AIDING MODERN TEXTUAL SCHOLARSHIP USING A VIRTUAL HINMAN COLLATOR

A Thesis

by

GAURAV KEJRIWAL

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Richard Furuta
Committee Members,	Frank Shipman
	Laura Mandell
Head of Department,	Nancy Amato

May 2014

Major Subject: Computer Science

ABSTRACT

Collation is an important step in textual criticism and is most often an arduous task for most scholars involved in scholarly edition. Finding variations is important for researchers in bibliography and book history as well. In the late 1940s Charlton Hinman invented a machine that became popular as the Hinman collator. Using optical means, the Hinman Collator allowed manual comparison of separate copies of a text in order to detect any differences that had been introduced. Although these mechanical collation systems are helpful, they still require a lot of manual labor and some scholars find them hard to use. Another approach used sometimes is to perform collation on OCR output of text. However the state-of-the-art OCR mechanisms for 15th/16th century books are not efficient to date (70-80% accurate). Also scholars doing textual criticism generally prefer to work on original copies or facsimiles rather than OCR versions of them because the accuracy and some of the nuanced details of the original copy are important to them

Thus there is a need of a tool that can reduce the effort required in the collation process while maintaining (and sometimes improving) the usefulness of the tool and allowing scholars to use original documents (high quality facsimiles). This research focuses on this aspect of scholarly work and explores various approaches for performing digital collation in a seamlessly easy manner. A prototype of the virtual Hinman (vHinman) collator was created and user evaluation was conducted amongst scholars experienced with collation work. Image-matching algorithms along with context information are used to match words and the tool was integrated into the creativity support environment CritSpace.

The tool was tested on books from early modern and late modern period for which

multiple copies with slight variations were available. The tool showed a high accuracy rate for the books tested. Most of the scholars found the tool very promising. This kind of tool can save a massive amount of time for scholars and set up a paradigm of digital collation encouraging even more scholars in finding new uses of collation in their work.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Richard Furuta for his excellent guidance, patience and support during the course of this research I would also like to thank my committee members Dr. Frank Shipman and Dr. Laura Mandell for their support and guidance in carrying out this research.

I would also like to thank undergraduate research scholar Ryan Olivieri for his amazing work of web-development for integrating the tool into CritSpace and also suggesting and implementing ideas related to improving the algorithm.

Thanks also to Neal Audenaert for his support regarding explaining the inbuilt architecture of CritSpace. I would also like to thank Luis Meneses at the CSDL for his guidance in various aspects.

Also, I would like to thank Ismet Zeki Yalnz for providing detailed explanations of the work described in their published paper on “An efficient framework for searching text in noisy document images”.

NOMENCLATURE

OCR	Optical Character Recognition
vHinman	Virtual Hinman Collator
TEI	Text Encoding Initiative
VISterms	Visual Terms

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

TABLE Page

1. INTRODUCTION

The Oxford English Dictionary defines the verb collate as comparing critically (a copy of a text) with other copies or with the original, in order to correct and emend it [Kuhn, 2010]. Unsworth includes collation as one of the scholarly primitives that have been basic to scholarship across eras and media [Unsworth, 2000]. Textual variation has been a pervasive problem affecting literary text since the invention of writing. It can arise in two forms - either due to repeated copying of a manuscript, such as the variants in the First Folio of Shakespeare, or those advertently inserted by the author/copyist such as the changes made in Mary Shelleys Frankenstein. In the first case collation aids the scholar in generating a critical edition. In the latter case, collation can help the scholar understand the authors purpose. Finding variations is important for researchers in bibliography and book history as well. It is commonly known that in the 15th/16th century print press, books were proofread while the prints were done so no one copy could be considered as the authoritative text. Hence collating multiple copies of these works helps in figuring out the authoritative text.

Collation is usually an arduous task for most scholars involved in scholarly edition, although technology has enabled scholars to access original facsimiles of rare documents without having to travel to the libraries and museums. Most of the focus in digital humanities till now has been on making documents available digitally and making standards like TEI for easing preparation and interchange of electronic texts. Much less focus has been laid on actually supporting the process of scholarly research. The area of collation too awaits a lot more from technology. Most of the humanists still perform paper-based collation, which is prone to errors and consumes a lot of manual effort.

In the early days, collation was done by reading one word at a time (aloud if two people performed collation) or by keeping fingers on the particular word on both the texts. This is a process where mistakes are inevitable as the collator has to read not just one but two texts correctly at once. Mistakes can also arise while recording the differences correctly [Robinson, 1994]. In the late 1940s Charlton Hinman was assigned the task to create a scholarly edition for the First Folio of Shakespeare by collating the various available copies of it. To reduce the manual effort required in this process he invented a machine, which became popular as the Hinman collator [Smith, 2002]. Using optical means, the Hinman Collator allowed manual comparison of separate copies of a text in order to detect any differences that had been introduced. Mechanical collators in some variant form of the Hinman collator are still used today by scholars. Some of them are the Mcleod collator, the Lindstrand collator and the Hailey's Comet [Smith, 2002]. The Hinman collator was bought by around fifty-seven institutions and is still used in some institutions today. David Vander Meulen used the Hinman to collate copies of Pope's Dunciad and examined running titles to resolve the old question of which of the two 1728 issues came first [Smith, 2002]. R. Carter Hailey, examined around sixty copies of the three 1550 editions of Piers Plowman on the Haileys Comet for his dissertation related to the analysis of the work done by Robert Crowley [Bibliographical-Mirrors, 1999].

The basic principle behind all these tools is that they rely on optical phenomenon to make two images superimpose which makes the differences evident. The Hinman uses lights and shutters to present alternate images with a blinking effect, which highlights the differences [Smith, 2002]. In the Lindstrand, the researcher views two texts set up in separate cradles and positioned beneath a set of binocular optics. The optics, a set of mirrors and a prism puts the texts in a kind of virtual superimposition. When this effect is achieved, small differences between the texts seem to stand above

the similarities in 3D [Smith, 2002].

Although these mechanical collation systems are helpful, they still require a lot of manual labor and some scholars find them physically/mentally exhausting [Raabe, 2008]. They are mostly expensive and not portable (with the exception of McLeods collator). Also these machines can be damaging to the books. Moreover these tools are inefficient if there are differences in the font sizes, typeface, and alignment of the pages being compared.

Another approach that is sometimes used is to perform collation on the OCR output or transcription of text. Popular systems incorporating this approach include Collate 2.0 by Peter Robinson [Raabe, 2008], Juxta by NINES [NINES, 2011] and Versioning Machine [Schreibman, 2000]. However the state-of-the-art OCR mechanisms for 15th/16th century books are not efficient to date (70-80% accurate). Transcription is also not practical if the scholar has to collate a huge number of copies (say 50) and it is bound to produce human errors.

Also these tools dont allow scholars to use facsimiles of original documents that are important to them because of some of the nuanced details of the original copy [Audenaert and Furuta, 2010]. Researchers usually rely on digital facsimiles for most of their time-consuming research work while only going to the libraries/museum for the final proofing work which saves a lot of travel time (and money). In certain cases, the digital objects may fully satisfy the researchers needs [Audenaert, 2011].

There is another approach being researched upon where optical collation can be achieved using image registration techniques. The HUMI project at Keio University Japan tried to collate copies of Gutenberg Bible using this approach. The pages were hand-flattened using bamboo rods to reduce the warping effect, which isnt safe as we are dealing with precious ancient documents. The project aimed to collate copies of the Gutenberg Bible only, hence it is not practical [CDH, 2009]. The Virtual Light

Box project at MITH used a similar image-registration approach but relied on the user to align the images [CDH, 2009]. Another notable project was the Sapheos project at the Center of Digital Humanities, University Of Southern Carolina, which later evolved into the currently ongoing Paragon project [CDH, 2012]. They are trying to unwarp the images and automatically register them using SIFT key points. This approach is good for collating books where the variants are very minute and the text can be theoretically registered. It can be put to use in many cases where the mechanical collators are useful. However, it wont be effective in copies of the same book with changes made by the author himself, for instance, the copies of Mary Shelleys Frankenstein.

Most commonly, todays digital collators allow comparison of two documents. However the scholar generally consults many more than two sources in carrying out a collation. Consequently, a further goal of the work is to allow collation of multiple copies at once. Most of these collation tools are standalone tools which dont support collaborative work among multiple scholars and the scholars usually need to use multiple other tools (like text editors) simultaneously to perform their research. Thus there is a need of a tool that can reduce the effort required in the collation process while maintaining (and sometimes improving) the usefulness of the tool and allowing scholars to use original documents (high quality facsimiles).

This thesis focuses on this aspect of humanities research and in figuring out ways to best support the collation process digitally while blending it into the other tasks of the scholars work. The collation process is a combination of two steps, the manual part of comparing text word by word (including punctuations etc.) and the scholarly part of inferring what those differences mean (either in scholarly edition or bibliographic history).

This research focuses on making that first step as automated as possible so that

the scholar can focus solely on inferring what those differences mean and making implications out of it. Its worth noting that we want the tool to be an aid to the scholar, while still giving the final power of deciding its implications to the scholar thus only being an unobtrusive supporting tool in scholars work.

The aim of this research is to create a digital equivalent of the popular Hinman collator, invented in the late 1940s [Smith, 2002], which can reduce the manual effort that is required in the current collation process. The tool will also enable scholars to perform collation on facsimiles of original documents. We analyze how scholars perform their collation work and what kinds of differences are important to them. Section 2 describes our various approaches to this problem and also describes the interface whereby the tool was integrated into CritSpace [Audenaert et al., 2010]

Section 3 describes the results of a user-evaluation conducted at the Department Of English summarizing their ways of performing collation and their views on the tool.

Section 4 presents a conclusion of the work and presents ideas for future work on the tool.

2. METHODOLOGY

The work focused on creation of a vHinman tool, incorporated into CritSpace. In the process of this research, we developed and evaluated various approaches towards comparing page-images:

• Made two page images superimpose one over another and varied the z-index of the top image to blink the images one over other making the differences visible. This approach is a mimicking of the optical method employed in the mechanical collators and requires the images to be registered first.

• Made the opacity of the top page swing from high to low using a slider that made the differences more prominent.Please see figure 2.1

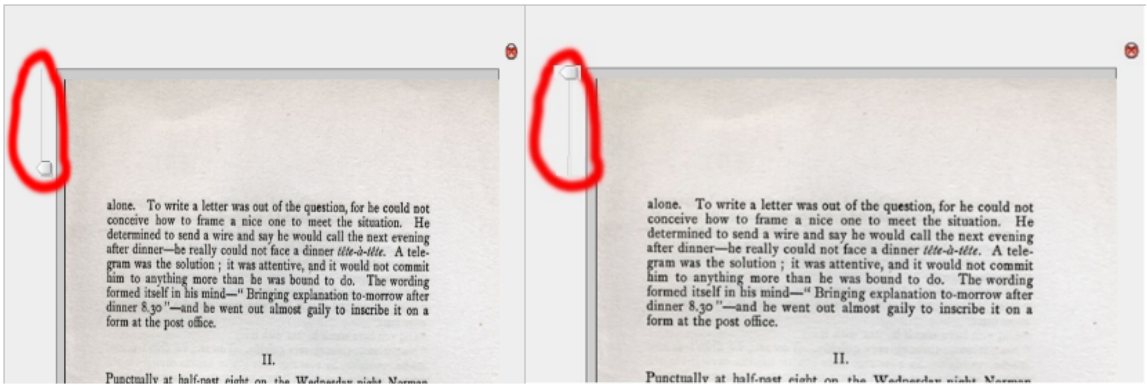


Figure 2.1: Screenshot of the opacity slider in two states.

• Used imagemagick [ImageMagick, 2012] tools inbuilt comparison methods to compare two images and highlight the difference. The comparison method works by subtracting the pixel intensity values of one image from another, which results in the differences being highlighted. Imagemagick does not have any scale and rotation

invariant comparison method. Hence, the images need to be manually registered (using imagemagicks other functions) to the same scale and rotation for the comparison to work effectively [Figure 2.2].

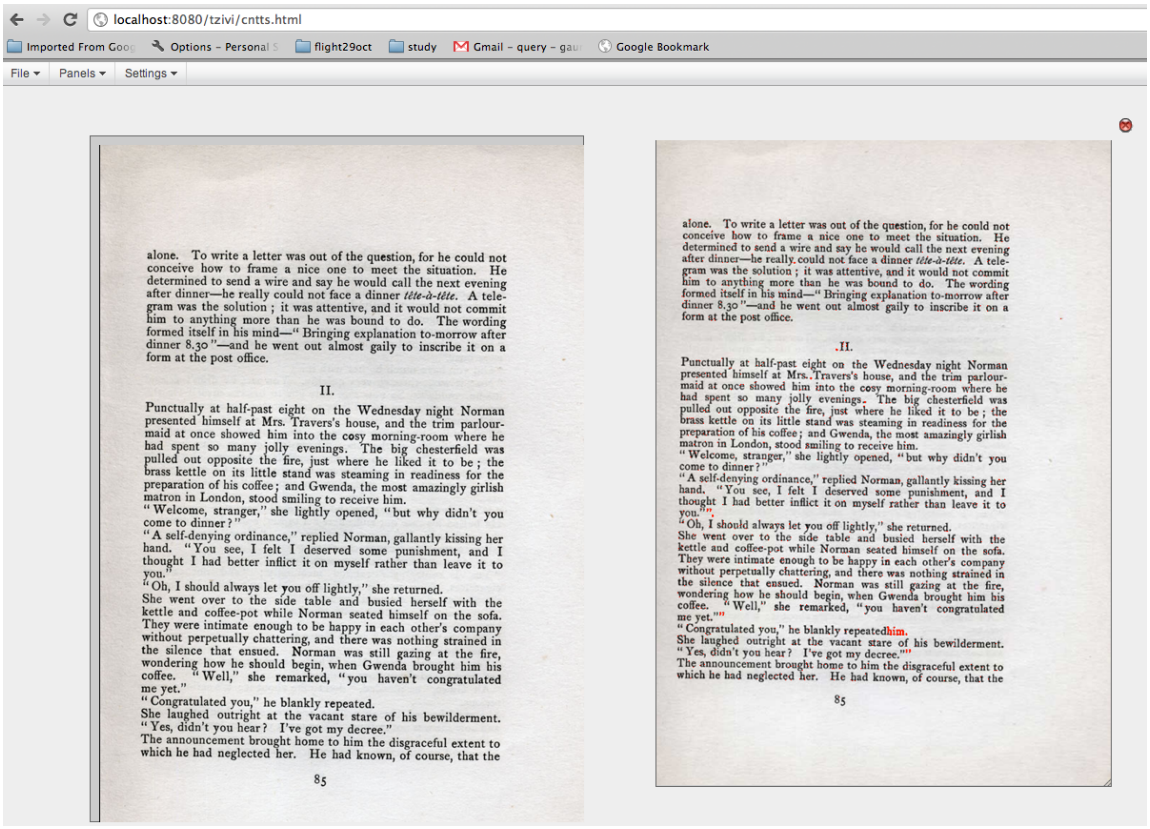


Figure 2.2: Screenshot of the collation result obtained using ImageMagick

The above methods work well only when the images are pre-registered and hence require the user to manually change the scale and rotation of the pages and wont be practical if the pages have different alignments and different font-sizes. Consequently we used image processing techniques and image matching algorithms to perform automated comparison of images. We followed an approach similar to [Yalniz and Manmatha, 2012] to compare word images amongst two scanned pages.

This approach uses the word bounding-box information to compare word shapes with one another and then uses context information to filter matches and find the exact match for that word. Thus if there is no exact match for a word it is highlighted as a difference. In this approach first we segment the words out of a scanned page-image (we wrote our own segmentation code for this purpose which worked well for one book but not for some books, so then we used Abbyy Fine Readers segmentation output because word-segmentation is an easier problem than OCR and the standard solutions for this work pretty well). We first pass all the images to the Abbyy Reader, which generates a DJVU format XML file which contains the coordinates for every word in that image. Then the corner key points for every word are extracted using the FAST algorithm. Before that, we first convert the image to grayscale, apply Gaussian blur and binarize it using a threshold. We noticed that we need to blur the image again after binarization as the number of detected corner points remains low if we dont blur it again. Then we calculate the SIFT feature vectors for all these key points. A subset of these feature vectors are then used to create a vocabulary tree using hierarchical kmeans algorithm. The rest of the vectors are then quantized to the nearest centroid in this tree. Thus for every word image weve obtained a sequence of VISterm IDs which depict the cluster IDs of the feature vectors. This sequence of vis-terms for every word image is stored in a text file in the server. A typical text file looks like a dictionary with word image number as key and value as a vector of corresponding cluster IDs, for example:

- 1.tif 120 130 1 11 1233 1212
- 2.tif 121 111

,

,

- 4190.tif 121 3434 2112 1212 13 3434 121 99

In the final step, the system takes any two page images as an input and starts comparing each word in that page to every word in the other page to find the most matching word. To calculate this, we use a combination of two scores - coverage score and configuration matching score. The coverage score between two words (x, y) is the ratio of matching vis-terms to the number of vis-terms, adjusted by multiplying it with the ratio of sizes of the two words:

coverage score = ((match/size1+match/size2)/2)*width-weight

where,

match = number of matching vis-terms

size1 = number of vis-terms in word1

size2 = number of vis-terms in word2

Width-weight = ratio of width of the two words

Using this coverage score we filter out top ten words for every word in the query page and calculate the top five matches for these using the configuration score. Configuration score is the ratio of longest common subsequence of cluster IDs between any 2 images to the number of key points in the query image. To make the calculation of LCS faster we remove those vis-terms from the sequence that are not present in both sequences as they are not going to affect the LCS size. After getting the configuration score, we devise a final matching score between the two words by a weighted sum of the configuration score and coverage score:

Final Score = (Lambda)* Configuration score + (1-Lambda)*Coverage score

For deciding the number of clusters in this step, we tried a statistical approach. We plotted a graph for the coverage score for all the words in one page for a particular cluster number and compared with another, as shown in figure 2.3

The chart shows that the coverage scores almost peak around 350 clusters and are almost same for 250, 300 and 350 clusters. Hence, we decided to choose 350

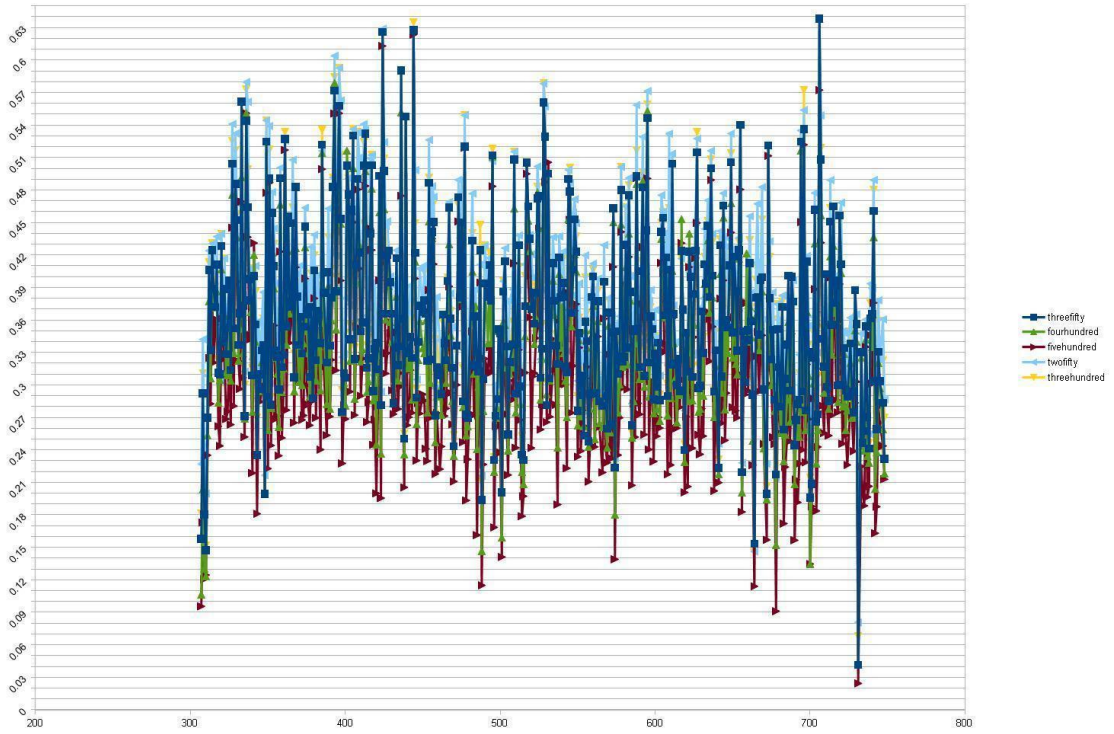


Figure 2.3: Graph showing the variation in coverage score of all words with number of clusters

clusters.

Thus we obtain and store the top ten matches for every word and use the positional context information to find if any of the top ten matches fits into the surrounding context of the word. Else we take that word as a difference. We tried two different approaches for the positional context part, which I explain in detail below:

1. First we calculate an offset of match for the first five words in original document. If the offset is positive we conclude that the target document contains a part of the original document and it starts somewhere after the beginning of the target document. If it is negative we can say that part of the query document is contained in the target document and we find where in the query document this part starts.

To find this offset we make all possible patterns with the top five matches of the first five words and see which of the patterns fall into a continuously increasing sequence with an increment of one. For this we find the length of the LCS of every possible pattern with the pattern [1, 2, 3, 4, 5] and return the pattern with the longest length.

Now once we have an offset I start with rest of the words in the query document and for every query word we look if any of the top ten candidates lie between the offset + query offset +- error tolerance . Here query offset is the position of the query word w.r.t. the first word in the query image. If any of the candidates falls within this range then we take it as the best match for that query word. If none of the candidate satisfies this condition then we assume the query word is a difference.

This approach seems to work fine with simple cases where the text is almost similar and the only major task is to find the offset. But there can be cases where even after finding the offset we are not guaranteed to find the best match as there may come a few dozen additional words after a sequence of correct matches and it will be difficult to discern where this ends.

2. In this approach, we take every six consecutive words and label it as a query pattern of [1, 2, 3, 4, 5, 6]. Then we take top ten candidates of each of these words and make all possible combinations of these matches which results in about 6$\hat{6}$10 such patterns.

Now we take the length LCS of each of the remaining patterns with our query pattern of [1,2,3,4,5,6] and return the pattern with the highest length of LCS. Then we look at the result pattern and see which of the members is a match with the query pattern or is in close vicinity to be a match. Then we map the ones which have a match to the query word and highlight the rest as differences.This approach seems to have a very high accuracy but is slow mostly because of the high number of possible result patterns. One approach to solve this is to filter the number of patterns by considering patterns that fall within a certain range.

Figure 2.4 shows the effect of clustering the key points.

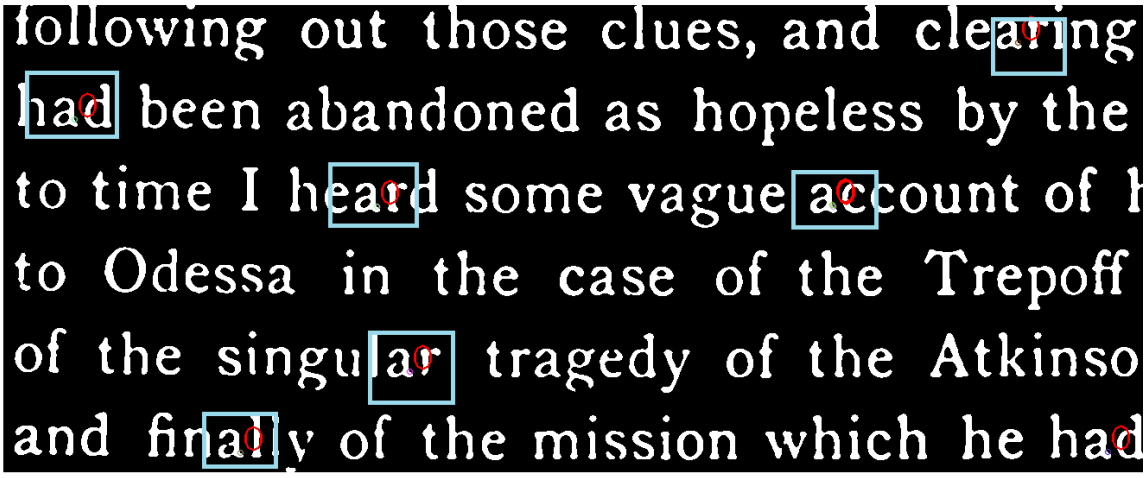


Figure 2.4: The outlined boxes show the keypoints in the same cluster

As can be seen in the figure all these outlined points belong to the same cluster

and they represent the same shape which is the bottom right curve of a in this case.

2.1 Integration into CritSpace

As Peter Robinson notes, the single greatest effect of the digital revolution is that it is empowering a new model of collaboration, and hence new modes of readership and study, among scholars, and between scholars and readers.[Robinson, 2009] In sync with this, the broad goal of the project was to integrate this tool into the creativity support environment CritSpace as its usefulness would be greatly enhanced when used in conjunction with such a tool.

CritSpace [Audenaert, 2011] is a creativity support environment which uses spatial information management strategies [Marshall and Shipman III, 1997] as one direction for supporting the early stages of humanities scholarship along with some supporting technologies. It was designed to support analysis by digital scholars during open-ended research tasks. It is a platform for building web-based visual interfaces which can be integrated into existing digital libraries easily. The interface can be easily customized by an institution to fit a particular groups specific needs.

In CritSpace [Figure 2.5], a workspace is the top-level unit of work created by users and provides the display context for rendering and interacting with panels. The base panel object provided by the CritSpace framework communicates basic information about its current state using the repository proxy. Any number of custom panels can be added and the CritSpace framework provides the functionality to do so easily.

The user-interface was planned keeping in mind the needs of the digital scholars so that an effortless user-experience could be generated. In the user-interface in brief below,

A new collation panel was added to the existing CritSpace environment. A Start compare button was added in a default container panel. On clicking this button,

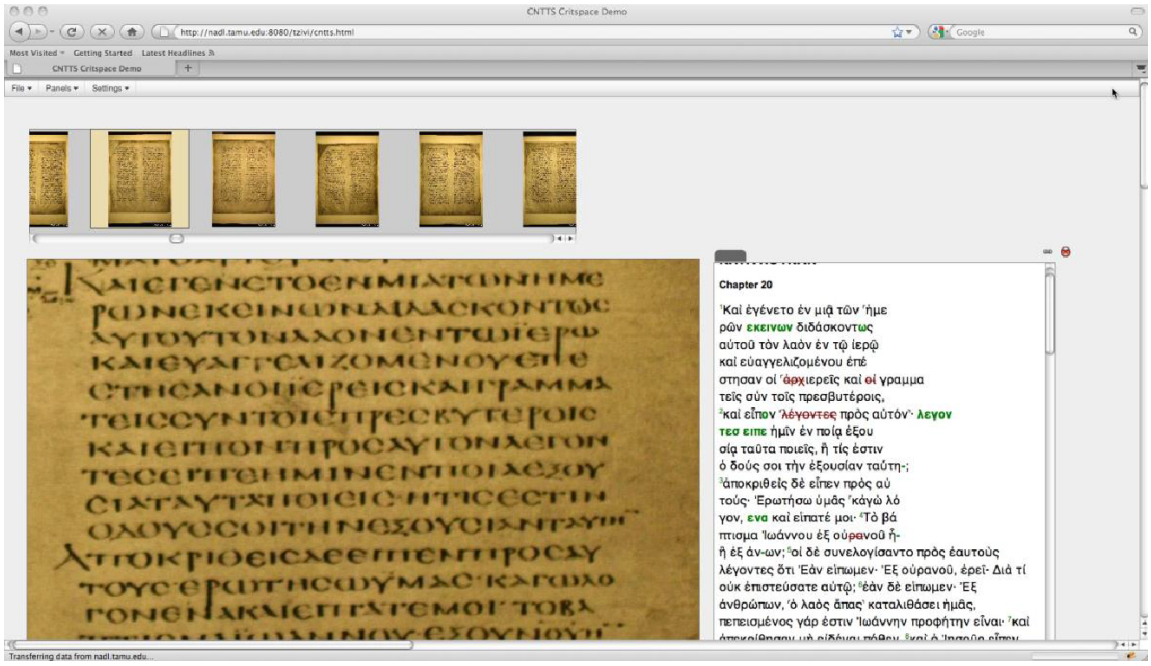


Figure 2.5: Sample workspace with a text panel, image panel and facsimile viewer

two tzivi-image panels pop out which have two page images selected by default. At this point the differences in both the pages will be highlighted around the bounding regions[Figure 2.6].

The benefit of using the tzivi panel is that the scholar can zoom into any part of the page-image to analyze the structure of the word. In addition, a dial was added onto both of the panels to aid the scholar in selecting particular page-images in the book.

The tool also has a feature to track the matches for any word on any of the page images. On switching on the Enable Tracking mode whenever a user hovers over a word in one of the page images, its best match (or best n matches) is highlighted in all the other panels [Figure 2.7]. Thus this feature will also act as a good evaluation tool to verify the accuracy of the matching.

We also added another feature to support adding annotations to a particular word

in any of the pages. Once the scholar enables annotation mode and clicks on a word a box will appear above it where the scholar can type his thoughts [Figure 2.8]. Work can be done to export these annotations to the server in a particular format so that they can be viewed whenever the user visits the workspace again.Another feature was added whereby the user could select any rectangular region in one of the pages by mouse clicking and the differences within that rectangle would be highlighted in both the pages.

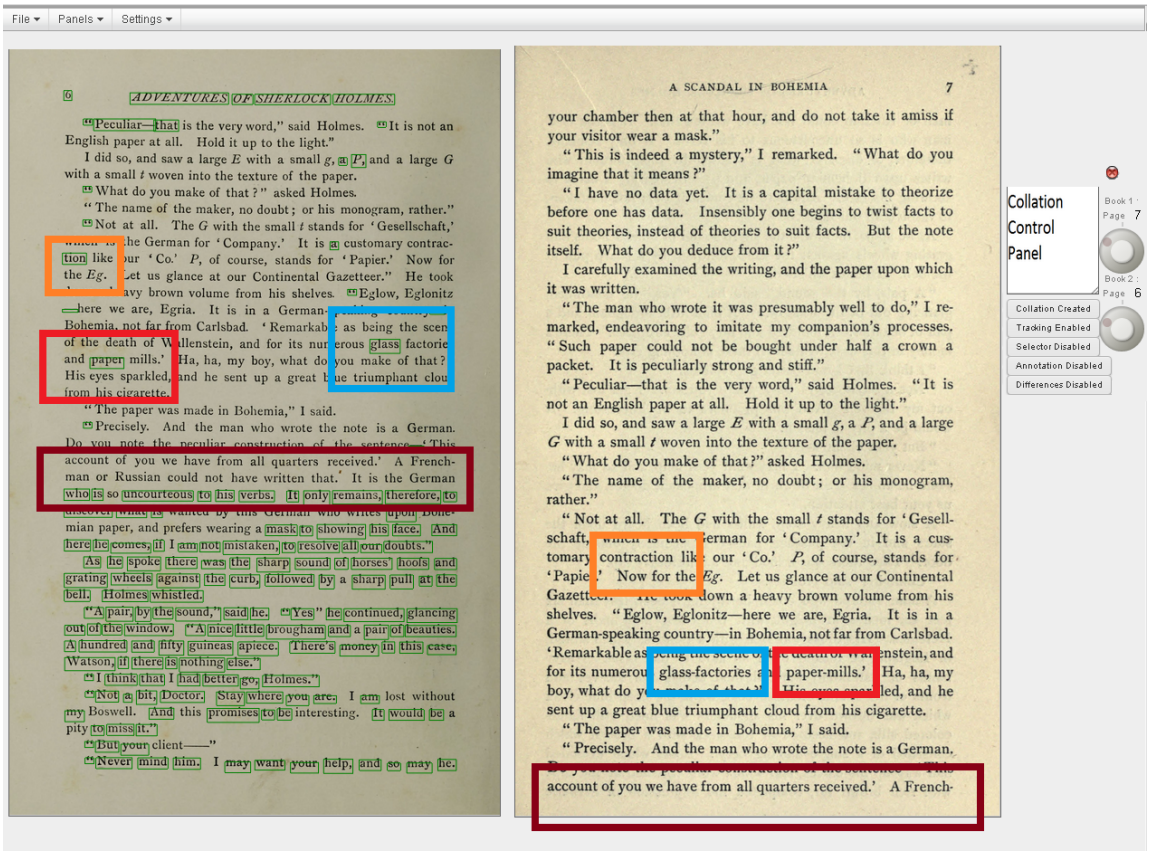


Figure 2.6: Screenshot highlighting the differences in green. Notable differences like missing hyphens are outlined.

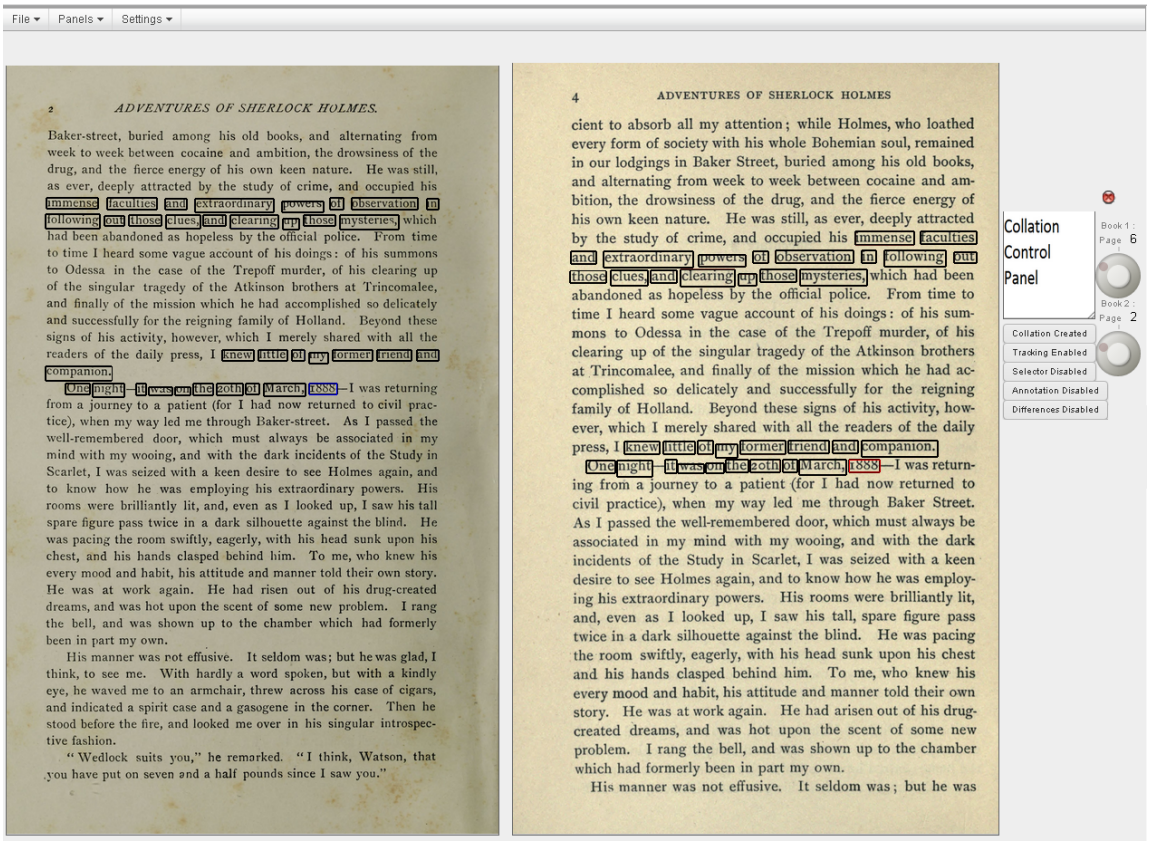


Figure 2.7: Screenshot demonstrating the tracking feature. When the user hovers the mouse over any block of word its corresponding match is highlighted in the other page in red. The ones which have already been checked are turned black

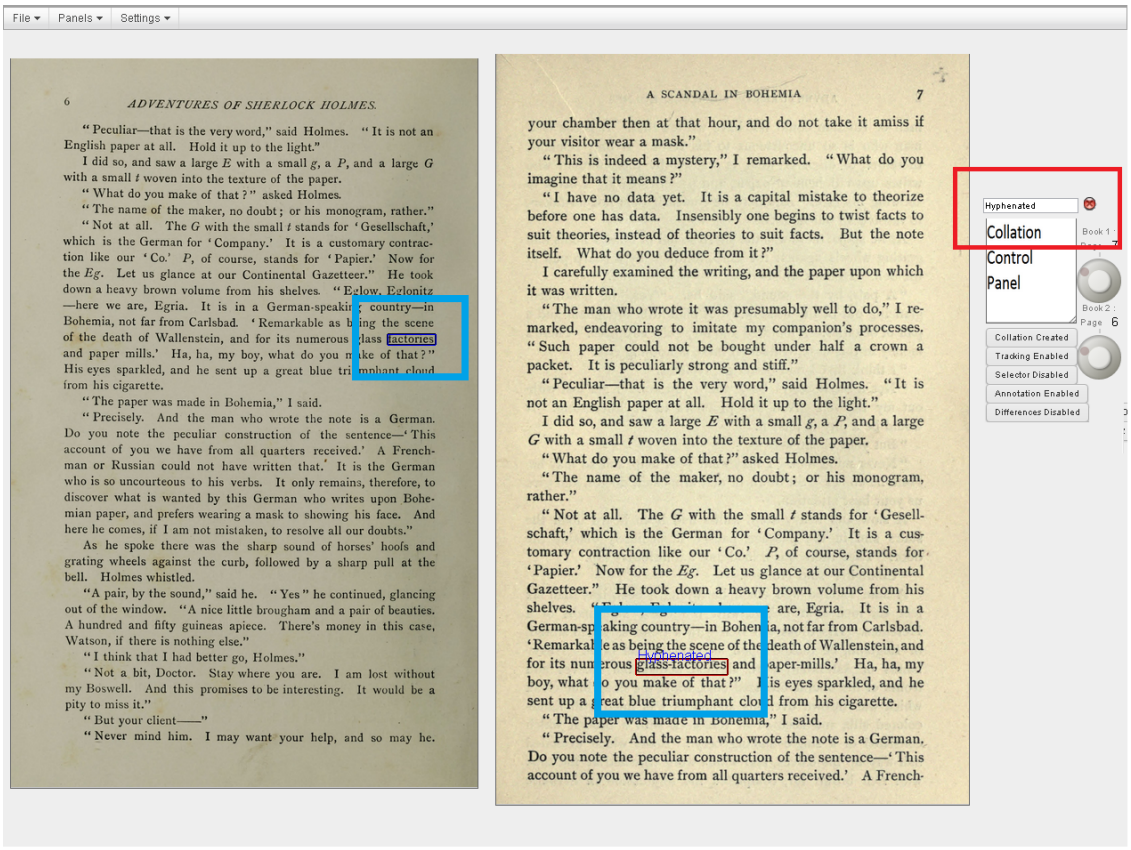
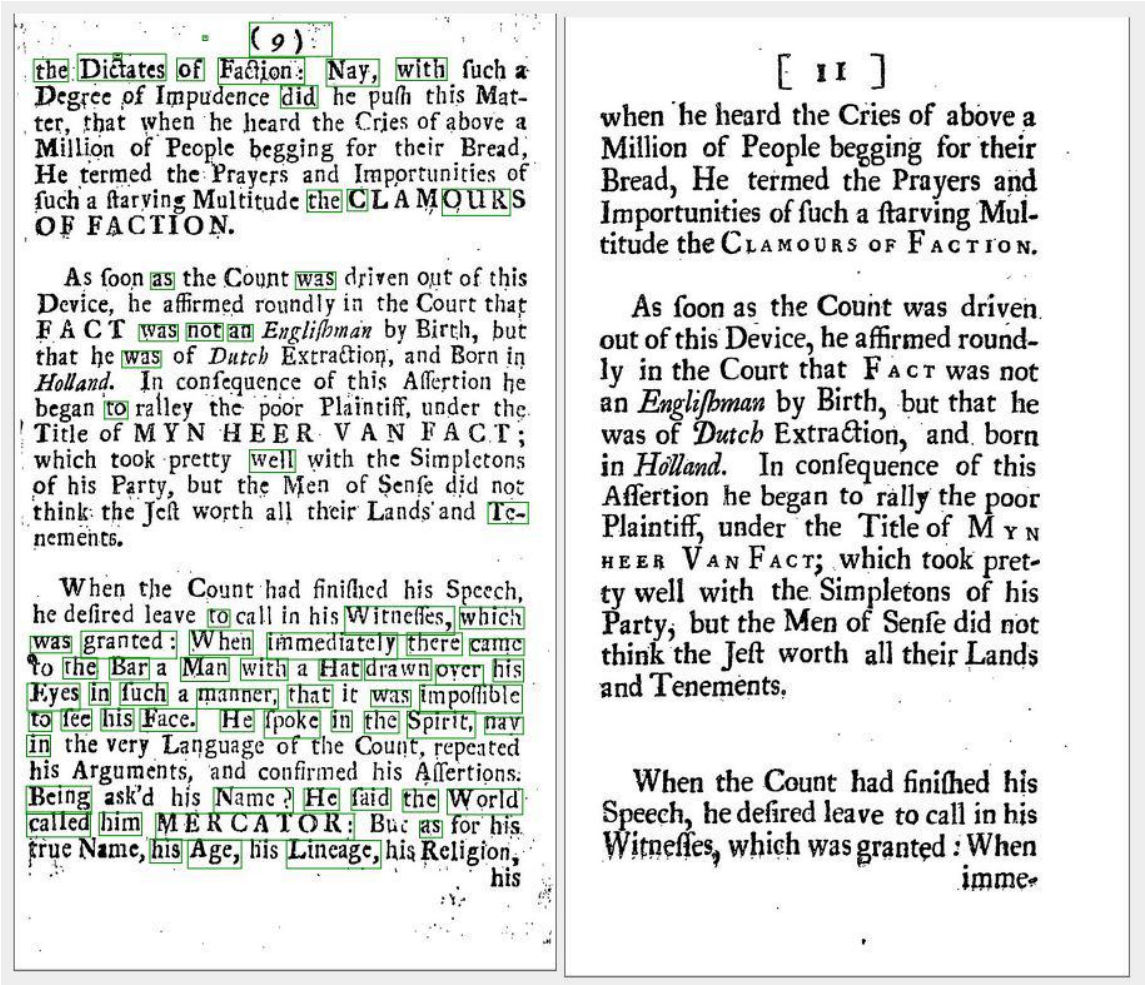


Figure 2.8: Screenshot of the annotation feature. On enabling annotation mode, the user can select a word and a text box will appear. The text is displayed above the word every time annotation mode is set. A sample use-case has been outlined.



(9)

the Dictates of Faction: Nay, with ſuch a Degree of Impudence did he puſh this Matter, that when he heard the Cries of above a Million of People begging for their Bread, He termed the Prayers and Importunities of ſuch a ſtarving Multitude the CLAMOURS OF FACTION.

As ſoon as the Count was driven out of this Device, he affirmed roundly in the Court that FACT was not an *Engliſhman* by Birth, but that he was of *Dutch* Extraction, and Born in *Holland*. In conſequence of this Aſſertion he began to ralley the poor Plaintiff, under the Title of MYN HEER VAN FACT; which took pretty well with the Simpletons of his Party, but the Men of Senſe did not think the Jeſt worth all their Lands and Tenements.

When the Count had finiſhed his Speech, he deſired leave to call in his Witneſſes, which was granted: When immediately there came to the Bar a Man with a Hat drawn over his Eyes in ſuch a manner, that it was impoſſible to ſee his Face. He ſpoke in the Spirit, nay in the very Language of the Count, repeated his Arguments, and confirmed his Aſſertions. Being ask'd his Name? He ſaid the World called him MERCATOR: But as for his true Name, his Age, his Lineage, his Religion, his

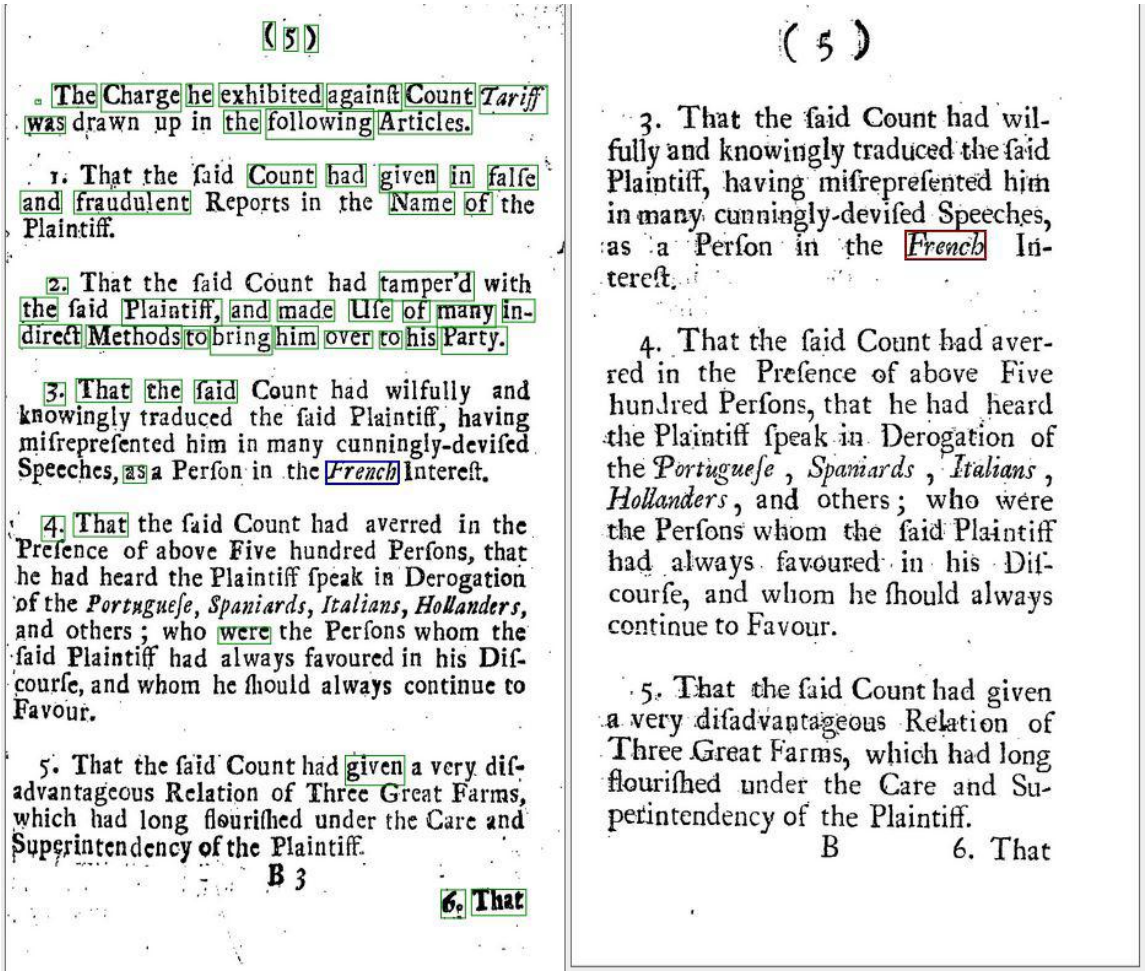
[11]

when he heard the Cries of above a Million of People begging for their Bread, He termed the Prayers and Importunities of ſuch a ſtarving Multitude the CLAMOURS OF FACTION.

As ſoon as the Count was driven out of this Device, he affirmed roundly in the Court that FACT was not an *Engliſhman* by Birth, but that he was of *Dutch* Extraction, and born in *Holland*. In conſequence of this Aſſertion he began to rally the poor Plaintiff, under the Title of MYN HEER VAN FACT; which took pretty well with the Simpletons of his Party, but the Men of Senſe did not think the Jeſt worth all their Lands and Tenements.

When the Count had finiſhed his Speech, he deſired leave to call in his Witneſſes, which was granted: When imme-

Figure 2.9: Screenshot of collation output of two 17th century versions of The Late Tryal and conviction of Count Tariff.



(5)

The Charge he exhibited againſt Count *Tariff* was drawn up in the following Articles.

1. That the ſaid Count had given in falſe and fraudulent Reports in the Name of the Plaintiff.

2. That the ſaid Count had tamper'd with the ſaid Plaintiff, and made Uſe of many indirect Methods to bring him over to his Party.

3. That the ſaid Count had wilfully and knowingly traduced the ſaid Plaintiff, having miſrepreſented him in many cunningly-deviſed Speeches, as a Perſon in the *French* Intereſt.

4. That the ſaid Count had averred in the Preſence of above Five hundred Perſons, that he had heard the Plaintiff ſpeak in Derogation of the *Portugueſe*, *Spaniards*, *Italians*, *Hollanders*, and others; who were the Perſons whom the ſaid Plaintiff had always favoured in his Diſcourſe, and whom he ſhould always continue to Favour.

5. That the ſaid Count had given a very diſadvantageous Relation of Three Great Farms, which had long flouriſhed under the Care and Superintendency of the Plaintiff.

B 3

6. That

(5)

3. That the ſaid Count had wilfully and knowingly traduced the ſaid Plaintiff, having miſrepreſented him in many cunningly-deviſed Speeches, as a Perſon in the *French* Intereſt.

4. That the ſaid Count had averred in the Preſence of above Five hundred Perſons, that he had heard the Plaintiff ſpeak in Derogation of the *Portugueſe*, *Spaniards*, *Italians*, *Hollanders*, and others; who were the Perſons whom the ſaid Plaintiff had always favoured in his Diſcourſe, and whom he ſhould always continue to Favour.

5. That the ſaid Count had given a very diſadvantageous Relation of Three Great Farms, which had long flouriſhed under the Care and Superintendency of the Plaintiff.

B 6. That

Figure 2.10: Collation output of another pair of pages from The Late Tryal and conviction of Count Tariff.

2.2 Dataset

We tested the vHinman tool on various scanned texts available on the Internet Archive website and within TAMU collections. These include digital copies of Sherlock Holmes, The Late Tryal[Figure 2.9,Figure 2.10] and conviction of Count Tariff and multiple editions of poems of John Donne. These works have many print and edition variants and are suitable samples for collation work. The accuracy in tracking the matches is very high for Sherlock Holmes and John Donnes poems at above 90%.The accuracy for the work of The Late Tryal and Conviction of Count Tariff is also above 80% which is good considering there are font variations in its multiple copies .For example the words French which are shown to be matching in Figure 2.11 have font variations as shown in the figure 2.12.

Figure 2.11: Font variations in two versions of word "French". This version doesn't have long endings in its letters.

Figure 2.12: This version has long endings in its letters.

The poems of John Donne and the work of The Late Tryal were written in 17th century,hene the accuracy in matching is respectable considering the OCR accuracy for these books is not good. The copies of Mary Shelleys Frankenstein obtained from the Internet Archive were also tested with the tool but the accuracy isnt as good which is probably because of the vast variations in the fonts of the two copies. The current tool can be good for collating editions with similar fonts but new approaches can be tried for getting higher accuracy with vastly different fonts.

3. USER EVALUATION

A user study was conducted to evaluate the usefulness of the tool. We contacted many researchers at our university for this evaluation. Five subjects were chosen to participate in this study [Table 3.1] which was a mix of semi-structured interview regarding the experience of scholars on collation, followed by a demo of the prototype and questions about the feedback of the tool and suggestions for its improvement.

Most of the subjects had prior experience with collation either in their scholarly research or for some classroom activities. Some of the subjects had used the mechanical collators like Hinman or Lindstrand for their work but found them to be very cumbersome to use and stressful to the eyes. Also they agreed that these tools are only useful if the concerned text can be aligned easily which is often not the case. Some of them had used the software based-collators like JUXTA but mostly dont find it so useful because of the inherent OCR or transcription errors that arise in the documents.

Many of the subjects still prefer the paper-based manual collation method because they find the supporting tools either inaccurate or too cumbersome to use or both. The need for collation in the subjects research varied from the traditional scholarly editing process to bibliographic research and book history research.

Table 3.1: Demographics of the user study participants

ID	Area of Interest	Career Stage
S1	Eighteenth Century Literature	Senior
S2	Bibliography	Senior
S2	Scholarly Editing	Senior
S2	Scholarly Editing	Senior
S2	Book History, Linguistics	Senior

S4 pointed out that he didnt have the resources to do the transcription for each of the documents he works on and also said that they are prone to errors. S1 pointed out the need to be able to find differences in font-styles, ligatures like the move from using the long s to the current s. S2 liked the idea of integrating the collator into CritSpace, which can foster collaborative work. She also liked the idea that the tool could have multiple panels (more than two). She pointed out that while supporting multiple images we can display the n-images in the form of medium sized thumbnails as is seen in Google images, where the scholar can select any two panels to collate at a time. She noted that the tool could bring forward new uses of collation and could get collation adopted by scholars who currently dont focus much on it attributing the manual effort and inherent inaccuracies in the current method.

S5 suggested a novel use of the tool in verifying the authorship of a poem. For this, he said we can look at the frequency of the common words used by that author and see if the frequency in the query poem matches with the authors generally known frequency of these words in his well-known poems. Another property that could be looked up is the average distance of the same word in the documents as a particular author used to have a known pattern of repetition of particular words.

Some of the subjects felt the need to be able to point small differences like punctuation because this is important for a critical edition. Although our tool currently supports identifying only word differences, punctuation support can be added. S4 felt that the current implementation can quicken the collation process by addressing textual differences while punctuation can be addressed separately. The subjects in general liked the ability to use the original facsimile of the document via the tool rather than a transcription or a somewhat inaccurate OCR version of it.

Most of the subjects really liked the tool and could think of ways in which the tool could be useful in scholarly research. These ways range from figuring out the

authorship of a work to making a critical edition of a work to book history research. They feel that such a tool could save lots of dull manual effort. The subjects in general liked the ability to use the original facsimile of the document via the tool rather than a transcription or a somewhat inaccurate OCR version of it. In conclusion, we found that the tool has huge potential and can revolutionize the current collation process if the accuracy is high for all kinds of documents.

4. CONCLUSION

This work has investigated the way humanities scholars perform collation work and what role does collation play in their research output. Collation is known to be a laborious and monotonous task unaided by technology so far. To address this problem, a prototype was developed to perform collation in an automated manner so that the scholars dont have to go through the dull manual collation or the mentally straining mechanical collators. Image matching techniques are employed in building this prototype so that the scholars can directly use the original facsimiles of the documents rather than the OCR output or the transcriptions of the documents, which may be somewhat inaccurate. The tool was integrated into the creativity support environment CritSpace, which uses spatial hypertext strategies to support the early stages of humanities scholarship. This provided a web-interface for the digital collator tool thus enabling collaboration among scholars, which can be a heavy asset in scholarly research. Finally, a user evaluation was conducted where scholars with prior collation experience were selected. The prototype of the tool was demonstrated and a semi-structured interview was conducted to judge the usefulness of the tool and understand the way they perform their research.

In summary, the tool looks very promising to the scholars and also has a high accuracy rate for the books tested so far. This kind of tool can save a massive amount of time for scholars and set up a paradigm of digital collation encouraging even more scholars in finding new uses of collation in their work. It extends the Hinmans principles by allowing collating multiple editions of a book in addition to multiple copies of same edition having minor differences. Since it is has application in creating a critical edition, bibliography and book history research, this tool has

the capability of gaining widespread adoption.

4.1 Future Work

Beyond printed material, it will be interesting to evaluate the tool for handwritten documents and make it robust for such documents. Also it will be great to test the tool for non-English documents. We can try out different visualization formats for ways the scholars can use the output in their work. A detailed usability study can be conducted where scholars can perform some real collation work on few pages and compare their traditional method and the vHinman. Also the accuracy could be tested for warped images as most of the unobtrusive scanning methods produce some warping on the images. Also we can use a GPU implementation of SIFT, which can greatly speed up the processing time for a page which will be useful in case of large books.

REFERENCES

[Audenaert, 2011] Audenaert, N. (2011). *CritSpace: An Interactive Visual Interface to Digital Collections of Cultural Heritage Material*. PhD thesis, Texas A&M University, College Station, Texas, USA.

[Audenaert and Furuta, 2010] Audenaert, N. and Furuta, R. (2010). What humanists want: how scholars use source materials. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 283–292, Gold Coast,Australia. ACM.

[Audenaert et al., 2010] Audenaert, N., Lucchese, G., and Furuta, R. (2010). Critspace: a workspace for critical engagement within cultural heritage digital libraries. In *Research and Advanced Technology for Digital Libraries*, pages 307–314, Glasgow,UK. Springer.

[Bibliographical-Mirrors, 1999] Bibliographical-Mirrors (1999). In *Lehigh University Information Resources*, Bethlehem, Pennsylvania,USA.

[CDH, 2009] CDH (2009). Sapheos. `http://sapheos.org/`. University of Southern Carolina.

[CDH, 2012] CDH (2012). Paragon. `http://cdh.sc.edu/paragon`. University of Southern Carolina.

[ImageMagick, 2012] ImageMagick (2012). `http://www.imagemagick.org/`. ImageMagick Studio LLC.

[Kuhn, 2010] Kuhn, J. (2010). "A hawk from a handsaw:" collating possibilities with the Shakespeare Quartos Archive. In *Renaissance Society of America conference*, Universit ca' Foscari, Venice, Italy.

[Marshall and Shipman III, 1997] Marshall, C. C. and Shipman III, F. M. (1997). Spatial hypertext and the practice of information triage. In *Proceedings of the eighth ACM conference on Hypertext*, pages 124–133, Southampton, UK. ACM.

[NINES, 2011] NINES (2011). Juxta. `http://www.juxtasoftware.org/`.

[Raabe, 2008] Raabe, W. (2008). Collation in scholarly editing: An introduction. `http://wraabe.wordpress.com/2008/07/26/collation-in-scholarly-editing-an-introduction-draft`.

[Robinson, 1994] Robinson, P. (1994). Collation, textual criticism, publication, and the computer. In *Textual Cultures*, volume 7, pages 77–94. Indiana University Press, Bloomington,Indiana,USA.

[Robinson, 2009] Robinson, P. (2009). Towards a scholarly editing system for the next decades. In *Sanskrit Computational Linguistics*, pages 346–357, Providence,Rhode Island ,USA. Springer.

[Schreibman, 2000] Schreibman, S. (2000). Versioning machine. `http://v-machine.org/`.

[Smith, 2002] Smith, S. E. (2002). "Armadillos of invention": A census of mechanical collators. In *Studies in Bibliography*, volume 55, pages 133–170. Bibliographical Society of the University of Virginia, Charlottesville,Virginia,USA.

[Unsworth, 2000] Unsworth, J. (2000). Scholarly primitives: what methods do humanities researchers have in common, and how might our tools reflect this. In *Humanities Computing, Formal Methods, Experimental Practice Symposium*, Kings College, London, UK.

[Yalniz and Manmatha, 2012] Yalniz, I. Z. and Manmatha, R. (2012). An efficient framework for searching text in noisy document images. In *Document Analy-*

sis Systems (DAS), 2012 10th IAPR International Workshop, pages 48–52, Gold Coast,Australia. IEEE.