A COLLATION PANEL FOR HUMANISTS

An Undergraduate Research Scholars Thesis

by

RYAN CHRISTOPHER OLIVIERI

Submitted to Honors and Undergraduate Research
Texas A&M University
in partial fulfillment of the requirements for the designation as

UNDERGRADUATE RESEARCH SCHOLAR

Approved by
Research Advisor: Richard Furuta

May 2014

Major: Computer Engineering

TABLE OF CONTENTS

Page

ABSTRACT

A Collation Panel for Humanists. (May 2014)

Ryan Christopher Olivieri
Department of Computer Science and Engineering
Texas A&M University

Research Advisor: Dr. Richard Furuta
Department of Computer Science and Engineering

Comparing a set of similar texts, often called collation, gives a scholar insight into the author's thoughts and motives and into the culture of the time period. Multiple editions can be published by an author for a diverse number of reasons such as political pressure or to further character development. Collating a set of editions can help discover the author's motives, but involves methodically comparing text to find minute differences. This task, which is currently done by hand due to unstandardized printing practices in historic novels, is time consuming, and prone to human error due to fatigue. We propose a Virtual Hinman Collator that automates the comparison process and highlights the differences between a set of pages. This is done through an image processing approach that is less susceptible to the issues of optical character recognition on poor quality images. This tool saves the user an immense amount of time which allows more time to focus on analyzing the text and drawing conclusions instead of finding the differences. With this increased productivity, a better understanding of past cultures can be found through recognizing the motivation and thoughts behind the author's works.

DEDICATION

This thesis is dedicated to my parents, Lisa and Mark Olivieri, for their continued encouragement and inspiration.

ACKNOWLEDGMENTS

I would like to thank Dr. Richard Furuta for the opportunity to conduct my research. I would also like to thank Gaurav Kerjiwal, a graduate student, for his guidance and his knowledge of the project.

CHAPTER I
INTRODUCTION

Background

Humanists play an important role in society by understanding the culture of the past. Some humanists do this through creating witnesses and critical editions of novels. When an author is collating a set of novels, they go through the copies word by word looking for differences. Then once the differences are found notes on the differences are made. Once the whole novel is finished, notes are compiled into one copy. Finally, the work is published in a journal or peer reviewed outlet. Humanists are able to use their knowledge of the era and context to understand why a change was added. Then the decision can be made if it was due to political pressure or if the author wanted to change the view of a character to the reader, or a variety of other reasons. This information can help our generation have a better understanding of the past by better understanding the motives behind the authors novels.

Problem

The main issue with the task of creating a witness or critical edition of a novel is the time it takes to complete. A great amount of experience is needed to analyze the differences between editions of a novel. This leaves a limited number of people that are capable of coming to educated conclusion about the differences in the editions. With this in mind, it does not make sense to have these bright minds completing a time consuming and mechanical task of finding the differences. This mundane task exists, because printing standards in the 17th and 18th century were in their infancy compared to today. Currently, this process cannot be automated with OCR (Optical Character Recognition) because the margins are not consistent, the font sizes differ from edition to edition, some pages have splotches and other defects not present today. Also in some editions the words on the backside of the page bleed through. Further whole sentences can be omitted or altered between editions, which causes the words on each edition to be offset from each other. All of these factors and many others make this task difficult to automate, and this is precisely why many humanists today still use the traditional way of comparing printed out copies of each novel. To convince humanists to change to a new method it needs to be more efficient, easy, to use, accurate, and reliable. Current attempts have not been able to satisfy all of these requirements for widespread adoption.

Previous solutions

This problem has been addressed a multitude of times, but while each solution may solve a particular area each has limitations. The Hinmann Collator [2] was one of the first solutions to this problem and was invented in the 1940's. This machine works by having two books to be collated placed in the machine. Then the pages are superimposed over each other. The user only sees one copy of the page if they are exactly the same, and any differences between the pages shows up blurry. The Hinman collator is useful when comparing the same editions for typeface changes. This works to the Hinman Collator's strengths. When the editions have differences the Hinman Collator is not useful because of different font sizes, added text in a page, or other vast differences. These strengths and weaknesses make the Hinmann Collator great at a small subset of books, but not practical for collating the majority of novels.

Another attempt is with OCR. This technique takes the images of the books and converts them to a text file. Once the pages are made into text files, the comparison is made with commonly available text comparison tools. The problem comes when the image quality of the book is too poor to work properly. If this is the case the result of the OCR conversion is a useless text file and inaccurate results are given. Unfortunately, many of the novels written in an earlier time period suffer from early printing practices. Splotches and misaligned text can be troublesome for OCR to decipher correctly, which makes OCR not a valid option for many novels. OCR also strips the words from the page and correspondingly the font used, the colorations of the paper, and all of the other important aspects of the page.

These aspects give the humanists context and a glimpse into the life at the time the book was written. Currently, the technical issues that OCR is facing limit the usefulness of the approach because of the need of accurate results.

Another variation of this approach involves one person reading the novel to a person who copies the novel to a text document. This task still takes a large amount of time and can suffer from the inaccuracies of humans reading the novel out load. If a good text version of a set of novels is produced then text comparison tools can be used to quickly find the differences. Some novels are digitized through manually going through and typing the whole novel. This technique also suffers from human error and involves a large amount of work.

Our approach

Taking all of this information into consideration we propose a Virtual Hinman Collator (vHinman) to find the differences between the set of novels. This digital collator will work with facsimiles in order to preserve the pages of the books. Instead of using OCR to find the differences an image processing approach will be used that is less susceptible to many of the problems OCR is afflicted with. This tool will also be integrated into CritSpace [1] a web application that creates an environment specifically for humanist to analyze facsimiles of novels. A CritSpace workspace allows the user to move panels containing images of the novels, a text editor, and a filmstrip to select which images to view. This environment promotes the collaboration of the humanists with their peers by using a web platform. By integrating the collator into CritSpace it will benefit from the existing support environment as opposed to being a stand alone tool. This tool is able to accurately and conveniently find the differences between the pages selected by the user. This takes the most time consuming part of collation out of the equation and lets humanists spend their limited time analyzing differences and collating new novels instead of finding the differences hidden among thousands of words.

CHAPTER II
METHODOLOGY

Approach

Our tool uses an image processing approach similar to the work of Dr. Yalniz and Dr. Manmatha at the University of Massachusetts [6]. It involves taking an image of a novel and separating it into the different words. Currently, the image is separated into words with ABBYY FineReader, a commercial OCR engine. We are using the word segmentation from the OCR output, which is not affected as much by the image quality as the rest of the OCR process. Word segmentation is one of the first steps in the OCR process and produces a set of coordinates describing the location of the words on the page. Then the FAST algorithm [4] is used to find keypoints on each letter. A keypoint is a pixel that is unique and is able to describe the characteristics of the curve. An example of one can been seen in Figure II.1 where a keypoint has been identified on the curve of the letter b. Keypoints are generally found on the edges and at the corners of a letter. Then SIFT [3] is used to find the descriptors of each keypoint. This provides the context that the keypoint was in and identifies a characteristic of the letter. Then a hierarchal k-means is used to cluster the thousands of keypoints into several hundred clusters. This allows similar curves and characteristics to be labeled with the same identifier. Then each word is comprised of a series of identifiers called cluster ids. Each word is identified by the cluster ids sorted from left to right. Finally words are matched by comparing their string of cluster ids by using longest common subsequence (LCS). A probability is produced for each word compared to every other word and the highest one is deemed the match. Then the results from the matching are fed into our position matching algorithm that takes into account the context of each word to correctly match the words.

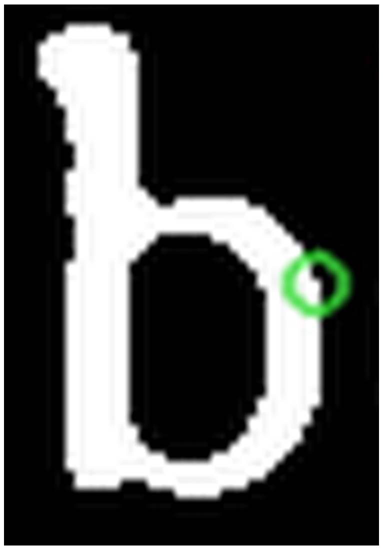


Fig. II.1. A keypoint that has been identified on the letter b that is marked by a green circle.

Finding the correct number of clusters

The way the algorithm works currently requires an optimal number of clusters to work accurately. The correct number is affected by a variety of factors. The type of language, quality of the image, font, and font size can all change the number of keypoints. Ideally, all the same type of keypoints would be in their own cluster. Originally, the correct number of clusters was not known, so trial and error was used starting at 4096 clusters. To aid finding the correct number of clusters, a tool was made that displays all of the points in a cluster on a page. Sample output from this tool can be seen in Figure II.2. This helps greatly because a page can have more than 40,000 keypoints on it and it becomes hard to distinguish the keypoints. The tool allows the grouping of the key points to be evaluated and if too many keypoints in a cluster were different then more clusters are needed to allow each cluster to have only the keypoints describing a unique characteristic. However, if the keypoint that describes the same unique characteristic is in different clusters then the number of clusters needs to be reduced. We found that with the poor printing practices of the books we are working with the scanned images have splotches and the type face is grainy. This affected the number of keypoints on the page resulting in less accurate results. To alleviate this the FAST algorithm [4] to find the keypoints was executed before a threshold, which is used to remove background noise, was applied. This prevented the threshold from taking to much of the detail from the image resulting in less keypoints. The SIFT algorithm [3] was then used to find the descriptors like before. After using the debugging tool to find where the keypoints were correctly distributed we found that 350 clusters was accurate for the novels being used.

Deciding if two words match

Our first step is to compare each word on the first page with all of the words from the second page. This ensures that every combination is taken into account. A top coverage score is calculated for each word from the first page.

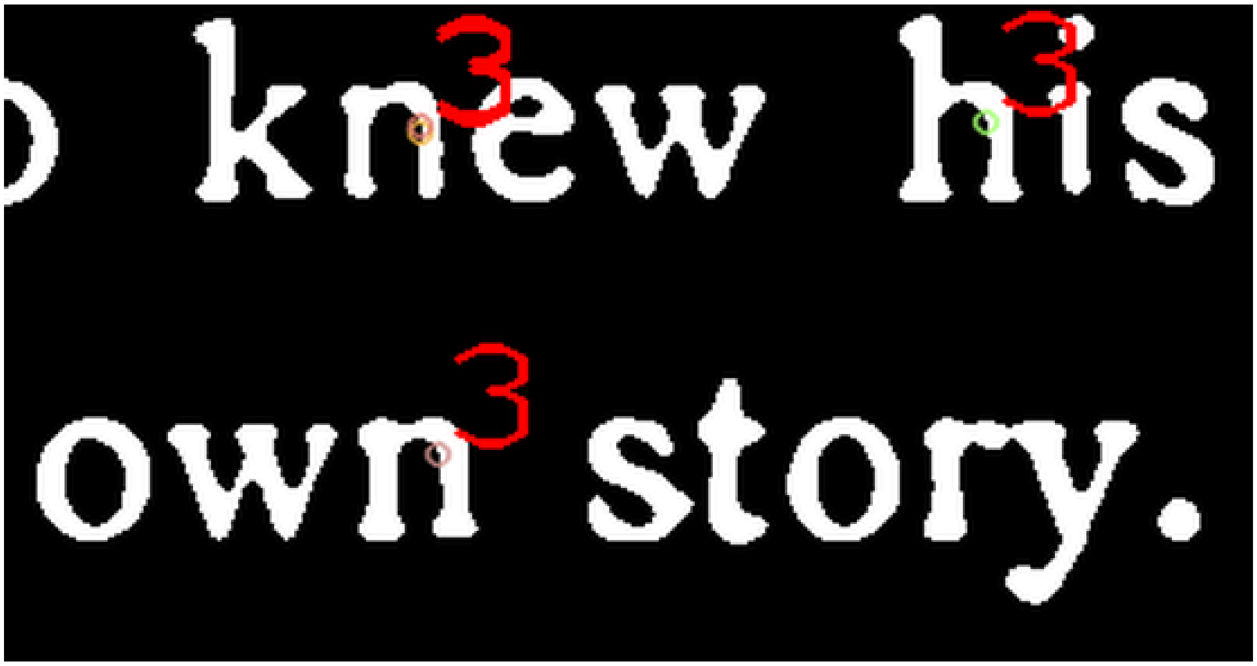


Fig. II.2. Output from the clustering tool that shows the keypoints in cluster 3.

$$TopCoverageScore = \frac{\frac{matches}{size1} + \frac{matches}{size2}}{2} * \frac{widthS}{widthL}$$

Matches is the size of the intersection of the two words. Size refers to the number of keypoints each word has. Width is how wide each word is in pixels.

This score is used to filter the number of words each word could match. It does this by giving word sets that do not have many cluster ids in common a low score. The low score is essentially the same as throwing the word out. The word sizes of the words is also taken into account. The equation always has the smaller word's pixel width divided by the larger word's pixel width. This ratio will give a low score to words with largely different widths and an almost unchanged score to similar width words. The width in pixel of the two words is also scaled to the dimensions of each page. This ensures that differing qualities of the scanned images or different font sizes is taken into account when comparing the widths. This helps prevent the algorithm from matching words together that are similar, but completely different. For example “the” and “thesaurus” are alike since “the” is present in both, but they are not the same and shouldn't match. With the length of the word taken into account these words will not match.

Once all of the scores have been computed a sorted list with the highest scores first is compiled. A list of the top twenty matches is then used to calculate the LCS score. By only comparing a word to the top twenty matches a vast amount of time is saved considering that there can be anywhere between 200 and 500 words on a page depending on the font size. This greatly speeds the matching up as the LCS function is a bottleneck. After the LCS score is calculated a final weighted score is made.

$$ProbabilityWeighted = LCSScore * LAMBDA + topcoveragescore * (1 - LAMBDA)$$

This final score is used to determine the match for each word. The score will range from 0 to 1 and the higher is better. This is the first step of the matching algorithm. The problems this step still has not addressed are dealing with multiple occurrences of the same word, small words, and similar words. These problems require a more in depth solution. Our first step is accurate for longer, more unique words.

Position matching

To increase the accuracy and thus the usefulness of our tool we decided to implement position based matching. Different approaches were tried including finding the offset between two pages and using a one to one matching from there. This approach could not handle added text between the two editions. Another way attempted involved taking the words around the word to be matched and comparing the top twenty matches of each of the surrounding words. This approach was abandoned due to the extensive time require to compute the thousands of potential matches. The time constraint of the data being available in minutes and the inaccuracy of the first approach led us to take a multi-step approach.

Small words are particularly difficult to match since they have very few keypoints and are often found inside other words. For example the letter "a" is a commonly used vowel that appears often in words. Finding the correct match for these words requires a different approach. Currently, words one or two letters in length are ignored by the algorithm and will not be highlighted.

We chose to use the context matching approach in the first step, but this time only the top five matches for each word will be considered instead of the top 20 matches. This greatly reduces the amount of time it takes to be computed. In this step we first take six words and find the smallest two words and set them to -1. Setting them to -1 means that we do not know the match currently and will find the match later once we have more information. Then two more words are added to the group that are bigger than the threshold set by the smallest words. Any words that are found that are smaller than the current threshold are set to -1. Then all of the possible sequences of the six words each with the top five matches is computed. The sequence that represents a linearly increasing model the best is selected as the correct pattern. Every pattern is checked by a function that sets values that are twenty indexes aways from the median of the sequence to -1. This allows outliers, which are most likely errors to be corrected later in the process. Once all the sequences have been considered the pattern with the most number of matches is selected. Then the process is started over with the next set of words until all of the words have been completed.

The next step addresses the words with “-1” as their matching word. In this step we find the sections that have “-1” and take the words before and after’s match that is not “-1” and set these as anchors. The anchors are checked to make sure they are not outliers. Then the “-1” matches are filled with the linearly increasing index. This sequence is then checked to see if it is a valid sequence and if it isn’t the result is set “-1”. This step relies on being able to accurately find anchors. From the testing we have done on the novels we have been pleased with the accuracy. The anchors are larger words that are more unique. The previous step has set smaller words to “-1”, so an anchor should not be a small word, which are more likely to be incorrectly matched.

CHAPTER III
COLLATION TOOL ADDED TO CRITSPACE

CritSpace

CritSpace is a creativity support environment, previously developed in our lab, that uses spatial information management strategies to help increase the users productivity. The usefulness of CritSpace comes from being able to move a panel that can contain an image, a text editor, or a facsimile viewer that lets one select the next image anywhere on the web page. An example layout of CritSpace can be seen in Figure III.1. This custom layout combined with multiple tools in one platform cater to the process of collation. We decided to incorporate our collation tool into CritSpace to add the flexibility that CritSpace provides to our collation tool.

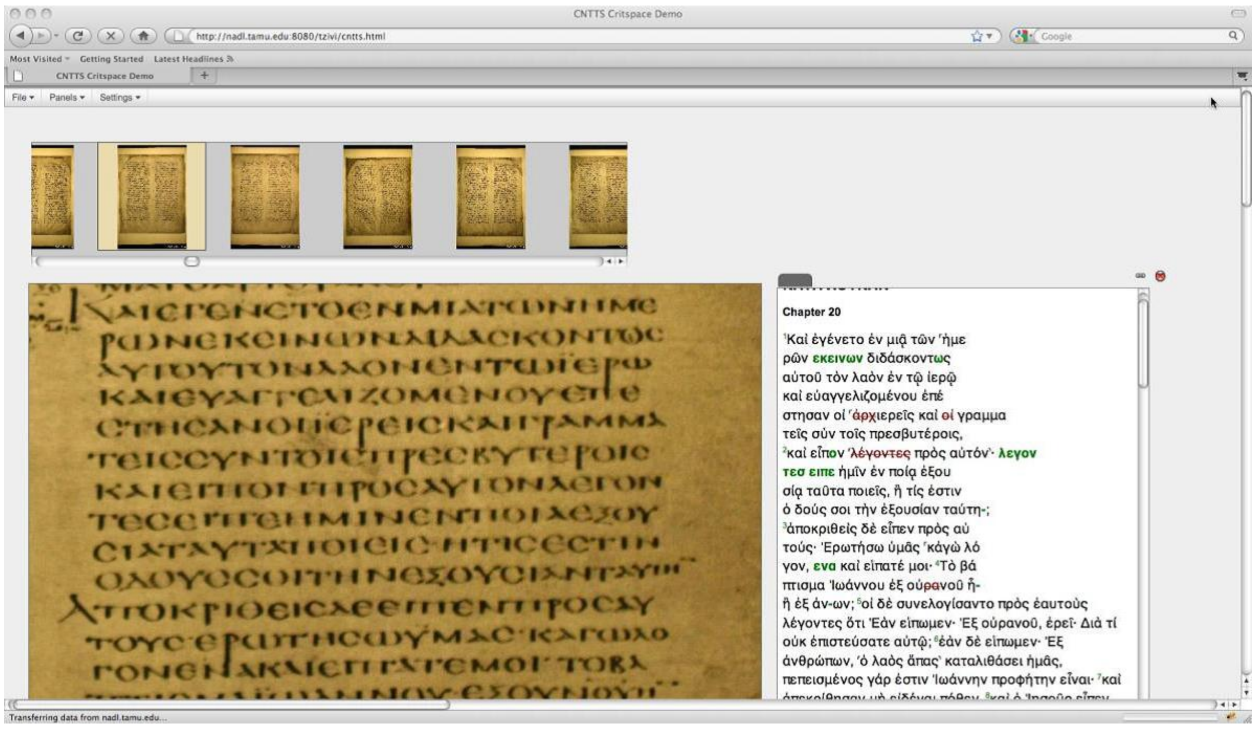
Collation tool

Control panel

Our Collation tool is comprised of a control panel and image panels that are created by the control panel. The control panel allows the user to select what page for each book they wish to compare. This is done by rotating the dial as seen in Figure III.2 to the desired page number. After the page numbers are set the user clicks the first button called Collation. This will create two image panels in the workspace. Since the tool is integrated into CritSpace the user is able to enlarge/shrink the images and order the images to their preference. Also multiple collation panels can be open simultaneously allowing greater flexibility.

Collation modes

The collation tool currently supports three different modes to aid in the collation process. The first mode, seen in Figure III.3, is the differences that displays the differences between the two pages from each edition in green. This mode is the main collation mode and allows


CNTTS Critspace Demo
http://nadl.tamu.edu:8080/tzivi/cntts.html
File
Panels
Settings
Transferring data from nadl.tamu.edu...

Chapter 20

1Καὶ ἐγένετο ἐν μιᾷ τῶν ʽἡμε
ρῶν εκεινων διδάσκοντως
αὐτοῦ τὸν λαὸν ἐν τῷ ἱερῷ
καὶ εὐαγγελιζομένου ἐπέ
στησαν οἱ ʽἀρχιερεῖς καὶ οἱ γραμμα
τεῖς σὺν τοῖς πρεσβυτέροις,
2καὶ εἶπον ʽ~~λέγοντες~~ πρὸς αὐτόν˺· λεγον
τεσ ειπε ἡμῖν ἐν ποίᾳ ἐξου
σίᾳ ταῦτα ποιεῖς, ἢ τίς ἐστιν
ὁ δούς σοι τὴν ἐξουσίαν ταύτη-;
3ἀποκριθεὶς δὲ εἶπεν πρὸς αὐ
τούς· Ἐρωτήσω ὑμᾶς ʽκἀγὼ λό
γον, ενα καὶ εἴπατέ μοι· 4Τὸ βά
πτισμα Ἰωάννου ἐξ οὐρανοῦ ἢ-
ἢ ἐξ ἀν-ων; 5οἱ δὲ συνελογίσαντο πρὸς ἑαυτοὺς
λέγοντες ὅτι Ἐὰν εἴπωμεν· Ἐξ οὐρανοῦ, ἐρεῖ· Διὰ τί
οὐκ ἐπιστεύσατε αὐτῷ; 6ἐὰν δὲ εἴπωμεν· Ἐξ
ἀνθρώπων, ʽὁ λαὸς ἅπας˺ καταλιθάσει ἡμᾶς,
πεπεισμένος γάρ ἐστιν Ἰωάννην προφήτην εἶναι· 7καὶ

Fig. III.1. An example of the Critspace environment with an image viewer panel, text editor panel and a facsimile viewer panel to select images.

the user to quickly identify differences between two pages. In Figure III.3 the orange, red, and blue boxes point out the differences found between the two editions of the novel. The words in the first book are hyphenated where as the second book omits hyphens. Also the maroon box shows that the first page is offset from the second page.

The next mode is the tracking mode which, when chosen, allows the user to move the mouse over a word in the first image and its matching word will be highlighted on the second book. This can been see in Figure III.4 where the user moved his mouse over the words on the left page and the matches were highlighted on the second page. The words on the first book are highlighted blue first and then changed to black after a new word is chosen to empathize the currently selected word. The words on the second page are highlighted red instead of blue.

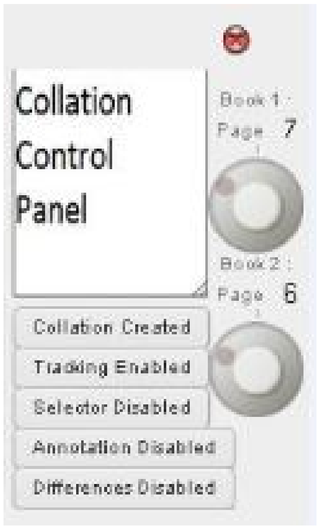
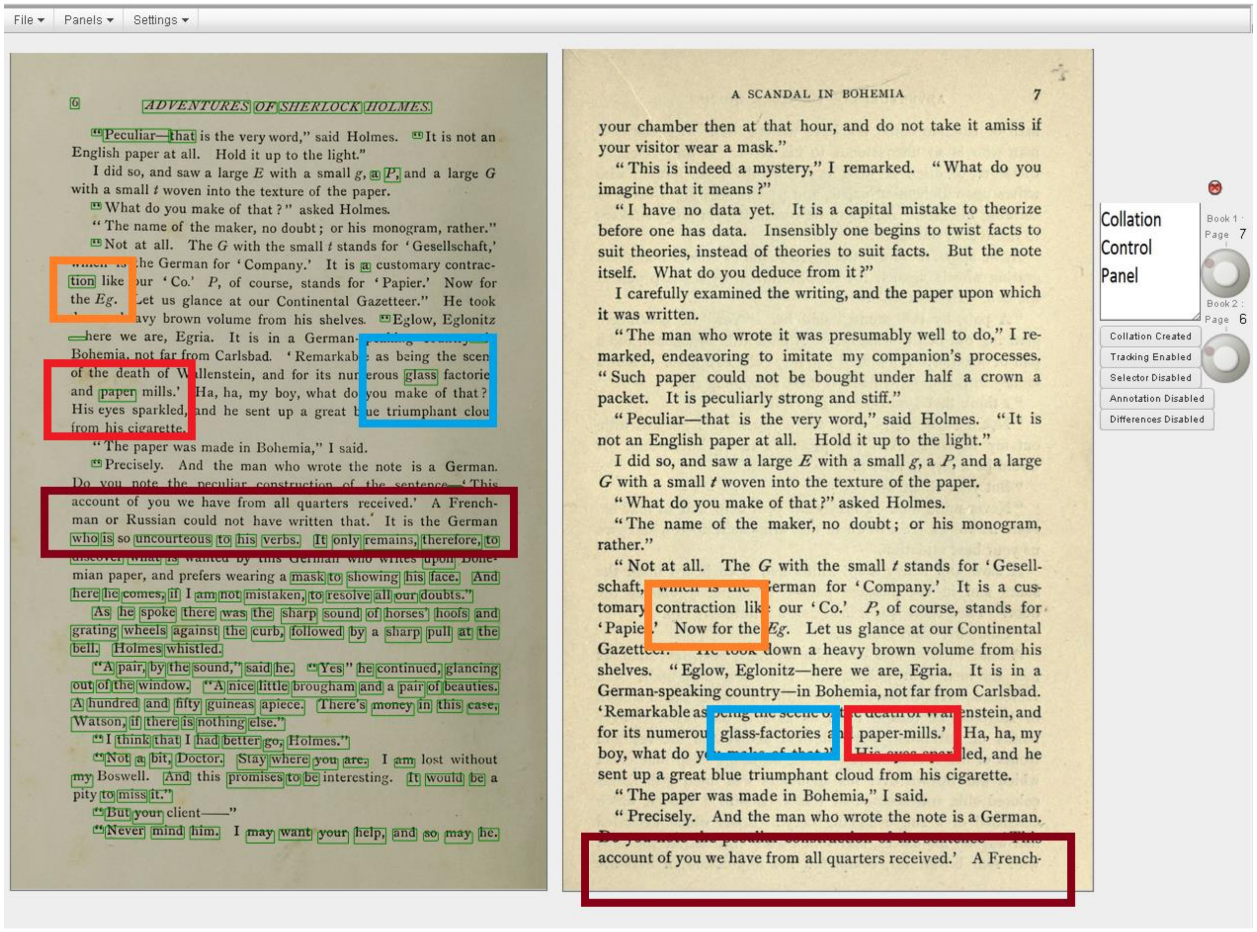


Fig. III.2. Control Panel for the collation tool that lets the user select the page numbers and mode.



6 *ADVENTURES OF SHERLOCK HOLMES.*

"Peculiar—that is the very word," said Holmes. "It is not an English paper at all. Hold it up to the light."

I did so, and saw a large *E* with a small *g*, a *P*, and a large *G* with a small *t* woven into the texture of the paper.

"What do you make of that?" asked Holmes.

"The name of the maker, no doubt; or his monogram, rather."

"Not at all. The *G* with the small *t* stands for 'Gesellschaft,' [illegible] the German for 'Company.' It is a customary contraction like [illegible]our 'Co.' *P*, of course, stands for 'Papier.' Now for the *Eg*. [illegible]et us glance at our Continental Gazetteer." He took [illegible] avy brown volume from his shelves. "Eglow, Eglonitz—here we are, Egria. It is in a German-[illegible] Bohemia, not far from Carlsbad. 'Remarkab[illegible] as being the scen[illegible] of the death of W[illegible]llenstein, and for its nu[illegible]erous glass factorie[illegible] and paper mills.' Ha, ha, my boy, what do you make of that? His eyes sparkled, and he sent up a great b[illegible]ue triumphant clou[illegible] from his cigarette.

"The paper was made in Bohemia," I said.

"Precisely. And the man who wrote the note is a German. Do you note the peculiar construction of the sentence—'This account of you we have from all quarters received.' A Frenchman or Russian could not have written that. It is the German who is so uncourteous to his verbs. It only remains, therefore, to [illegible] upon Bohemian paper, and prefers wearing a mask to showing his face. And here he comes, if I am not mistaken, to resolve all our doubts."

As he spoke there was the sharp sound of horses' hoofs and grating wheels against the curb, followed by a sharp pull at the bell. Holmes whistled.

"A pair, by the sound," said he. "Yes" he continued, glancing out of the window. "A nice little brougham and a pair of beauties. A hundred and fifty guineas apiece. There's money in this case, Watson, if there is nothing else."

"I think that I had better go, Holmes."

"Not a bit, Doctor. Stay where you are. I am lost without my Boswell. And this promises to be interesting. It would be a pity to miss it."

"But your client——"

"Never mind him. I may want your help, and so may he.

A SCANDAL IN BOHEMIA 7

your chamber then at that hour, and do not take it amiss if your visitor wear a mask."

"This is indeed a mystery," I remarked. "What do you imagine that it means?"

"I have no data yet. It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts. But the note itself. What do you deduce from it?"

I carefully examined the writing, and the paper upon which it was written.

"The man who wrote it was presumably well to do," I remarked, endeavoring to imitate my companion's processes. "Such paper could not be bought under half a crown a packet. It is peculiarly strong and stiff."

"Peculiar—that is the very word," said Holmes. "It is not an English paper at all. Hold it up to the light."

I did so, and saw a large *E* with a small *g*, a *P*, and a large *G* with a small *t* woven into the texture of the paper.

"What do you make of that?" asked Holmes.

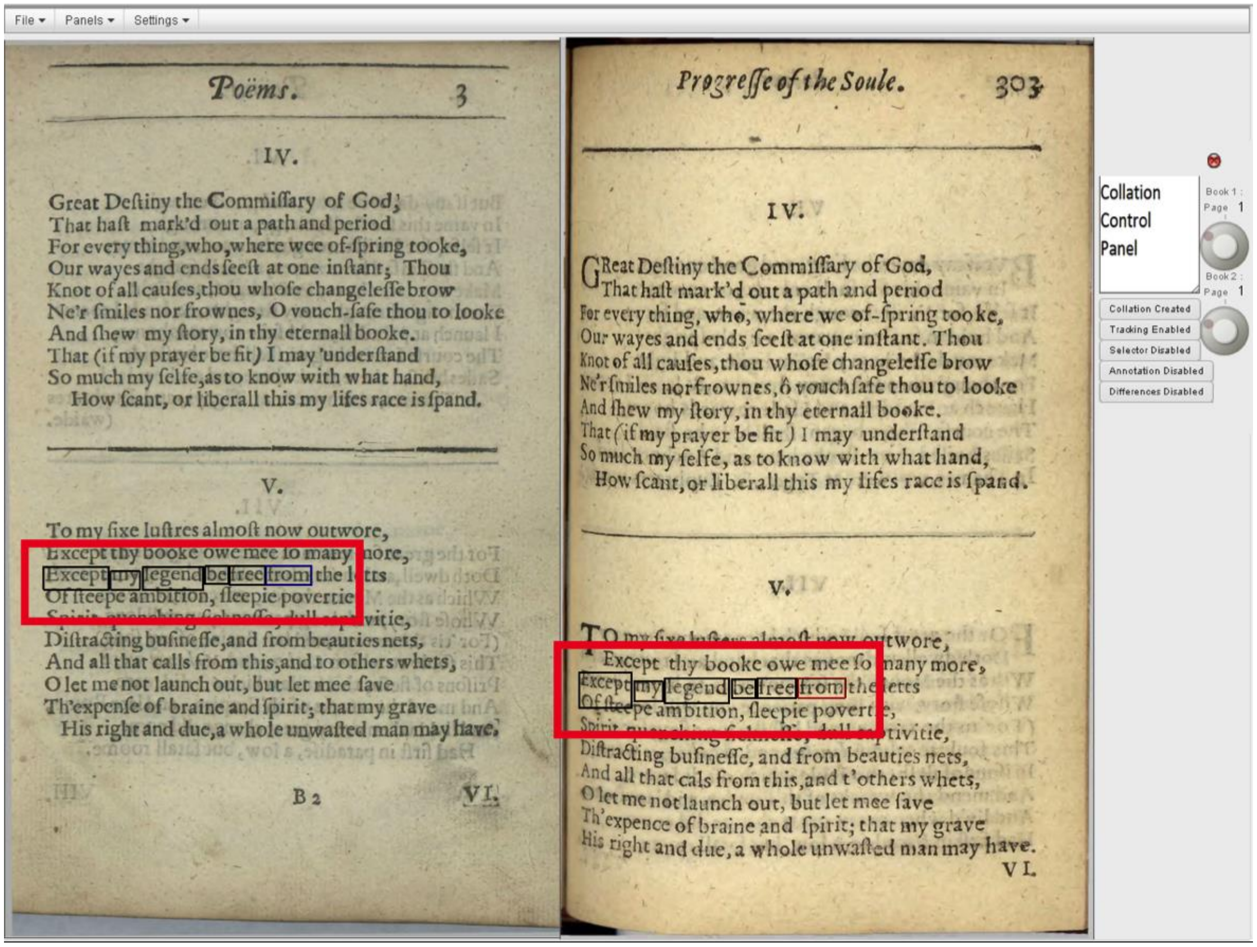
"The name of the maker, no doubt; or his monogram, rather."

"Not at all. The *G* with the small *t* stands for 'Gesellschaft, [illegible]erman for 'Company.' It is a customary contraction lik[illegible] our 'Co.' *P*, of course, stands for 'Papie[illegible].' Now for the *Eg*. Let us glance at our Continental Gazett[illegible] down a heavy brown volume from his shelves. "Eglow, Eglonitz—here we are, Egria. It is in a German-speaking country—in Bohemia, not far from Carlsbad. 'Remarkable as [illegible]enstein, and for its numerou[illegible] glass-factories [illegible] paper-mills.' Ha, ha, my boy, what do y[illegible]led, and he sent up a great blue triumphant cloud from his cigarette.

"The paper was made in Bohemia," I said.

"Precisely. And the man who wrote the note is a German. [illegible] account of you we have from all quarters received.' A French-

Fig. III.3. The differences between two pages are shown.



Poëms. 3

IV.

Great Deſtiny the Commiſſary of God,
That haſt mark'd out a path and period
For every thing,who,where wee of-ſpring tooke,
Our wayes and ends ſeeſt at one inſtant; Thou
Knot of all cauſes,thou whoſe changeleſſe brow
Ne'r ſmiles nor frownes, O vouch-ſafe thou to looke
And ſhew my ſtory, in thy eternall booke.
That (if my prayer be fit) I may 'underſtand
So much my ſelfe,as to know with what hand,
How ſcant, or liberall this my lifes race is ſpand.

V.

To my ſixe luſtres almoſt now outwore,
Except thy booke owe mee ſo many more,
Except my legend be free from the letts
Of ſteepe ambition, ſleepie povertie,
Spirit-quenching ſickneſſe, dull captivitie,
Diſtracting buſineſſe,and from beauties nets,
And all that calls from this,and to others whets,
O let me not launch out, but let mee ſave
Th'expenſe of braine and ſpirit; that my grave
His right and due,a whole unwaſted man may have.

B 2 VI.

Progreſſe of the Soule. 303

IV.

Great Deſtiny the Commiſſary of God,
That haſt mark'd out a path and period
For every thing, who, where we of-ſpring tooke,
Our wayes and ends ſeeſt at one inſtant. Thou
Knot of all cauſes,thou whoſe changeleſſe brow
Ne'r ſmiles nor frownes, ô vouchſafe thou to looke
And ſhew my ſtory, in thy eternall booke.
That (if my prayer be fit) I may underſtand
So much my ſelfe, as to know with what hand,
How ſcant, or liberall this my lifes race is ſpand.

V.

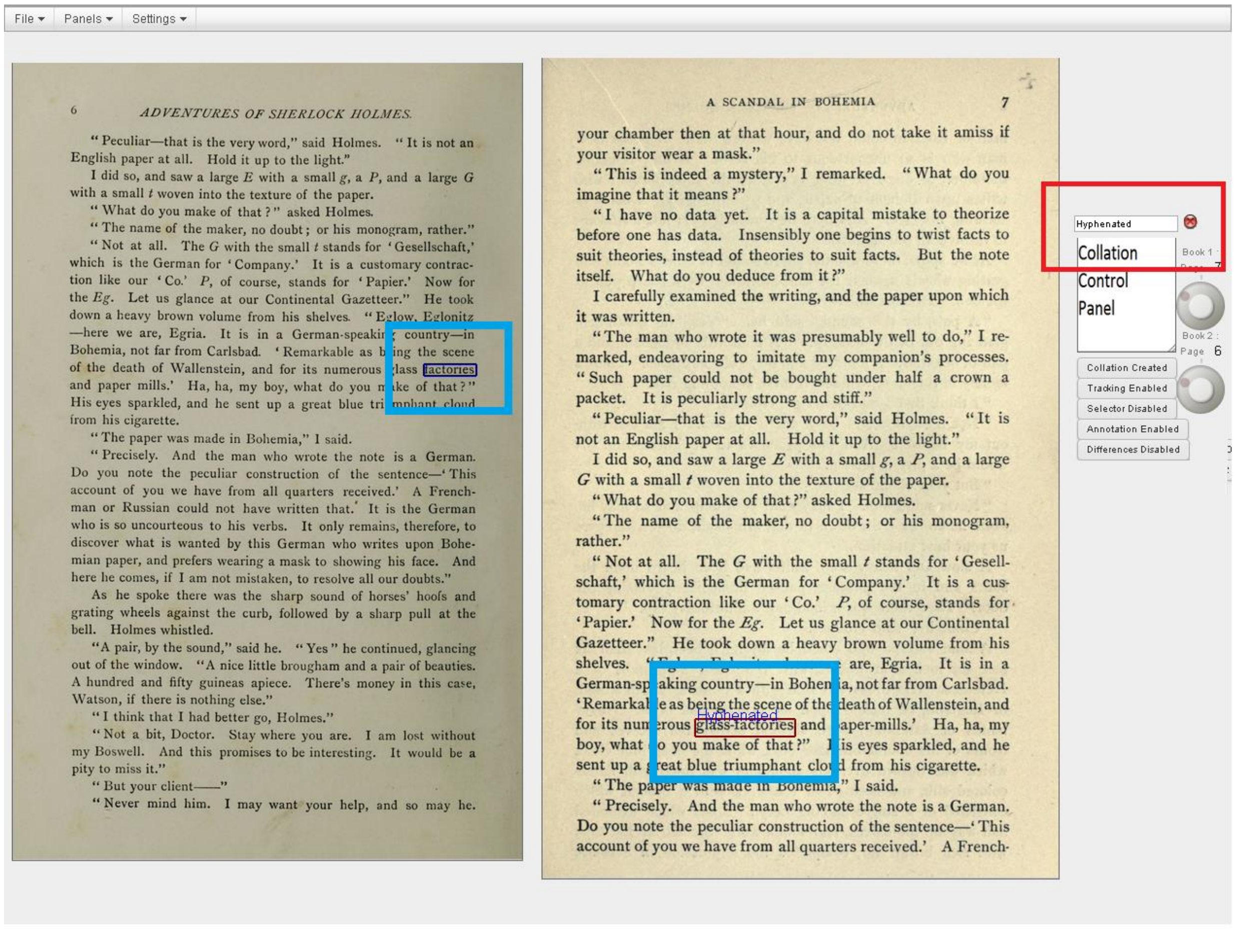
To my ſixe luſtres almoſt now outwore,
Except thy booke owe mee ſo many more,
Except my legend be free from the letts
Of ſteepe ambition, ſleepie povertie,
Spirit-quenching ſickneſſe, dull captivitie,
Diſtracting buſineſſe, and from beauties nets,
And all that cals from this,and t'others whets,
O let me not launch out, but let mee ſave
Th'expence of braine and ſpirit; that my grave
His right and due, a whole unwaſted man may have.

VI.

Fig. III.4. Tracking feature displayed on the Donne Poems.

This feature is helpful to easily find the offset between the two pages as well as verifying that the tool is matching words correctly.

The last mode is the annotation mode and can be used in conjunction with the other modes. When the annotation mode button is clicked a text box appears at the top of the control panel as seen in Figure III.5. The user simply types the note into the text box and then clicks on a word on the second page to add the note to the selected word. To use this mode in conjunction with the tracking and differences mode the user needs to already be in either the tracking or differences mode.



6 *ADVENTURES OF SHERLOCK HOLMES.*

"Peculiar—that is the very word," said Holmes. "It is not an English paper at all. Hold it up to the light."

I did so, and saw a large *E* with a small *g*, a *P*, and a large *G* with a small *t* woven into the texture of the paper.

"What do you make of that?" asked Holmes.

"The name of the maker, no doubt; or his monogram, rather."

"Not at all. The *G* with the small *t* stands for 'Gesellschaft,' which is the German for 'Company.' It is a customary contraction like our 'Co.' *P*, of course, stands for 'Papier.' Now for the *Eg*. Let us glance at our Continental Gazetteer." He took down a heavy brown volume from his shelves. "Eglow, Eglonitz—here we are, Egria. It is in a German-speaking country—in Bohemia, not far from Carlsbad. 'Remarkable as being the scene of the death of Wallenstein, and for its numerous glass factories and paper mills.' Ha, ha, my boy, what do you make of that?" His eyes sparkled, and he sent up a great blue triumphant cloud from his cigarette.

"The paper was made in Bohemia," I said.

"Precisely. And the man who wrote the note is a German. Do you note the peculiar construction of the sentence—'This account of you we have from all quarters received.' A Frenchman or Russian could not have written that. It is the German who is so uncourteous to his verbs. It only remains, therefore, to discover what is wanted by this German who writes upon Bohemian paper, and prefers wearing a mask to showing his face. And here he comes, if I am not mistaken, to resolve all our doubts."

As he spoke there was the sharp sound of horses' hoofs and grating wheels against the curb, followed by a sharp pull at the bell. Holmes whistled.

"A pair, by the sound," said he. "Yes" he continued, glancing out of the window. "A nice little brougham and a pair of beauties. A hundred and fifty guineas apiece. There's money in this case, Watson, if there is nothing else."

"I think that I had better go, Holmes."

"Not a bit, Doctor. Stay where you are. I am lost without my Boswell. And this promises to be interesting. It would be a pity to miss it."

"But your client——"

"Never mind him. I may want your help, and so may he.

A SCANDAL IN BOHEMIA 7

your chamber then at that hour, and do not take it amiss if your visitor wear a mask."

"This is indeed a mystery," I remarked. "What do you imagine that it means?"

"I have no data yet. It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts. But the note itself. What do you deduce from it?"

I carefully examined the writing, and the paper upon which it was written.

"The man who wrote it was presumably well to do," I remarked, endeavoring to imitate my companion's processes. "Such paper could not be bought under half a crown a packet. It is peculiarly strong and stiff."

"Peculiar—that is the very word," said Holmes. "It is not an English paper at all. Hold it up to the light."

I did so, and saw a large *E* with a small *g*, a *P*, and a large *G* with a small *t* woven into the texture of the paper.

"What do you make of that?" asked Holmes.

"The name of the maker, no doubt; or his monogram, rather."

"Not at all. The *G* with the small *t* stands for 'Gesellschaft,' which is the German for 'Company.' It is a customary contraction like our 'Co.' *P*, of course, stands for 'Papier.' Now for the *Eg*. Let us glance at our Continental Gazetteer." He took down a heavy brown volume from his shelves. "Eglow, Eglonitz—here we are, Egria. It is in a German-speaking country—in Bohemia, not far from Carlsbad. 'Remarkable as being the scene of the death of Wallenstein, and for its numerous glass-factories and paper-mills.' Ha, ha, my boy, what do you make of that?" His eyes sparkled, and he sent up a great blue triumphant cloud from his cigarette.

"The paper was made in Bohemia," I said.

"Precisely. And the man who wrote the note is a German. Do you note the peculiar construction of the sentence—'This account of you we have from all quarters received.' A French-

Fig. III.5. Annotation feature

How the matching algorithm is executed

Users require a quick and responsive interface that serves data in a timely manner. The definition of timely can vary vastly, but for this project the goal is to have a match show up in a few seconds. Currently, from start to finish finding the differences between two pages can take several hours. Fortunately, the difference computation does not need to be computed in real time like other applications may call for. To meet the time requirements of the project the segmentation and clustering process with k-means is computed for each pair of editions when they are added to the program. The results are stored in text files that are read in by the matching portion of the program. When the user selects the pair of pages to be computed and clicks the create button a message is sent to a servlet written in Java. This servlet takes the page numbers and forwards them to the matching program, that is then executed. After it is finished the results text file and other supporting text files are loaded into a JSON structure by the servlet and are sent back to CritSpace. Then, the message is parsed and used to display the results on the images.

Testing material

Currently, the collation tool has been tested on *The Adventures of Sherlock Holmes* and the *Donne Poems* written in 1633 and 1635. *The Sherlock Holmes* set of editions was used during the development of the tool because of the high resolution facsimiles available and the notable differences between the editions. One edition has a larger font size ,which results in an offset between the pages, and there are different words and capitalizations throughout the novels. The *Donne Poems* show the tools ability to work on less than desired printing conditions. The text is bled through from the other side and splotches and random marks are rampant.

CHAPTER IV
CONCLUSION

The vHinman tool addresses many of the problems that previous solutions have yet to solve by using the elegance and strength of technology. It is able to automate the process of finding the differences between editions and is wrapped into an interface that supports the user throughout the collation process. Also, the vHinman can be used on a wider set of novels than other techniques with the flexibility from the image processing approach. Humanists now have more time to decipher the differences, and come to conclusions that will help understand why the changes were made. Also more collations will be able to be completed with the increased speed. With integration into CritSpace the collation tool is in a platform that supports collaboration which will be helpful in larger collation efforts. Ultimately, this tool will help humanists complete their tasks easier and more efficiently.

CHAPTER V
FUTURE WORK

Supporting multiple editions

With the current configuration of the collation tool, comparing three or more editions to each other would be difficult. Since it is set up for two editions, the three novels would have to be split into pairs to collate which results in added work. Luckily, with the customizable panel layout of CritSpace adding another image panel to support more editions to be collated at once would not be difficult. With multiple copy collation the user would need to select which text to be the baseline that all of the others would be compared to. The combinations between the baseline copy and each additional copy could be precomputed to have quick results for the user. Adding this to the collation tool would help greatly in collating Shakespeare's *First Folio* which has have multiple different editions. Some of the editions had corrections made during the print runs [5]. This tool would be helpful in quickly finding the differences between the editions.

New novels

The collation tool could be expanded to new novels that are in different languages than English. The image processing approach [6] has been shown to work with the Telugu language, which is not based on a Latin alphabet, but a Brahmi alphabet. This is a significant advantage over OCR because our collation tool can be much more language independent than OCR which needs to support each language. This can open up a vast number of novels to be collated that previously would need to be done by hand. In order to work with a new novel the number of clusters may need to be adjusted depending on the uniqueness of the alphabet, but this is an easy calibration task.

REFERENCES

[1] Michael Neal Audenaert. *CritSpace: An Interactive Visual Interface to Digital Collections of Cultural Heritage Material*. PhD thesis, Texas A&M University, 2011.

[2] Gordon Lindstrand. Mechanized textual collation and recent designs. *Studies in Bibliography*, pages 204–214, 1971.

[3] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[4] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In *Computer Vision–ECCV 2006*, pages 430–443. Springer, 2006.

[5] Alice Walker. *Textual problems of the First Folio*. CUP Archive, 1953.

[6] Ismet Zeki Yalniz and R Manmatha. An efficient framework for searching text in noisy document images. 2011.