications*, vol. 77, no. 21, 2018, pp. 28077–28100., doi:10.1007/s11042-018-6009-2.

[84] Vasistareddy. (n.d.). Vasistareddy/pythonRLSA. Retrieved from https://github.com/Vasistareddy/pythonRLSA/tree/master/pythonRLSA/test_images