FEASIBILITY OF MELVILLE MARGINALIA AUTHORSHIP DIFFERENTIATION

by

Aaron Burdin

A thesis
submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Electrical Engineering
Boise State University
August 2017

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the thesis submitted by

Aaron Burdin

Thesis Title: Feasibility of Melville Marginalia Authorship Differentiation

Date of Final Oral Examination: 8 May 2017

The following individuals read and discussed the thesis submitted by student Aaron Burdin, and they evaluated his presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Elisa H. Barney Smith, Ph.D.	Chair, Supervisory Committee
Steven Olsen-Smith, Ph.D.	Member, Supervisory Committee
Said Ahmed-Zaid, Ph.D.	Member, Supervisory Committee

The final reading approval of the thesis was granted by Elisa H. Barney Smith, Ph.D., Chair, Supervisory Committee. The thesis was approved by the Graduate College.

ACKNOWLEDGMENTS

I would to thank Elisa H. Barney Smith, Ph.D. for her tireless patience and encouragement. I would also like to thank Steven Olsen-Smith Ph.D. for his insight into Melville's world.

ABSTRACT

Aaron Burdin, Electrical Engineering, Boise State University

Abstract of Bachelors Thesis, Submitted 8 May 2017:

Feasibility of Melville Marginalia Authorship Differentiation

We examine the feasibility of using image processing techniques to determine differentiation in authorship of historical pencil marks. Pencil marks with unattributed and attributed authorship are segmented from digital images of historical books. Analysis is performed on five features that are extracted from the ”vertical” pencil marks, with those features used as a basis for authorship of marks. These marks consist of single stroke marks that are interspersed in the same document. We describe the challenges of the digital format that we were given and the steps taken in using autonomous segmentation to save pixel locations of marks. Five mark features are chosen and extracted: Average Intensity, Stroke Width, Blurriness, Stroke Curvature, and Stroke Angle. Features are then analyzed with the use of different histograms, 2D scatter plots of feature space, and comparing and contrasting the two groups of marks. C-means clustering is performed on the feature spaces of both groups. Semi-supervised clustering is used to test if we can predict the clustering. We then use two forms of cluster validity, Davies-Bouldin Index and Silhouette, in order to produce a confidence value on the number of clusters and their membership. Then we look at the histograms and 2D scatter plots with the Melville's Marginalia Online attributed and unattributed labels applied. Extracting features show patterns and trends within the marks that could be used to group marks. Specifically, Stroke Curvature became a dominant feature that showed promises of differentiating marks created by different authors. Extracting features has the potential to be used with high confidence in separating marks by author.

TABLE OF CONTENTS

ABSTRACT . . . v

LIST OF FIGURES . . . ix

1 Introduction . . . 1

2 Methods and Tools . . . 5

2.1 Image Data . . . 6

2.2 Mark Segmentation . . . 8

2.2.1 Unattributed Mark Segmentation . . . 9

2.2.2 Marks from the Attributed Book . . . 15

2.3 Categorizing and Mark Distributions . . . 17

2.4 Feature Extraction . . . 19

2.4.1 Average Intensity . . . 20

2.4.2 Stroke Width . . . 20

2.4.3 Blurriness . . . 21

2.4.4 Stroke Curvature . . . 23

2.4.5 Stroke Angle . . . 25

2.5 Clustering . . . 27

2.5.1 Visual Clustering . . . 28

2.5.2 C-means Clustering . . . 28

2.5.3 Cluster Validity . . . 30

3 Experiments 33

3.1 Individual Feature Distribution Analysis 33

3.1.1 Possible Multimodal Distributions 34

3.1.2 Average Intensity 34

3.1.3 Stroke Curvature 36

3.2 Comparing Histograms of Attributed and Unattributed Marks 39

3.2.1 Average Intensity 39

3.2.2 Stroke Width 39

3.2.3 Blurriness 40

3.2.4 Stroke Curvature 41

3.2.5 Stroke Angle 42

3.3 2D Feature Space: Comparing Attributed and Unattributed Marks . . . 44

3.4 C-means Clustering Results 53

3.5 Semi-Supervised Learning 54

3.6 Melville's Marginalia Online Attributed Labels 57

3.6.1 Book Sections 58

3.6.2 Assigned and Unassigned 63

4 Conclusion 68

4.1 Summary of Work 68

4.2 Conclusion 69

4.3 Future Work 71

REFERENCES 74

LIST OF FIGURES

1.1 Example of pages from both book sources **a)** Dante's The Vision, **b)** Shakespeare Anthology 4

2.1 Example of area of useful information 7

2.2 **a)** Unattributed mark zoom, **b)** Attributed mark zoom 8

2.3 **a)** Original page image, **b)** Enlarged view of mark 10

2.4 **a)** Image converted to gray scale (intensity), **b)** Enlarged view of mark 11

2.5 **a)** Global threshold of > 80 was performed, **b)** Enlarged view of mark area 11

2.6 **a)** Dilation: 255×255 structuring element, **b)** Dilation with 25×25 structuring element 12

2.7 **a)** Full page, background reference area was within the red rectangle, **b)** Background reference area 12

2.8 **a)** Text pixels assigned the intensity value of the background, **b)** Enlarged view of mark 13

2.9 **a)** Tsinghua adaptive thresholding applied, **b)** Enlarged view of mark . 14

2.10 **a)** After application of median filter: 5×5 structuring element, **b)** Enlarged view of mark 14

2.11 **a)** Dilation: 15×15 structuring element, **b)** Enlarged view of mark . . . 15

2.12 **a)** Erode: 12×12 structuring element, **b)** Enlarged view of mark 15

2.13 Mark isolated from page 16

2.14 Flowchart showing steps for mark segmentation 16

2.15 Example of Foxing on page 17

2.16 Example of types of marks **a)** Horizontal **b)** Multi-stroke, **c)** Symbol, **d)** Hand writing 18

2.17 Example of mark intensity range. **a)** Unattributed mark in RGB, **b)** Grayscale average intensity:145.5, **c)** Attributed Mark in RGB, **d)** Grayscale average intensity:179.5 21

2.18 Thinning Structuring Element. **a)** Starting, **b)** Rotated 45^o, **c)** Rotated 90^o 22

2.19 Example of mark being skeletonized. **a)** Initial mark, **b)** 3 iterations, **c)** 6 iterations, **d)** stroke width of 9 22

2.20 Example of mark being skeletonized, **a)** mark, **b)** 1 iteration, **c)** 3 iterations, **d)** 4 iterations 23

2.21 Example of mark being skeletonized. **a)** mark, **b)** 4 iterations, **c)** 7 iterations, **d)** 11 iterations 23

2.22 Blurriness reference **a)** Mark, **b)** Histogram of mark intensities, with 20, 60, and 80 percent shown, **c)** Mark by the pixels outside 20 and 80 percent removed 24

2.23 Example of types of paths a stroke can take **a)** Snake-like, **b)** Small tick, **c)** Check, **d)** Bracket 25

2.24 Example of curve feature measurement, stroke curvature=24.6, **a)** Original mark, **b)** Longest Line, **c)** Longest line superimposed on original mark, **d)** Connected **e)** Filled area, **f)** Filled area superimposed on original mark 26

2.25 Example of curve feature, stroke curvature= 15.4078, **a)** Original mark, **b)** Longest Line, **c)** Longest line superimposed on original mark, **d)** Connected, **e)** Filled area, **f)** Filled area superimposed on original mark 27

2.26 Hotelling example used for stroke angle feature. Stroke Angle: **a)** 0.285, **b)** 0.0025 28

3.1 Histograms of features **a)** Blurriness, **b)** Stroke Angle, **c)** Stroke Width, **d)** Stroke Curvature, **e)** Average Intensity 34

3.2 Average Intensity Histograms **a)** Average Intensity, **b)** Labeled into four groups 35

3.3 Histograms of features a,b,c,d. Each was color-coded with the group from average intensity in Figure 3.2b **a)** Stroke Width, **b)** Stroke Curvature, **c)** Stroke Angle, **d)** Blurriness 36

3.4 Stroke Curvature Histograms **a)** Stroke Curvature, **b)** Labeled into two groups 37

3.5 Histograms of features a,b,c,d. Each was color-coded with the group from Stroke Curvature in Figure 3.4b **a)** Average Intensity, **b)** Blurriness, **c)** Stroke Angle, **d)** Stroke Width 38

3.6 Histogram of intensity of marks from attributed and unattributed marks, **a)** Attributed marks, **b)** Unattributed marks 40

3.7 Histogram of stroke width of attributed and unattributed marks, **a)** attributed mark, **b)** unattributed marks 40

3.8 Histogram of Blurriness of attributed and unattributed marks, **a)** Attributed mark, **b)** Unattributed marks 41

3.9 Unattributed marks with lowest Blurriness value, **a)** Blurriness: 0.549, **b)** Blurriness: 0.618, **a)** Blurriness: 0.624 41

3.10 Histogram of Stroke Curvature of attributed and unattributed marks, **a)** Attributed mark, **b)** Unattributed marks 42

3.11 Example of Unattributed marks with large Stroke Curvature values, **a)** CN: 40.2872, **b)** CN: 84.1715, **c)** CN: 101.2632 42

3.12 Histogram of Stroke Angle of attributed and unattributed marks, **a)** Attributed marks, **b)** Unattributed marks 43

3.13 Attributed and unattributed marks max valued Stroke Angle, **a)** Attributed marks, Stroke Angle: 0.571, **b)** Unattributed mark, Stroke Angle: 1.28 43

3.14 Comparing Unattributed and Attributed marks, Average Intensity vs. Stroke Width **a)** 2D plot of Average Intensity vs. Stroke Width, **b)** Unattributed mark, stroke width: 11, **c)** Attributed mark, stroke width: 8, **d)** Unattributed mark, stroke width: 5, **e)** Attributed mark, stroke width: 4 45

3.15 Comparing Unattributed and Attributed marks, Blurriness vs. Average Intensity **a)** 2D plot of Blurriness vs. Average Intensity, **b)** Unattributed mark, Blurriness: 0.694, **c)** Attributed mark, Blurriness: 0.847, **d)** Unattributed mark, Blurriness: 0.953, **e)** Attributed mark, Blurriness: 0.963 46

3.16 Comparing Unattributed and Attributed marks, Average Intensity vs. Stroke Curvature **a)** 2D plot of Average Intensity vs. Stroke Curvature, **b)** 2D plot of Stroke Curvature vs. Stroke Width, **c)** Unattributed mark, Stroke Curvature: 1.55, Average Intensity: 180, **d)** Attributed mark, Stroke Curvature: 1.87, Average Intensity: 160, **e)** Unattributed mark, Stroke Curvature: 101, Average Intensity: 158, **f)** Attributed mark, Stroke Curvature: 52.9, Average Intensity: 167 47

3.17 Unattributed marks with large Stroke Curvature **a)** Stroke Curvature: 105, **b)** Stroke Curvature: 111, **c)** Stroke Curvature: 108, **d)** Stroke Curvature: 108 48

3.18 Comparing Unattributed and Attributed marks, Stroke Curvature vs. Blurriness. **a)** 2D plot of Stroke Curvature vs. Blurriness, **b)** Unattributed mark, Stroke Curvature: 1.55, Blurriness: 0.953, **c)** Attributed mark, Stroke Curvature: 4.54, Blurriness: 0.963, **d)** Unattributed mark, Stroke Curvature: 47.6, Blurriness: 0.753, **e)** Attributed mark, Stroke Curvature: 23.6, Blurriness: 0.876 49

3.19 Comparing Unattributed and Attributed marks, Stroke Angle vs. Average Intensity. **a)** 2D plot of Stroke Angle vs. Average Intensity, **b)** Unattributed mark, Stroke Angle: 1.55, Average Intensity: 0.953, **c)** Attributed mark, Stroke Angle: 4.54, Average Intensity: 0.963, **d)** Unattributed mark, Stroke Angle: 47.6, Average Intensity: 0.753, **e)** Attributed mark, Stroke Angle: 23.6, Average Intensity: 0.876 50

3.20 Comparing Unattributed and Attributed marks, Stroke Curvature vs. Stroke Angle. **a)** 2D plot of Stroke Curvature vs. Stroke Angle, **b)** Unattributed mark, Stroke Curvature: 2.46, Stroke Angle: 0.042, **c)** Attributed mark, Stroke Curvature: 3.74, Stroke Angle: 0.055, **d)** Unattributed mark, Stroke Curvature: 97.2, Stroke Angle: 0.789, **e)** Attributed mark, Stroke Curvature: 24, Stroke Angle: 0.199 51

3.21 Comparing Unattributed and Attributed marks, Stroke Angle vs Blurriness. **a)** 2D plot of Stroke Angle vs Blurriness, **b)** Unattributed mark, Stroke Angle: 0.063, Blurriness: 0.974, **c)** Attributed mark, Stroke Angle: 0.004, Blurriness: 0.971, **d)** Unattributed mark, Stroke Angle: 1.16, Blurriness: 0.76, **e)** Attributed mark, Stroke Angle: 0.106, Blurriness: 0.847 52

3.22 Comparing Unattributed and Attributed marks, Stroke Angle vs. Stroke Width. **a)** 2D plot of Stroke Angle vs. Stroke Width, **b)** Unattributed mark, Stroke Angle: 0.089, Stroke Width: 5, **c)** Attributed mark, Stroke Angle: 0.082, Stroke Width: 5, **d)** Unattributed mark, Stroke Angle: 0.46, Stroke Width: 12, **e)** Attributed mark, Stroke Angle: 0.156, Stroke Width: 9 53

3.23 The two attributed marks that were placed into cluster 2 and their matched unattributed marks **a)** Attributed, **b)** Unattributed, **c)** Attributed, **d)** Unattributed 56

3.24 Histogram of features with book section labels applied **a)** Average Intensity, **b)** Blurriness, **c)** Mark Angle, **d)** Mark Curvature, **e)** Stroke Width 60

3.25 2D plots with book section labels applied **a)** Average Intensity vs. Stroke Width, **b)** Blurriness vs. Average Intensity, **c)** Blurriness vs. Stroke Width, **d)** Mark Curvature vs. Average Intensity, **e)** Mark Curvature vs. Blurriness 61

3.26 2D plots with book section labels applied, continued **a)** Mark Curvature vs. Mark Angle, **b)** Mark Curvature vs. Stroke Width, **c)** Mark Angle vs. Average Intensity, **d)** Mark Angle vs. Blurriness, **e)** Mark Angle vs. Stroke Width 62

3.27 Histogram of features with assigned and unassigned labels applied **a)** Average Intensity, **b)** Blurriness, **c)** Mark Angle, **d)** Mark Curvature, **e)** Stroke Width 65

3.28 2D plots with assigned and unassigned labels applied **a)** Average Intensity vs. Stroke Width, **b)** Blurriness vs. Average Intensity, **c)** Blurriness vs. Stroke Width, **d)** Mark Curvature vs. Average Intensity, **e)** Mark Curvature vs. Blurriness 66

3.29 2D plots with assigned and unassigned labels applied, continued **a)** Mark Curvature vs. Mark Angle, **b)** Mark Curvature vs. Stroke Width, **c)** Mark Angle vs. Average Intensity, **d)** Mark Angle vs. Blurriness, **e)** Mark Angle vs. Stroke Width 67

CHAPTER 1

INTRODUCTION

Herman Melville (1819-1891), an American author in the 19th century, is best known for his writing of Moby Dick which was published in 1851. By the end of his lifetime Melville had collected a library of around a thousand books, and had access to many more books by borrowing from others. Melville's works were not appreciated at his death, only much later when a "Melville Revival" happened in the time period of 1910-1920. This led to many of Melville's library books getting dispersed at the time of his death, either being sold or given away. Recently, some of these books have been brought back together in digital form by the Melville Marginalia Online project [13]. It was discovered that Melville would mark passages and even write notes in books as he read them. Dr. Steven Olsen-Smith, a faculty member at Boise State University, has determined that there exists a correlation between the comments and written marks that Melville put in books and his own writing [11]. Linking his marks to his own works allows people a glimpse into what he was thinking and his process for writing.

Dr. Steven Olsen-Smith has been able to distinguish distinct handwriting samples from different annotators, but attribution for many of the markings in the volume has proven inconclusive. While most of the books that have been collected have marks that are mostly attributed to Melville, some books have marks authored from other readers. One example is a copy of Dante's The Vision (The Divine Comedy) [14], [12] (page example in Figure 1.1a). A second book was provided, a Shakespeare anthology volume. This book only has marks that are attributed to Melville as determined by Melville's Marginalia Online. This book will be used as a reference point of Melville mark creation. An example page from this book is seen in Figure 1.1b.

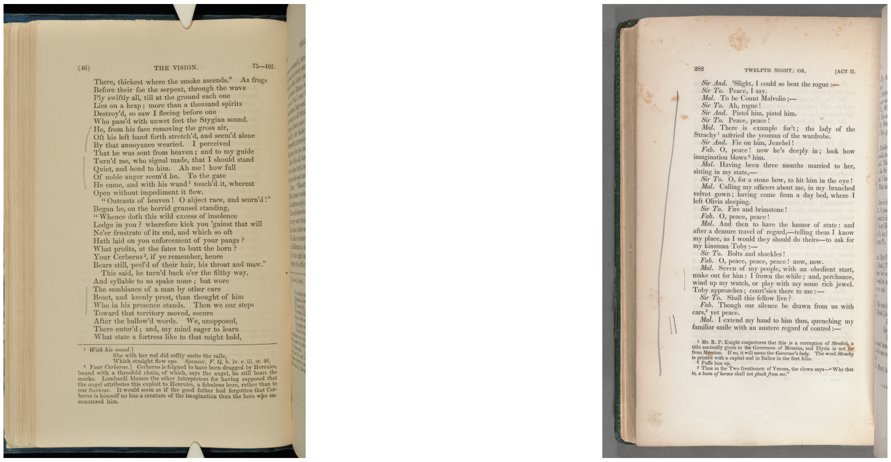
Freehand style writing has been shown to be unique to an individual [3]. Being able to distinguish between different authors is based on the hypothesis that each author's handwriting is unique. Work supporting this hypothesis has been done by Osborne [15], Huber [6], and Srihari [18]. Many aspects of life affect how we write: schooling, mood, age, time span, and whether writing from memory or making a copy. Authorship can be complicated when the authors share a similar time period (specifically during a time when handwriting form was strictly enforced), regional location, and standardized education [2].

When trying to identify or verify a writer, most techniques require either a large amount of text [16] such as a source document or from a longhand written document. With a source document the writer writes multiple copies of a standardized "letter" in order to reduce the complications associated with handwriting and to capture as many characters and character pairs as possible. Examples of such documents are the *London Letter* and *Dear Sam Letter* [15] along with the more modern *CEDAR-letter* [18]. With this baseline information, characters and character pairs are analyzed and compared. With historical documents, a generated source document is not possible. As such, features are extracted from the available written work.

When dealing with handwriting, typically characters and character pairs are analyzed in how they were formed. With handwriting there is a defined shape that a writer is creating and trying to convey that is constrained to the shapes defined by a given language. The author is attempting to recreate the character each time they write it. No external constraints are applied to making single stroke marks, instead the style and shape are self-constrained to the writing style of the author. Work similar to single stroke marks includes work done with Chinese handwriting recognition [19], but these are made from multiple strokes. Lutf's [8] work on Arabic handwriting identification looks at the diacritics, which are single strokes, but it takes multiples of them in combination to make an identification.

Melville could have created marks at different times in his life in the same book, making the age of a mark or writing instrument not sufficient to demonstrate the differences between annotators. Another potential problem was that a pencil was not a consistent writing instrument; there were variations in the writing as the pencil was worn down while used, thus changing the features of the mark as more marks are created.

The purpose of this thesis is to determine the feasibility of using image processing techniques in order to differentiate attributed marks from unattributed marks. In Chapter 2 (Methods and Tools) we will cover how the images were captured, how the marks were segmented from the page, categorizing marks, and then extracting the features that will be used in analysis. In Chapter 3 (Experiments) several methods of analysis are introduced. Mark features will have their histograms analyzed for patterns. Features from the books with unattributed and attributed marks will be compared, looking for commonalities as well as discrepancies between the marks. A



(46) THE VISION. 75—107.

There, thickest where the smoke ascends." As frogs
Before their foe the serpent, through the wave
Ply swiftly all, till at the ground each one
Lies on a heap; more than a thousand spirits
Destroy'd, so saw I fleeing before one
Who pass'd with unwet feet the Stygian sound.
He, from his face removing the gross air,
Oft his left hand forth stretch'd, and seem'd alone
By that annoyance wearied. I perceived
That he was sent from heaven; and to my guide
Turn'd me, who signal made, that I should stand
Quiet, and bend to him. Ah me! how full
Of noble anger seem'd he. To the gate
He came, and with his wand[1] touch'd it, whereat
Open without impediment it flew.
"Outcasts of heaven! O abject race, and scorn'd!"
Began he, on the horrid grunsel standing,
"Whence doth this wild excess of insolence
Lodge in you? wherefore kick you 'gainst that will
Ne'er frustrate of its end, and which so oft
Hath laid on you enforcement of your pangs?
What profits, at the fates to butt the horn?
Your Cerberus[2], if ye remember, hence
Bears still, peel'd of their hair, his throat and maw."
This said, he turn'd back o'er the filthy way,
And syllable to us spake none; but wore
The semblance of a man by other care
Beset, and keenly prest, than thought of him
Who in his presence stands. Then we our steps
Toward that territory moved, secure
After the hallow'd words. We, unopposed,
There enter'd; and, my mind eager to learn
What state a fortress like to that might hold,

[1] *With his wand.*]
She with her rod did softly smite the raile,
Which straight flew ope. *Spenser, F. Q.* b. iv. c. iii. st. 46.

[2] *Your Cerberus.*] Cerberus is feigned to have been dragged by Hercules, bound with a threefold chain, of which, says the angel, he still bears the marks. Lombardi blames the other interpreters for having supposed that the angel attributes this exploit to Hercules, a fabulous hero, rather than to our Saviour. It would seem as if the good father had forgotten that Cerberus is himself no less a creature of the imagination than the hero who encountered him.

(a)

282 TWELFTH NIGHT; OR, [ACT II.

Sir And. 'Slight, I could so beat the rogue:—
Sir To. Peace, I say.
Mal. To be Count Malvolio;—
Sir To. Ah, rogue!
Sir And. Pistol him, pistol him.
Sir To. Peace, peace!
Mal. There is example for't; the lady of the Strachy[1] married the yeoman of the wardrobe.
Sir And. Fie on him, Jezebel!
Fab. O, peace! now he's deeply in; look how imagination blows[2] him.
Mal. Having been three months married to her, sitting in my state,—
Sir To. O, for a stone bow, to hit him in the eye!
Mal. Calling my officers about me, in my branched velvet gown; having come from a day bed, where I left Olivia sleeping.
Sir To. Fire and brimstone!
Fab. O, peace, peace!
Mal. And then to have the humor of state: and after a demure travel of regard,—telling them I know my place, as I would they should do theirs—to ask for my kinsman Toby:—
Sir To. Bolts and shackles!
Fab. O, peace, peace, peace! now, now.
Mal. Seven of my people, with an obedient start, make out for him: I frown the while; and, perchance, wind up my watch, or play with my some rich jewel. Toby approaches; court'sies there to me:—
Sir To. Shall this fellow live?
Fab. Though our silence be drawn from us with cars,[3] yet peace.
Mal. I extend my hand to him thus, quenching my familiar smile with an austere regard of control:—

[1] Mr. R. P. Knight conjectures that this is a corruption of *Stratici*, a title anciently given to the Governors of Messina, and Illyria is not far from Messina. If so, it will mean the *Governor's lady*. The word *Strachy* is printed with a capital and in Italics in the first folio.

[2] Puffs him up.

[3] Thus in the Two Gentlemen of Verona, the clown says—"Who that is, *a team of horses* shall not *pluck from me*."

(b)

Figure 1.1: Example of pages from both book sources
a) Dante's The Vision, **b)** Shakespeare Anthology

clustering algorithm will be run and the resulting clusters will be compared. A semi-supervised learning algorithm will be run using knowledge gained from the clustering performed. Lastly we will look at grouping the marks from *The Vision* (Dante) based on the section of the book the mark was from and on Melville's Marginalia Online labeling the marks as attributed or unattributed to Melville. In Chapter 4 (Conclusion) we will then discus how this information could be used to determine possible marks that were created by an unknown author.

CHAPTER 2

METHODS AND TOOLS

In this section we will discuss and introduce the algorithm used in mark segmentation, how features were extracted from the marks, and finally the clustering method with clustering validation. The purpose of mark segmentation is to separate the mark from the page and machine printed text, specifically cataloging the pixels comprising the mark. By isolating the pixels that make up the marks, we were able to collect data on the mark composition. By isolating the marks we can extract mark features in isolation. Once the marks are isolated, features can be extracted, see Section 2.4. Making quantitative representative values, we can compare these values to each other in the hopes of differentiate the marks. Using that comparison, we can attempt to label marks as belonging to different groups (authors). This is done by using histograms, visual inspection, and clustering algorithms.

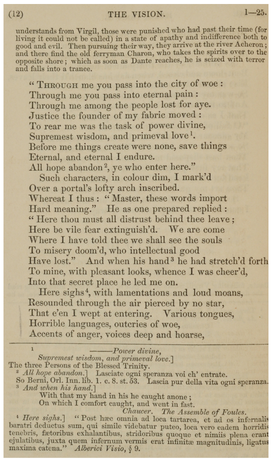
Marks in the books being analyzed were created using a pencil. The pencil marks were affected by how sharp the tip was, how long it had been used, how the graphite adhered to the page, the hardness of the graphite, and the manufacturers. The marks themselves were not uniform, creating complications and issues that have to be addressed and compensated for when dealing with mark features.

2.1 Image Data

We were given images from two books; *The Vision* (Dante) [9] which was assumed to contain marks from multiple authors, and a Shakespeare volume which was determined, by Dr. Olsen-Smith, to contain marks attributed to Melville. We were given 188 pages from *The Vision* (Dante) and 104 pages from the Shakespeare anthology book. In this section we will talk about how the pages were captured and stored.

The images of the pages from *The Vision* (Dante) were 2848 × 4288 pixels in size, and captured with a 12.2 Megapixel digital camera. The book pages had a physical size of 7 × 5 inches. These images were saved using the camera's built-in JPEG (Joint Photographic Experts Group) image format. The images captured from the books had been visually inspected by the Boise State University English department and sorted into pages that either contained pencil/pen objects and others that did not. The images of the pages from the Shakespeare anthology volume [10] were saved as a TIFF (Tagged Image File Format), at 2298 × 3825 pixels in size, and the book pages had a physical size of 9.75 × 6.5 inches. TIFF is a lossless image format that does not introduce compression artifacts.

Images can be taken and saved in a number of base formats including color, grayscale, or black and white. In a color image each pixel is stored as three 8-bit integer values in a three dimensional matrix; x, y, and color. These values are used in an additive color formula where they are added together in order to produce a single color. Saving it this way allowed us to have full control of the information, so that we could perform analysis based on the color contained in the image. Many image processing techniques utilize a grayscale or black and white image, therefore most algorithms discussed in



(12) THE VISION. 1—25.

understands from Virgil, those were punished who had past their time (for living it could not be called) in a state of apathy and indifference both to good and evil. Then pursuing their way, they arrive at the river Acheron; and there find the old ferryman Charon, who takes the spirits over to the opposite shore; which as soon as Dante reaches, he is seized with terror and falls into a trance.

"THROUGH me you pass into the city of woe:
Through me you pass into eternal pain:
Through me among the people lost for aye.
Justice the founder of my fabric moved:
To rear me was the task of power divine,
Supremest wisdom, and primeval love[1].
Before me things create were none, save things
Eternal, and eternal I endure.
All hope abandon[2], ye who enter here."
Such characters, in colour dim, I mark'd
Over a portal's lofty arch inscribed.
Whereat I thus: "Master, these words import
Hard meaning." He as one prepared replied:
"Here thou must all distrust behind thee leave;
Here be vile fear extinguish'd. We are come
Where I have told thee we shall see the souls
To misery doom'd, who intellectual good
Have lost." And when his hand[3] he had stretch'd forth
To mine, with pleasant looks, whence I was cheer'd,
Into that secret place he led me on.
Here sighs[4], with lamentations and loud moans,
Resounded through the air pierced by no star,
That e'en I wept at entering. Various tongues,
Horrible languages, outcries of woe,
Accents of anger, voices deep and hoarse,

[1] ———*Power divine,*
Supremest wisdom, and primeval love.]
The three Persons of the Blessed Trinity.

[2] *All hope abandon.*] Lasciate ogni speranza voi ch' entrate.
So Berni, Orl. Inn. lib. 1. c. 8. st. 53. Lascia pur della vita ogni speranza.

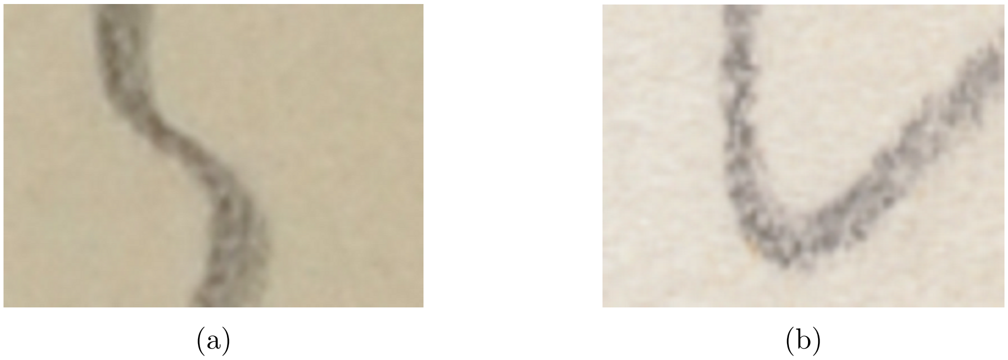
[3] *And when his hand.*]
With that my hand in his he caught anone;
On which I comfort caught, and went in fast.
Chaucer. The Assemble of Foules.

[4] *Here sighs.*] "Post hæc omnia ad loca tartarea, et ad os infernalis baratri deductus sum, qui simile videbatur puteo, loca vero eadem horridis tenebris, fætoribus exhalantibus, stridoribus quoque et nimiis plena erant ejulatibus, juxta quem infernum vermis erat infinitæ magnitudinis, ligatus maxima catena." *Alberici Visio*, § 9.

Figure 2.1: Example of area of useful information

this thesis utilize either a grayscale or a black and white image. A grayscale image is stored as a single matrix of 8-bit integers, each pixel is represented as an intensity value, having the extremes of the values representing black and white. A black and white image has its pixels stored as single bit values, with 1 and 0 representing black and white. Black and white images were required for some algorithms that were utilized in this thesis.

The specific JPEG parameters used by the camera were unknown. The JPEG standard (ISO/IEC 10918) is a lossy compression format. A lossy compression format loses information and introduces compression artifacts into the image during the compression process. The JPEG format uses a discrete cosine transform on 8 × 8 blocks; this will introduce blocking artifacts (blocking artifacts become more apparent at higher compression). The cosine transform acts like a smoothing function, softening edges by deleting high frequencies from the image, creating a general loss of sharpness. This is more apparent in images that contain a lot of high frequency information, like those found in text images. Where the image transitions from black to white, a "halo" or ringing artifact manifests as small "dots" around edges contained in the image. The pages of the book contained printed text and handmade marks, both of which contain high frequency content, namely the transition of each item to the background. The ringing artifacts created by the printed text can disrupt the true values of the marks, creating incorrect boundaries and pixel intensity values for marks. How the edge of the mark, or text, falls into that 8×8 grid will greatly affect the artifacts introduced; they were dependent on both the image and the parameters of the original compression and cannot be generally compensated for. While the marks had been distorted by the compression algorithm, the same algorithm was used on all unattributed marks. Figure 2.2 shows a zoomed-in view of marks from the unattributed marks and the attributed marks groups, highlighting the compression artifacts created by the JPEG compression format. Figure 2.2a shows a slight halo around the mark, a blurriness that was not present in the attributed mark in Figure 2.2b.



(a)

(b)

Figure 2.2: **a)** Unattributed mark zoom, **b)** Attributed mark zoom

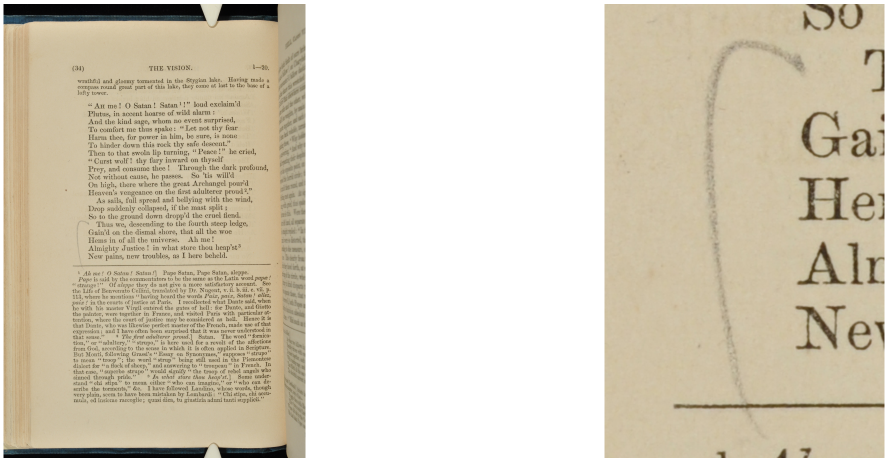
2.2 Mark Segmentation

The first step in analyzing the marks was to segment the marks from the rest of the page. The overall method used to extract the marks from the two books was the same, but some adjustments were necessary. In this section the base method used to extract the unattributed marks is described, then the modified algorithm for attributed marks will be described. The marks were located, isolated, and their pixel locations saved. The details of these steps are described next.

2.2.1 Unattributed Mark Segmentation

The first step was to separate the printed text from the background page. The pages were stored as an RGB color model using the JPEG format, see Figure 2.3a. Each page was converted to a grayscale image, see Figure 2.4a. This was done using the built-in *rgb2gray* function in MATLAB. This function uses a weighted average of the R, G and B intensity values to calculate the gray intensity value $I = 0.2989 * R + 0.5870 * G + 0.1140 * B$. Each page had pixels with intensity values ranging from 0-255. In the images, the printed text contained intensity values that were darker than the pixels that made up the pencil marks. A global threshold was used to get a close approximation of the darker center of each text character. By applying a global threshold of 80, every pixel that had a value greater (lighter) than 80 was set to a 0 (background paper) and every pixel that had a value less than (darker) or equal to 80 was set to 1 (foreground ink), see Figure 2.5a. A zoomed-in section of page that contained just text and background information was used as a starting value, then through experimentation we chose a threshold value of 80. The result was a compromise between setting mark pixels as the foreground (printed text) or background (white page). This method did not isolate all the pixels that represented the printed text. In order to acquire the rest of the printed text pixels further operations were needed.

In order to expand the printed text pixel locations, dilation was performed. Dilation is



(34) THE VISION. 1—20.

wrathful and gloomy tormented in the Stygian lake. Having made a compass round great part of this lake, they come at last to the base of a lofty tower.

"Ah me! O Satan! Satan[1]!" loud exclaim'd
Plutus, in accent hoarse of wild alarm:
And the kind sage, whom no event surprised,
To comfort me thus spake: "Let not thy fear
Harm thee, for power in him, be sure, is none
To hinder down this rock thy safe descent."
Then to that swoln lip turning, "Peace!" he cried,
"Curst wolf! thy fury inward on thyself
Prey, and consume thee! Through the dark profound,
Not without cause, he passes. So 'tis will'd
On high, there where the great Archangel pour'd
Heaven's vengeance on the first adulterer proud[2]."
As sails, full spread and bellying with the wind,
Drop suddenly collapsed, if the mast split;
So to the ground down dropp'd the cruel fiend.
Thus we, descending to the fourth steep ledge,
Gain'd on the dismal shore, that all the woe
Hems in of all the universe. Ah me!
Almighty Justice! in what store thou heap'st[3]
New pains, new troubles, as I here beheld.

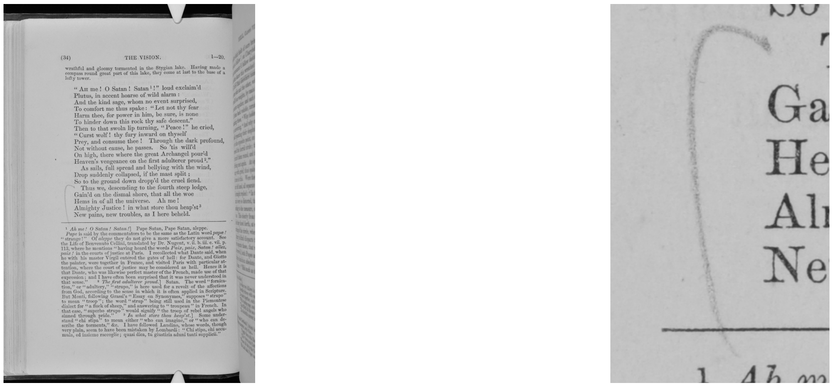
[1] *Ah me! O Satan! Satan!*] Pape Satan, Pape Satan, aleppe. *Pape* is said by the commentators to be the same as the Latin word *papæ!* "strange!" Of *aleppe* they do not give a more satisfactory account. See the Life of Benvenuto Cellini, translated by Dr. Nugent, v. ii. b. iii. c. vii. p. 113, where he mentions "having heard the words *Paix, paix, Satan! allez, paix!* in the courts of justice at Paris. I recollected what Dante said, when he with his master Virgil entered the gates of hell: for Dante, and Giotto the painter, were together in France, and visited Paris with particular attention, where the court of justice may be considered as hell. Hence it is that Dante, who was likewise perfect master of the French, made use of that expression; and I have often been surprised that it was never understood in that sense." [2] *The first adulterer proud.*] Satan. The word "fornication," or "adultery," "strupo," is here used for a revolt of the affections from God, according to the sense in which it is often applied in Scripture. But Monti, following Grassi's "Essay on Synonymes," supposes "strupo" to mean "troop"; the word "strup" being still used in the Piemontese dialect for "a flock of sheep," and answering to "troupeau" in French. In that case, "superbo strupo" would signify "the troop of rebel angels who sinned through pride." [3] *In what store thou heap'st.*] Some understand "chi stipa" to mean either "who can imagine," or "who can describe the torments," &c. I have followed Landino, whose words, though very plain, seem to have been mistaken by Lombardi: "Chi stipa, chi accumula, ed insieme raccoglie; quasi dica, tu giustizia aduni tanti supplicii."

(a)

(b)

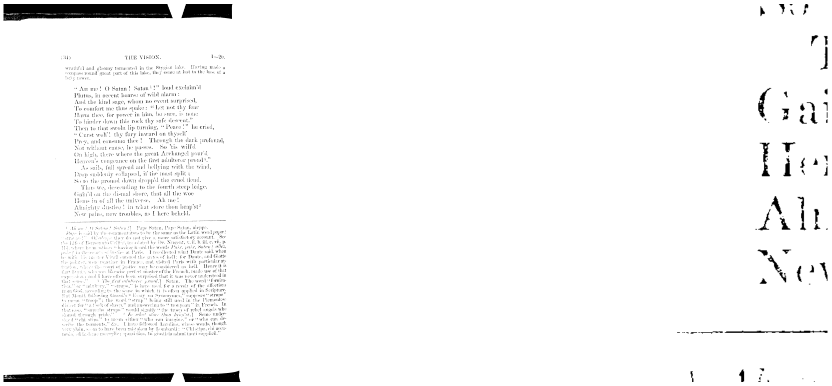
Figure 2.3: **a)** Original page image, **b)** Enlarged view of mark

the process of "growing" the boundary region of an object by the use of a structuring element; the original object's "growth" is determined by the size and shape of the structuring element. A large square 225 × 225 pixel structure element consisting of 1's was used to perform dilation (Figure 2.6a). The dilation operation connected neighboring pixels, thus connecting each printed text item to each of its neighbors, as well as connecting the background black bars and other noise in the images. Large connected components were created from the dilation. These objects consisted of the printed text in the middle of each page and scanning artifacts on the borders. An N_8 neighbor connected components algorithm was run in order to determine the size and location of each object. N_8 neighbor connected components looks at each pixel and labels the surrounded pixels as connected if they are one of the 8 neighboring pixels, this is repeated for each pixel location that contains a 1. A new label is created when a pixel is found that does not already have a label. Using prior knowledge of the structure of each page, objects near the borders were ignored, and a bounding box was set around the center box as the printed text location. A dilation operation was



(a) (b)

Figure 2.4: **a)** Image converted to gray scale (intensity), **b)** Enlarged view of mark

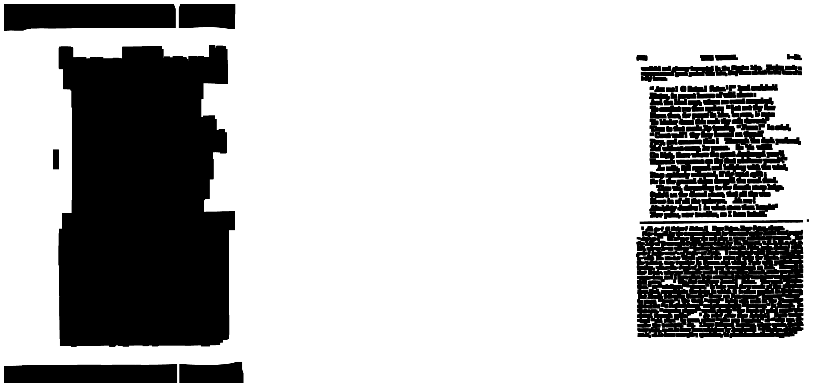


(a) (b)

Figure 2.5: **a)** Global threshold of > 80 was performed, **b)** Enlarged view of mark area

performed on the original black and white image (created by the global thresholding) to expand the printed text. A 25 × 25 structuring element consisting of 1's was used to expand the binary image to produce Figure 2.6b. This selected most of the printed text pixels while limiting the number of background pixels labeled as printed text pixels.

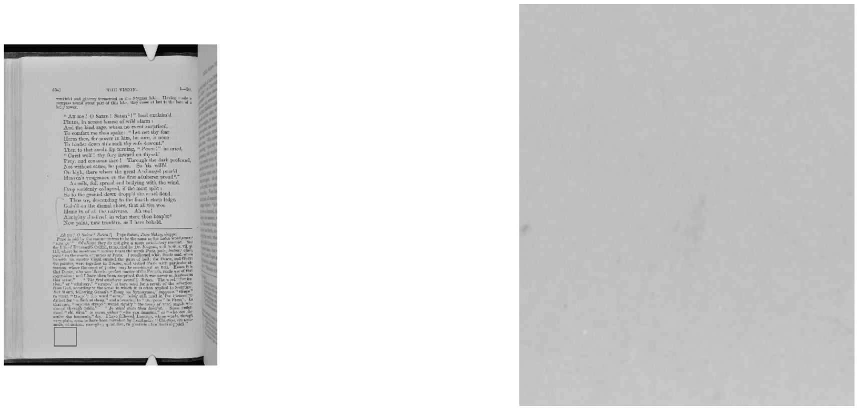
Up to this point we had focused on isolating the printed text. Now that the text was isolated we could make the printed text "disappear." The background color was



(a) (b)

Figure 2.6: **a)** Dilation: 255 × 255 structuring element, **b)** Dilation with 25 × 25 structuring element

sampled by taking an area of background pixels, see Figure 2.7. This area was used to generate a set of intensity values that were then randomly assigned to the printed text pixels. In this way the intensity values assigned to the printed text matched that of the background page, see Figure 2.8a.



(a) (b)

Figure 2.7: **a)** Full page, background reference area was within the red rectangle, **b)** Background reference area

With the printed text removed, the marks were left as the largest objects on the page, see Figure 2.8. To capture the mark pixel locations an adaptive thresholding algorithm was used. The adaptive thresholding algorithm that was used was developed at Tsinghua University [20]. Global thresholding uses a common threshold for all

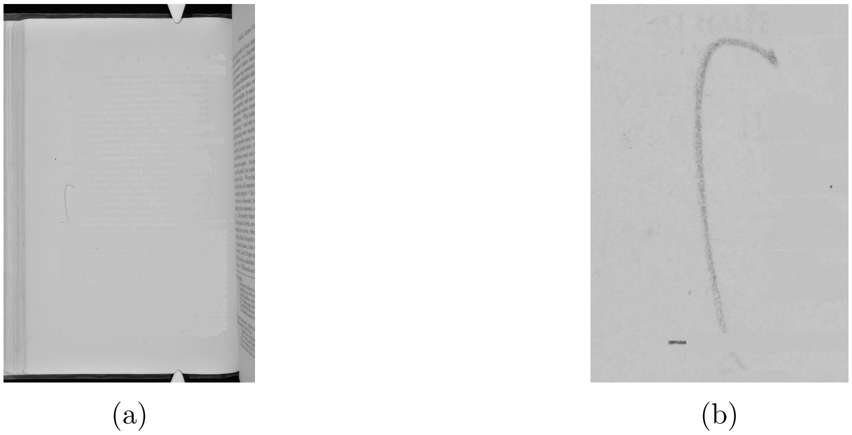
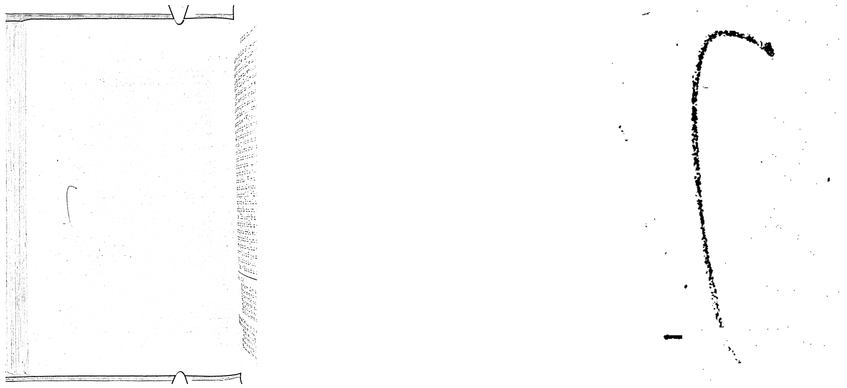

(a) (b)

Figure 2.8: **a)** Text pixels assigned the intensity value of the background, **b)** Enlarged view of mark

pixels, with adaptive thresholding the threshold changes dynamically over an image. The threshold for a pixel is determined by looking at the local neighborhood and calculating the mean, which is then used as the threshold and applied to the pixel. Adaptive thresholding was used to handle two problems: non-uniform yellowing of the page and show-through. The background of each page was not uniform; it contained non-uniform yellowing. This non-uniformity was still present after the image was converted to grayscale. The show-through was where the printed text from the back side of the sheet of paper was visible on the front side of the paper. All of the printed text on the next sheet of paper can add noise to the page being worked on. The result of this adaptive thresholding can be seen in Figure 2.9a.

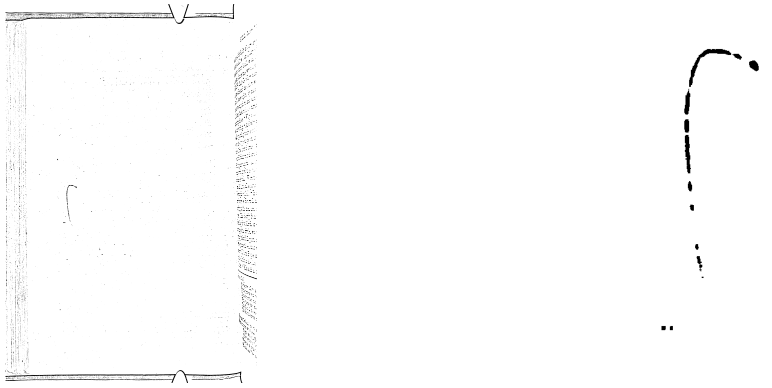
The resulting image after performing the adaptive threshold had salt and pepper noise, Figure 2.9b. To remove the salt and pepper noise, a 5×5 median filter was applied. The resulting image, Figure 2.10a, contained some noise, and a broken-up eroded mark.

In order to obtain as much information as possible about the mark, a 15×15 dilation



(a) (b)

Figure 2.9: **a)** Tsinghua adaptive thresholding applied, **b)** Enlarged view of mark



(a) (b)

Figure 2.10: **a)** After application of median filter: 5 × 5 structuring element, **b)** Enlarged view of mark

was performed, see Figure 2.11. This connected the stray pixels and parts of the mark, filling out the mark information. This process also widened the mark and caused it to contain multiple pixels that were background pixels. A 12 × 12 erosion was performed to shrink the boundary of the mark back to its original size, see Figure 2.12. Erosion leaves connected pixels, and shrank the mark to be thinner.

The final step was to perform an N_8 neighbor connected component algorithm. The largest objects were assumed to be marks; their pixels locations, mass (total number

(a) (b)

Figure 2.11: **a)** Dilation: 15 × 15 structuring element, **b)** Enlarged view of mark

(a) (b)

Figure 2.12: **a)** Erode: 12 × 12 structuring element, **b)** Enlarged view of mark

of pixels that make up the mark), and size (width and height) of the marks were stored and labeled. With the mark locations we could pull the mark information directly from the original unmodified image, see Figure 2.13. The mark extraction process is summarized in Figure 2.14.

2.2.2 Marks from the Attributed Book

We were given a Shakespeare anthology book that contained marks attributed to Melville. Many pages of this book showed signs of foxing, brown splotches and discoloration of the page, Figure 2.15 shows an example of this discoloration. The original

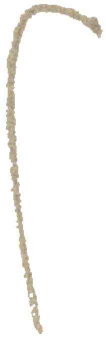


Figure 2.13: Mark isolated from page

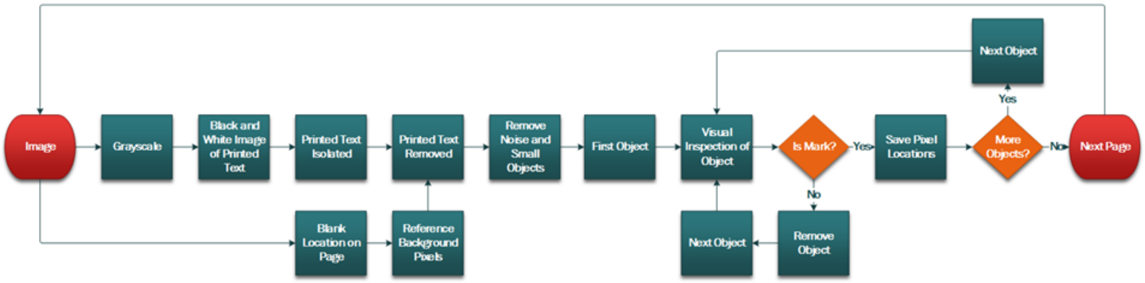


Figure 2.14: Flowchart showing steps for mark segmentation

mark extraction algorithm had to be modified in order to ignore these splotches. As the marks were not contained within these splotches, their features should not be affected.

The stains or splotches that appeared on the pages were unique in color, and did not match the rest of the page coloration. The splotches' RGB values were recorded, each image was then searched to find pixels that contained this color and any color in a small range around the selected RGB color spectrum. The found pixels were then changed to those of the background color. This technique was able to remove most splotches from the color image, but not all traces of them. Not all splotches were ignored by the algorithm, some were extracted as potential marks. These were later labeled as errors during manual visual sorting.

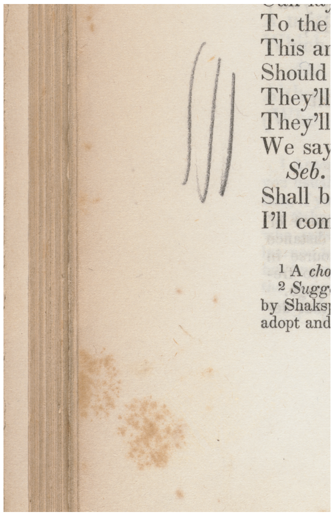


Figure 2.15: Example of Foxing on page

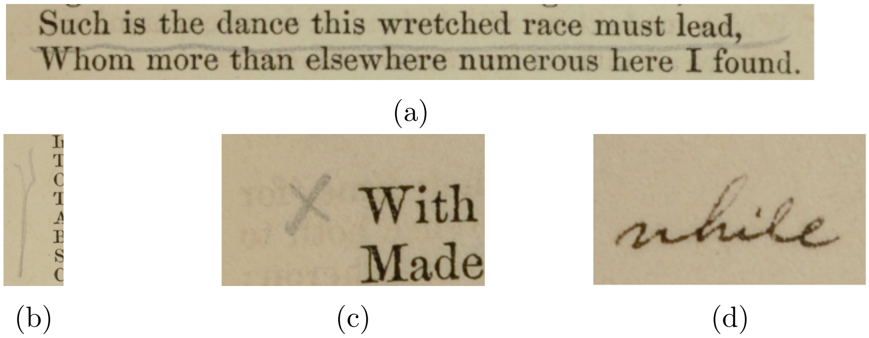
2.3 Categorizing and Mark Distributions

Using the algorithm described in Section 2.2, 566 marks from *The Vision* (Dante) and 115 from the Shakespeare anthology volume were segmented from the page and their pixel locations saved. All marks were then empirically categorized as belonging to one of six groups: vertical, horizontal, multi-stroke, symbols, handwritten object, and errors (Table 2.1). This was done manually by looking at each mark and assigning them to the correct category. An example of a vertical mark can be seen in Figure 2.15, while examples of all other types of marks can be seen in Figure 2.16.

Table 2.1: Unattributed and Attributed Mark Data

	Number of Marks		**Pages**	
Marks	**Unattributed**	**Attributed**	**Unattributed**	**Attributed**
Total	566	115	148	82
Vertical	261	63	123	48
Horizontal	93	7	57	6
Multi-stroke	79	16	54	14
Symbols	50	32	39	31
Hand-written Object	83	13	18	2

The majority of marks were vertical marks. Vertical marks in this context consist of marks that were placed to the left or right of the printed text and that travel generally vertical in reference to the printed text. Vertical marks were the "cleanest" marks found, they were typically unbroken, contained a single stroke, and were done in a freehand style.


Such is the dance this wretched race must lead,
Whom more than elsewhere numerous here I found.
With
Made
while

(a)

(b) (c) (d)

Figure 2.16: Example of types of marks
a) Horizontal **b)** Multi-stroke, **c)** Symbol, **d)** Hand writing

The second largest category of marks were horizontal marks. An example of this can be seen in Figure 2.16a. Horizontal marks were typically marks that were used to underline a section of printed text. Horizontal marks often passed through the printed text, leaving the mark broken and disjointed after the extraction process, and artificially increasing the count of the number of marks in this category. Horizontal mark segmentation was inconsistent, and would require interpolation in order to connect each component.

When a mark consisted of two or more strokes, or when two single stroke marks intersect, like in Figure 2.16b, the object was grouped as a multi-stroke mark. The issue with multi-stoke marks was how to separate them. Consider the multi-stroke mark in Figure 2.16b. When does one stroke end and the other start? How would we split this up in order to be comparable to a single stroke mark? The feature extraction algorithms that we created made a number of assumptions that multi-stroke marks break. As an example, the algorithms expect a mark to have two ends, while the multi-stroke mark in Figure 2.16b has three. Multi-stroke marks will not be looked at in this thesis.

Symbols are marks that were created to form a specific shape, such as an "x" or a Roman numeral. An example of a symbol is shown in Figure 2.16c. Symbols are a different category than handwriting as the intent was not a character but a shape. Check marks could be difficult to distinguish from vertical marks, as it was common to add a tail to the end of a line.

We were labeling handwriting marks that consist of multiple strokes put together to form standard characters (alphabetic) and combination of characters, as seen in Figure 2.16d. The authorship of the handwriting objects contained in the two books were not in dispute and as such any and all handwriting annotations were ignored.

In this thesis only vertical marks were considered. The largest group of marks was vertical marks and vertical marks do not require interpolation or other algorithms in order to maintain the integrity of the marks. As such it was believed that analyzing just the vertical marks would reveal information to determine if mark authorship can be obtained by the use of image processing techniques.

2.4 Feature Extraction

Each mark was created by a single stroke of a writing instrument. The features extracted were affected by the type of writing utensil that was used, its condition, the writing location, how the pencil was sharpened, the brand of pencil, the angle at which the pencil was used, whether the surface of the paper was different, if the marks were created while traveling, and even the pressure that was applied when the mark was created. All of these things could be indicators of different authors, but could also just be from the same author writing during different conditions and times. Features that were used for analysis were: average intensity, stroke width, blurriness, stroke curvature, and stroke angle. These features are described in the following sections.

2.4.1 Average Intensity

The average intensity feature was the average intensity of every pixel that had been designated as belonging to the mark, per Section 2.2 Mark Segmentation. To calculate the average mark intensity, the mark was converted from an RGB color image to a grayscale image that had pixel values ranging from 0 to 255. The marks that were collected had intensity ranges from 75 to 200. Figure 2.17 shows examples of marks with their average intensity values.

2.4.2 Stroke Width

Stroke width was the measurement of how many pixels wide a mark was. Each mark was converted to a binary image (black and white) by setting the background to 0 and the pixels of the mark to 1. This binary image was then skeletonized with the use of a morphological thinning process by means of erosion. Morphological thinning was done by taking the structuring element in Figure 2.18a and overlapping it at every possible position within the image. The structuring element was then rotated by forty-five degrees and the process was repeated until the structuring element was back to the starting position. This was counted as a single iteration. The number of iterations required before there was no change in the skeleton was referred to as the

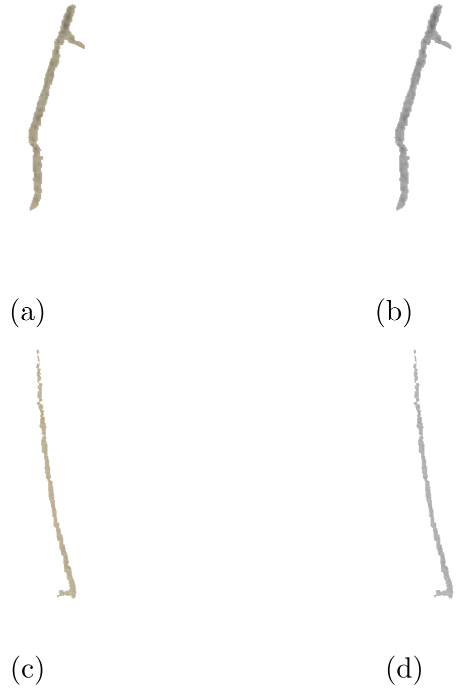


Figure 2.17: Example of mark intensity range.
a) Unattributed mark in RGB, **b)** Grayscale average intensity:145.5, **c)** Attributed Mark in RGB, **d)** Grayscale average intensity:179.5

stroke width, see Figure 2.19. Because stroke width measures the width of a mark by counting the layers of pixels, the resolution of the original image and size of the book effect the results. The stroke width value of the unattributed and the attributed marks cannot be directly compared, instead the distribution of the values could be compared.

2.4.3 Blurriness

The Blurriness feature looks at the uniformity of a mark's intensity. It was derived from the *ISO/IEC 13660:1997 Measurement of Image Quality Attributes for Hard Copy Output Standard Blurriness Attribute* [1]. The ISO 13660 standard was designed to evaluate the quality of printed objects. From the ISO 13660 standard, blurriness

-1	-1	-1
0	1	0
1	1	1

(a)

0	-1	-1
1	1	-1
1	1	0

(b)

1	0	-1
1	1	-1
1	0	-1

(c)

Figure 2.18: Thinning Structuring Element.
a) Starting, **b)** Rotated 45^o, **c)** Rotated 90^o

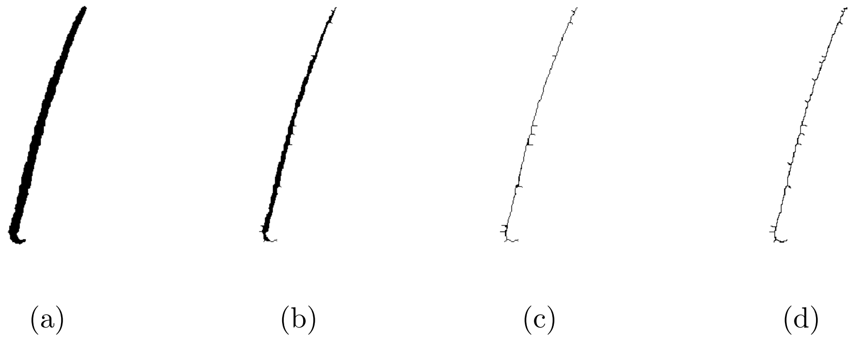


Figure 2.19: Example of mark being skeletonized.
a) Initial mark, **b)** 3 iterations, **c)** 6 iterations, **d)** stroke width of 9

was the measure of the average distance between the inner and outer boundary edges. Printed text was expected to be a solid uniform color, but there exists an edge transition between this solid color and the background page, where the ink or toner do not fully cover the paper's original color. Printed text quality is higher when the transition region from the background to the printed text intensity value is smaller. Unlike printed text, a pencil mark does not have a consistent intensity center or value. Often the maximum intensity value will not show up in every row, nor will it be consistent in its location within a row. This means that there was not a region to measure to create the standard blurriness feature. The blurriness feature used in this work looks at the percentage of the mark contained within the region

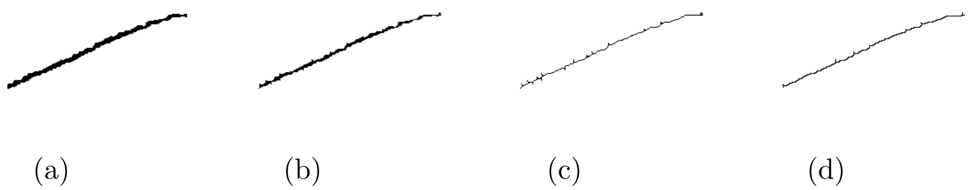


Figure 2.20: Example of mark being skeletonized,
a) mark, **b)** 1 iteration, **c)** 3 iterations, **d)** 4 iterations

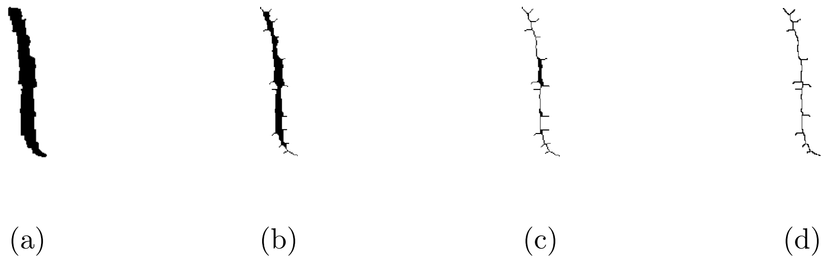
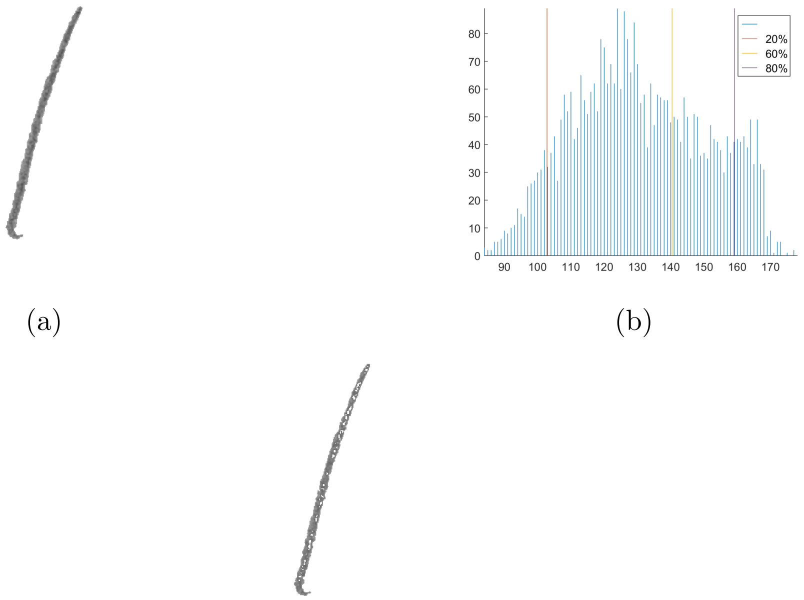


Figure 2.21: Example of mark being skeletonized.
a) mark, **b)** 4 iterations, **c)** 7 iterations, **d)** 11 iterations

of eighty to twenty percent of maximum intensity; these levels were taken from the ISO 13660 standard. The larger the value, the more gradual the line transition. The eighty percent of maximum intensity was calculated for the full mark and then the twenty percent of maximum was calculated. All pixels that fell within the 20 percent to 80 percent range were counted. This was then normalized by the size of the mark, producing a percentage that represents how many of the pixels contained by the mark were within that range, see Figure 2.22. Figure 2.22c shows what the mark looks like when the pixels outside the 20-80 range were removed. As can be seen, the maximum intensity was not consistent or present in every row.

2.4.4 Stroke Curvature

Stroke Curvature looks at the straight line path a mark takes to go from one end of the mark to the other end and the area that is created by connecting those points.



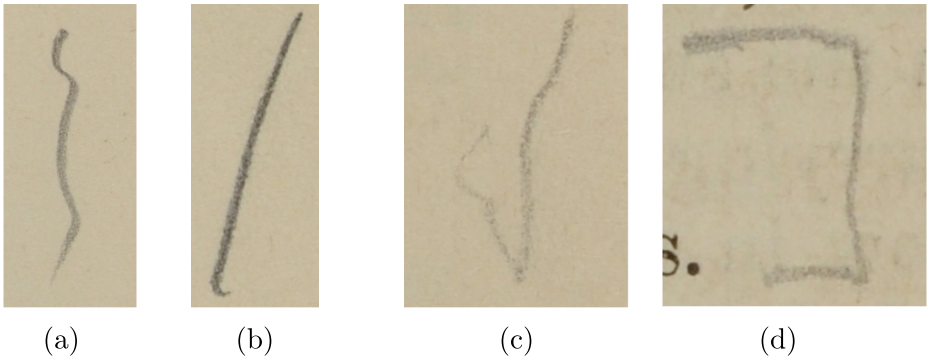
(a)

(b)

(c)

Figure 2.22: Blurriness reference
a) Mark, **b)** Histogram of mark intensities, with 20, 60, and 80 percent shown, **c)** Mark by the pixels outside 20 and 80 percent removed

There were a number of types of paths the strokes took: there were the snake-like shape as seen in Figure 2.23a, a small tick at the end of a relatively straight line as seen in Figure 2.23b, a large check mark as shown in Figure 2.23c, and a bracket as seen in Figure 2.23d. The curvature of a mark was determined by the position of the beginning of the mark and the path traveled to the end of the mark. To calculate this feature, a mark was first skeletonized, then each end point was identified. The longest path was identified and this was assumed to be the mark's "true" path, all other branching points were removed. A straight line was created by connecting the start and end of the longest line. The area that was contained between the longest line and the connecting line was calculated. This area was then normalized (divided) by the length of the connecting line. Figure 2.24 and Figure 2.25 show examples of how the stroke curvature feature was processed and calculated.



(a) (b) (c) (d)

Figure 2.23: Example of types of paths a stroke can take
a) Snake-like, **b)** Small tick, **c)** Check, **d)** Bracket

2.4.5 Stroke Angle

Stroke Angle is the measurement of the deflection of the mark past true vertical. A mark could be vertical, horizontal, or somewhere in-between. In order to calculate the angle of a mark, the Hotelling Transform (Karhunen-Loève theorem) [5] $y = A(x - m_x)$ was used, with $x = [x_1, x_2, ...x_n]^T$ representing all the mark pixel coordinates. The mean vector m_x was

$$m_x = E\{x\} = \frac{1}{K}\sum_{k=1}^{K} x_k \tag{2.1}$$

with K being the total number of vectors (points). C_x was defined as

$$C_x = E\{(x - m_x)(x - m_x)^T\} = \frac{1}{K}\sum_{k=1}^{K} x_k x_k^T - m_x m_x^T \tag{2.2}$$

and was a real and symmetric matrix. The Hotelling Transform returned the major and minor eigenvalues and eigenvectors of the pixel seeds, which were the major and minor axis of an edge on the mark, see Figure 2.26. A was determined by the covariance matrix C_x of the pixel coordinates. The rows of A were the eigenvectors

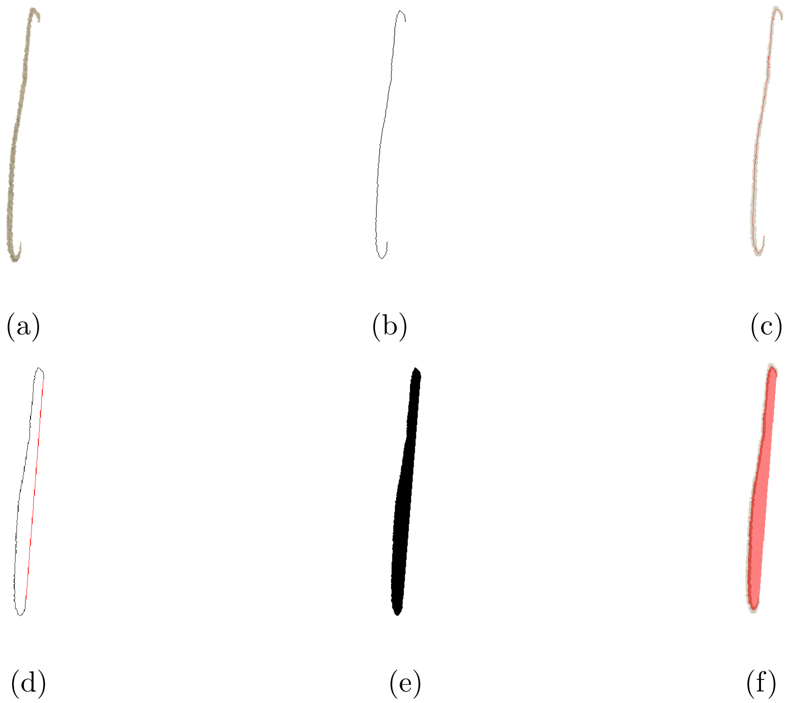


Figure 2.24: Example of curve feature measurement, stroke curvature=24.6,
a) Original mark, **b)** Longest Line, **c)** Longest line superimposed on original mark, **d)** Connected **e)** Filled area, **f)** Filled area superimposed on original mark

e of C_x in descending order from their corresponding eigenvalues. The deflection off the major axes was used as the feature value.

We described the algorithm for extracting pencil marks from an image out of a page from a book. This allowed us to quickly extract and process large numbers of marks automatically. The basis of this algorithm was used for both books, with only small modifications needed due to extra degradation of the pages contained in the Shakespeare volume, with some parameters adjusted for each book. Once the marks had been segmented from the page, the same feature extraction algorithm was used on marks from both books, no modification was necessary. The features were tailored for pencil marks, and could be used for any pencil marks of this type. These features could be used to create a profile of the marks contained in the books.

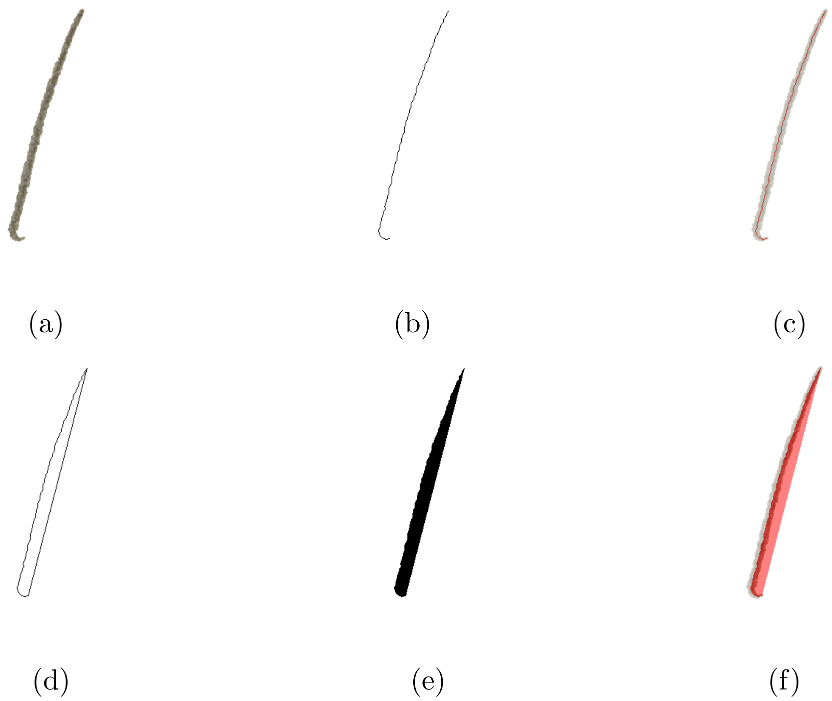
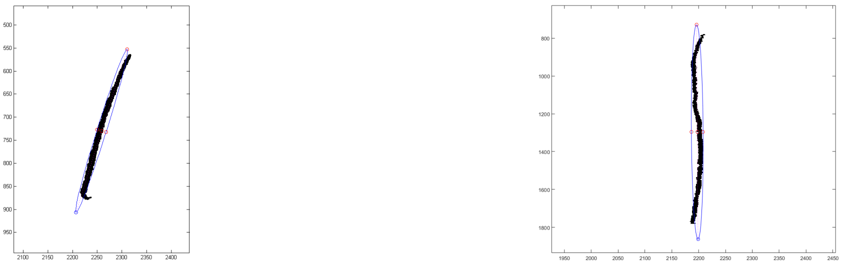

(a) (b) (c) (d) (e) (f)

Figure 2.25: Example of curve feature, stroke curvature= 15.4078,
a) Original mark, **b)** Longest Line, **c)** Longest line superimposed on original mark, **d)** Connected, **e)** Filled area, **f)** Filled area superimposed on original mark

2.5 Clustering

We have assumed that marks in *The Vision* (Dante) are unlabeled, attributed labeling by Melville's Marginalia Online will be discussed in section 3.6 (Melville's Marginalia Online Attributed Labels). Therefore an unsupervised learning algorithm, called Clustering, was used to divide them into groups of similar content. These groups could represent different authors. Two techniques of clustering were utilized: visual clustering, and C-means clustering.

With unsupervised classification, C-means clustering was chosen as the clustering method. C-mean clustering is a tool often used to analyze trends in features. In order to label and categorize the marks it was necessary to look at all features, as



(a)

(b)

Figure 2.26: Hotelling example used for stroke angle feature. Stroke Angle: **a)** 0.285, **b)** 0.0025

well as subsets of features. Each mark had a feature vector created that contains all the features collected for that mark. Each mark was classified without an initial label, which was done by using unsupervised classification.

2.5.1 Visual Clustering

With only a few features, two or less, it was possible to visually see how features were clustered with the use of histograms and scatter plots. This was done by visually analyzing histograms created for each feature, and the creation of two feature scatter plots with a feature on each axis.

2.5.2 C-means Clustering

C-means clustering [5] [7], also attributed as K-means or ISODATA clustering, was chosen because it is a non-hierarchical clustering algorithm; it allowed for augmentation of the algorithm and it is algorithmically simple. C-means is a clustering algorithm that can be used when the prior knowledge is limited; in this case, all parameters were unknown.

The C-means clustering algorithm starts by defining the C (centroids) vectors of each of the clusters. Membership of each cluster is determined by setting each feature to be a member of the cluster which has a center to which it is closest. A new cluster center is calculated based on the newly created sets. Using these new centers, each feature has its distances recalculated and is assigned to a new cluster based on the new smallest distance; this process is repeated until the cluster-set membership does not change.

For the algorithm of C-means clustering we defined the number of clusters as C. A data point was randomly selected from the dataset X; this was set as the first centroid c_1. The distance from each data point m to c_1 was calculated and labeled as d_{m_x,c_k}. The second centroid c_2 was selected at random from X with probability

$$c_2 = \frac{d^2_{(m_x,c_1)}}{\sum\limits_{k=1}^{n} d^2_{(x_k,c_1)}}. \tag{2.3}$$

The distance from each data point to each centroid was calculated with each assigned to the closest centroid. For $m = 1, ..., n \quad \text{and} \quad p = 1, ...k - 1$ centroid k was selected at random from X with probability

$$P = \frac{d^2_{(m_m,c_p)}}{\sum\limits_{h;x_h \in c_p} d^2_{(x_h c_p)}} \tag{2.4}$$

where C_p was the set of all data points closest to the centroid c_p and x_m belongs to C_p. This process was repeated for k centroids.

In order to utilize this algorithm, we had a number of choices in terms of parameters. The initial cluster center locations needed to be defined. Random points in the feature space could be used, but this ran the risk of centers being defined in locations that lead to empty clusters from all features having distances that were closer to other centers and were never assigned to one of the centers. A simple approach to assigning cluster centers was to take a random sample of the data-set. This has its own complications, such as having selected cluster centers too close to each other in the feature space. The type of distance that was used needed to be chosen (euclidean, cityblock, cosine), in this case squared euclidean was used.

2.5.3 Cluster Validity

After the clusters were created by C-means clustering, the validity of the clusters had to be determined. Two clustering validity methods were used, Davies-Bouldin Index and Silhouette. They will be described next.

2.5.3.1 Davies-Bouldin Index

The Davies-Bouldin Index [4] was based on a ratio of the distance between clusters and the inner distance between cluster members, defined as

$$DB = \frac{1}{k}\sum_{i=1}^{k} max_{j\neq i}\{D_{i,j}\}$$

and

$$D_{i,j} = \frac{(d_i + d_j)}{d_{i,j}}.$$

$D_{i,j}$ is the intercluster distance ratio between the i and j clusters, with d_i and d_j representing the average distance of the points contained within the respective clusters. This can be the average distance between points or it can be the maximum distance between any two cluster points. For the Davies-Bouldin Index, the smaller the returned value, the higher the confidence that it was the right number of clusters.

2.5.3.2 Silhouette

Silhouette [17] clustering validation was the measure of inter-cluster distances between each cluster member, indicating each cluster's own cohesion, and was defined

$$S_i = \frac{(b_i - a_i)}{max(a_i, b_i)}$$

with a_i as the average distance from the i point to other points in the same cluster, and with b_i as the minimum average distance from the i point to points in different clusters. The silhouette values ranged from -1 to +1, with +1 indicating a high confidence that i was well-matched to its own cluster.

Using an unsupervised learning algorithm like C-means allowed us to analyze the feature spaces that were too complex for visual inspection. We did not need to have prior knowledge of the marks before running the algorithm and placing the marks into different clusters. By performing validation techniques on these clusters, we could determine the confidence level of the number of clusters.

We were given two groups of marks: unattributed and attributed marks. We talked about the types of image formats we were given, and the challenges associated with those images. We disused on how marks were segmented from the pages, and the five features that were chosen and how they were collected. We also talked about some of the tools used to analyze the features. In the next section we talk about the experiments that we performed on the features collected, and the results of that analysis.

CHAPTER 3

EXPERIMENTS

Five features were collected from each mark. In this chapter we will discuss the experiments that were run and the results gained from these features. Five methods of analysis were used: individual feature distributions, comparing unattributed and attributed marks features, 2D features space comparing pairs of features from attributed and unattributed marks, C-means clustering, and semi-supervised learning.

3.1 Individual Feature Distribution Analysis

The first approach in analyzing the features was to look at the histograms of the five features extracted. We were looking for a bimodal or multimodal distribution, as the different distributions around means could be an indicator of different groupings of marks. If a multimodal distribution was found, the marks would be labeled into different groups. With these labels applied we would look at the histograms of the other features, with a hope of seeing patterns. The histograms for all features individually can be seen in Figure 3.1.

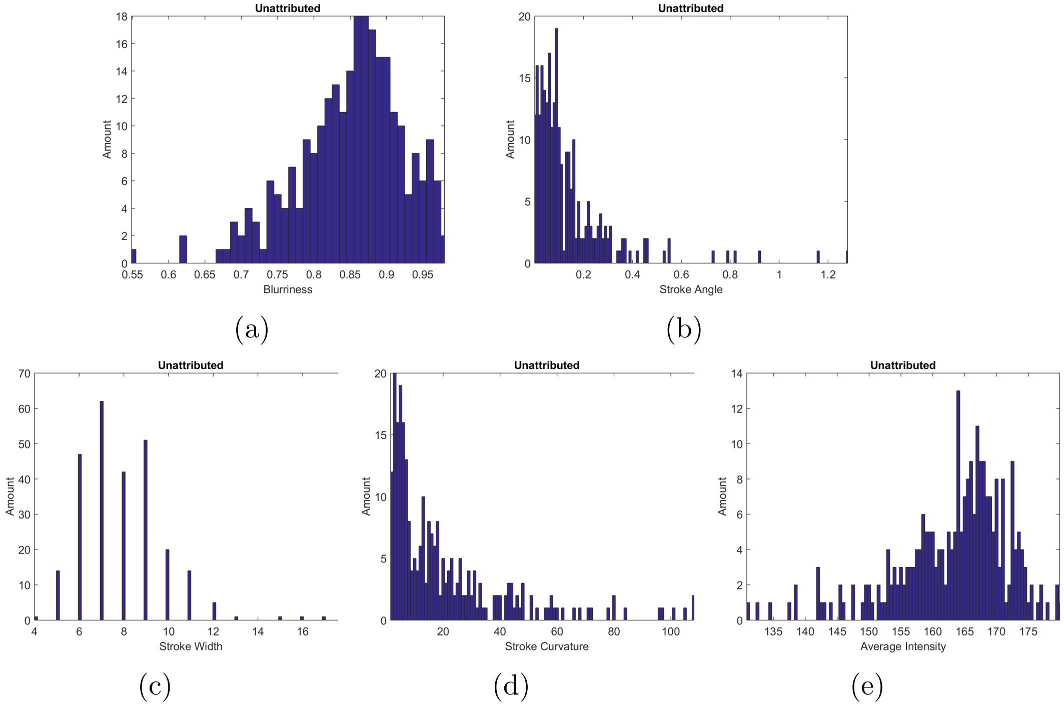


Figure 3.1: Histograms of features
a) Blurriness, **b)** Stroke Angle, **c)** Stroke Width, **d)** Stroke Curvature, **e)** Average Intensity

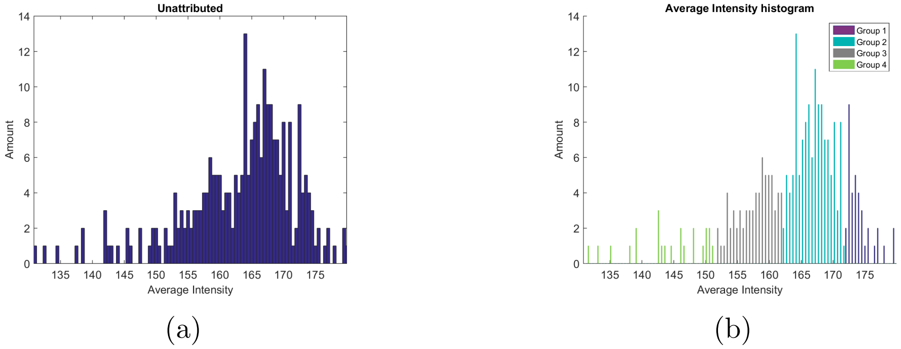
3.1.1 Possible Multimodal Distributions

To look for a multimodal distribution we were looking for patterns in the histogram of the features, multiple peaks, or groups of marks that were separate from each other. We found two features that contained a possible multimodal distribution: Average Intensity and Stroke Curvature. Marks were split into four groups based on the Average Intensity histogram, and into two groups based on the Stroke Curvature histogram. They will be discussed next.

3.1.2 Average Intensity

The average intensity was calculated by taking the grayscale image of a mark and finding the average intensity of the pixels that make up the mark, as discussed in

Section 2.4.1 Average Intensity. The Average Intensity histogram can be seen in Figure 3.2a. This histogram appears to contain a multimodal distribution, with four distinct peaks that could be segmented into defined groups. This split can be seen in Figure 3.2b, where each group is displayed with a different color. The exact values used for each group is shown in Table 3.1.



(a)

(b)

Figure 3.2: Average Intensity Histograms
a) Average Intensity, **b)** Labeled into four groups

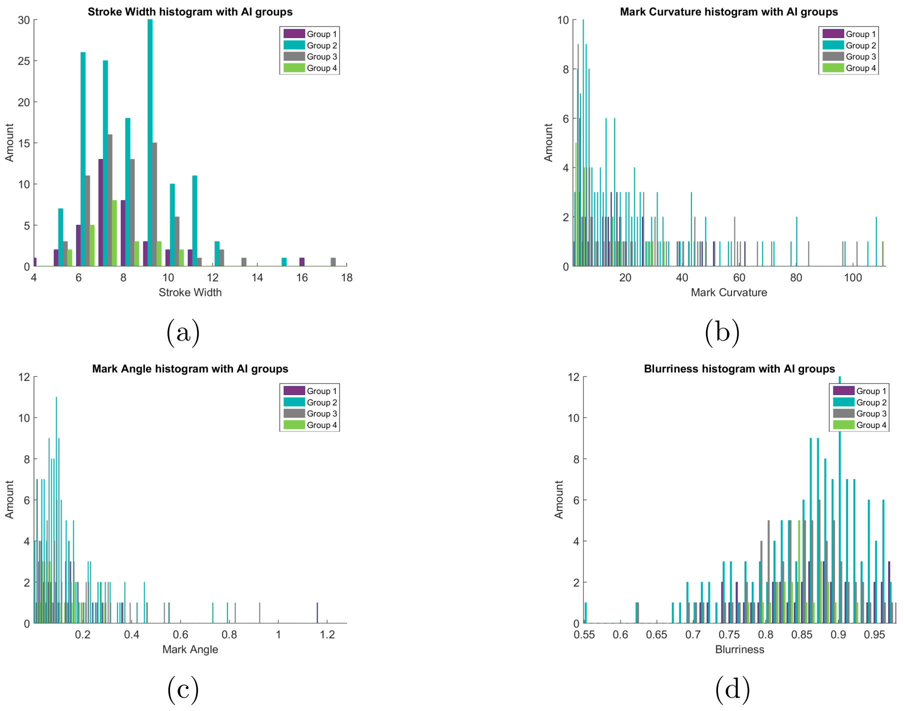
Table 3.1: Average Intensity Group Ranges

Group	Minimum Intensity (≥)	Maximum Intensity (<)
1	171.77	181
2	161.75	171.77
3	151	161.75
4	1	151

3.1.2.1 Average Intensity Applied to Other Features

The four group labels were applied to the histograms of the other four features. Stroke Width, Stroke Curvature, and Stroke Angle did not show any patterns, as seen in Figures 3.3a, 3.3b, and 3.3c.

The Blurriness feature showed some patterns when this grouping was applied, see Figure 3.3d. It can be seen that Group 4 and 3 have a Gaussian-like shape that was offset from Group 2's Gaussian shape. Group 1 has a uniform shape throughout the blurriness range.



(a)

(b)

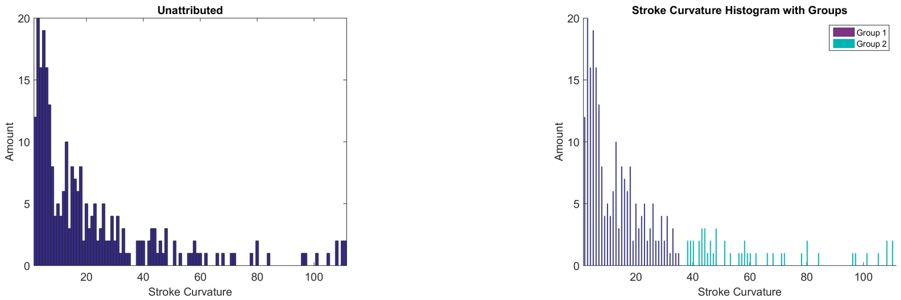
(c)

(d)

Figure 3.3: Histograms of features a,b,c,d. Each was color-coded with the group from average intensity in Figure 3.2b
a) Stroke Width, **b)** Stroke Curvature, **c)** Stroke Angle, **d)** Blurriness

3.1.3 Stroke Curvature

The Stroke Curvature feature looks at how much a mark deviates from a straight line. The more curvature a mark has, the larger the value. Figure 3.4a shows the histogram of Stroke Curvature. The histogram appears to have a possible bimodal distribution, with one group going from a value of 0 to 37, and the second group starting at 37 to a max value of 112, see Table 3.2. In order to explore this possible bimodal distribution, the curvature histogram in Figure 3.4a was split into two group labels, see Figure 3.4b.



(a)

(b)

Figure 3.4: Stroke Curvature Histograms
a) Stroke Curvature, **b)** Labeled into two groups

Table 3.2: Curvature Group ranges

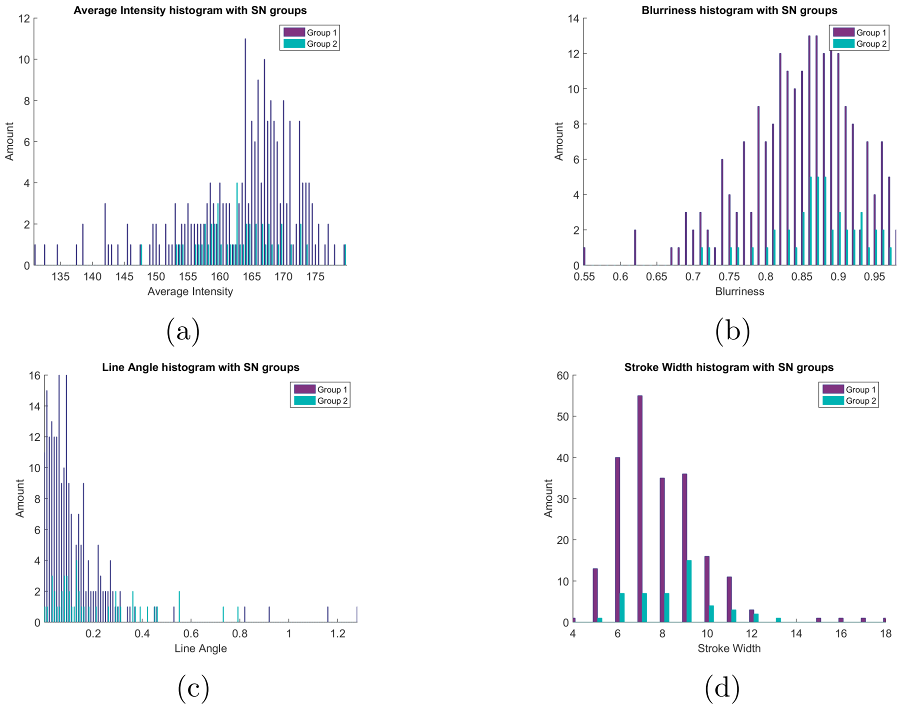
Group	Minimum Intensity (≥)	Maximum Intensity (<)
1	0	37
2	37	112

Figure 3.5a shows Average Intensity with Stroke Curvature group labels. Group 1 has the same basic pattern as seen without the grouping, but Group 2 was more normalized with some small peaks around the 163 range.

The histogram of Blurriness and Stroke Angle, Figures 3.5b and 3.5c, do not show any difference in patterns between the two groups. Both subgroups showed a similar pattern to each other and what was seen in the original histogram.

Figure 3.5d shows the histogram of Stroke Width. Both groups had a Gaussian distribution, but they were centered at two different points. Group 1 was centered at 7, while Group 2 was centered at 9. Group 1 had the largest values, but also had the smallest and more marks with a value of 5. Because they were spread out over a large range, we were unable to group them directly.

In summary, two of the features that we collected show possible bimodal or multi-



(a) (b) (c) (d)

Figure 3.5: Histograms of features a,b,c,d. Each was color-coded with the group from Stroke Curvature in Figure 3.4b
a) Average Intensity, **b)** Blurriness, **c)** Stroke Angle, **d)** Stroke Width

modal distributions. The average intensity feature had a possible multimodal distribution with four groups. These group labels were applied to the other features to look for patterns. Blurriness showed that Group 4 and 3 have a Gaussian-like shape that is offset from each other, see Figure 3.3d. The other feature that had a bimodal distribution was stroke curvature. Marks were split into two group labels based on the stroke curvature. When average intensity was split into two groups, it showed different styles of distribution, one group has a Gaussian distribution while the other has a uniform distribution. The stroke width groups both showed Gaussian distribution but had different peaks. Stroke curvature shows a possibility of determining the difference between different authors.

3.2 Comparing Histograms of Attributed and Unattributed Marks

We collected features from both unattributed and attributed marks and then histograms were created from the five features. In this section we will compare the histograms from the unattributed and attributed mark groups. We are looking for similarities and discreteness between the two sets of histograms. Table 3.3 shows the minimum, maximum, and 90th percentile, of feature values obtained from the attributed and unattributed marks. The largest difference between these two groups was the number of data points available.

Table 3.3: Unattributed and attributed mark feature info

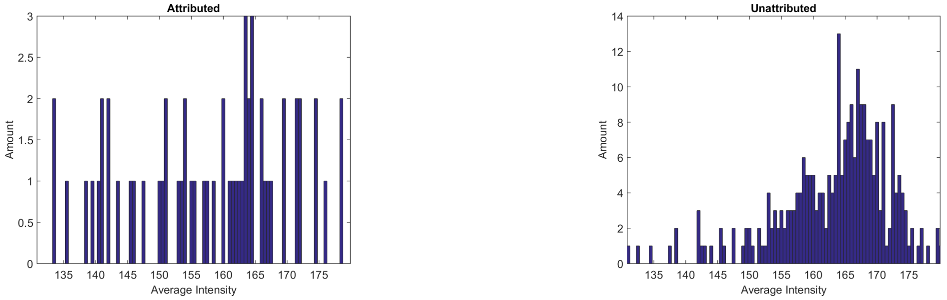
	90%		**Minimum**		**Maximum**	
Marks	**Attrib**	**Unattrib**	**Attrib**	**Unattrib**	**Attrib**	**Unattrib**
Avg Intensity	178.5	172.8	133.5	130.8	185.7	179.9
Stroke Width	8	10	4	4	11	18
Blurriness	0.882	0.7503	0.822	0.549	0.979	0.980
Stroke Curv.	23.7	49.4	1.62	1.55	52.9	111.5
Stroke Angle	0.146	0.295	0.002	.00009	0.571	1.28

3.2.1 Average Intensity

The Average Intensity histograms are compared in Figure 3.6, both histograms of this feature have similar patterns and spread.

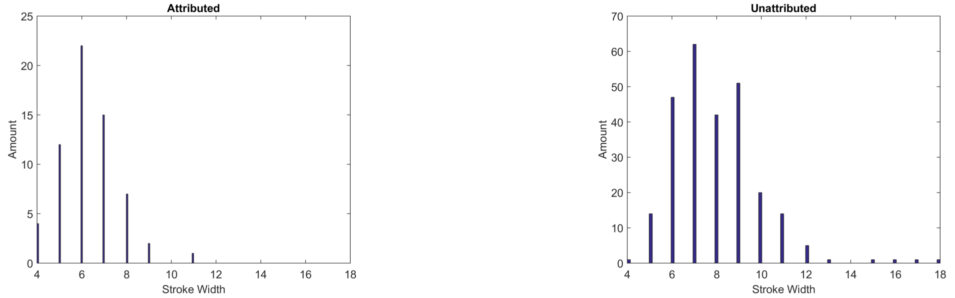
3.2.2 Stroke Width

Figure 3.7 compares the histograms of stroke width from attributed and unattributed marks. Unattributed marks had a larger distribution compared to attributed marks.



(a) (b)

Figure 3.6: Histogram of intensity of marks from attributed and unattributed marks, **a)** Attributed marks, **b)** Unattributed marks

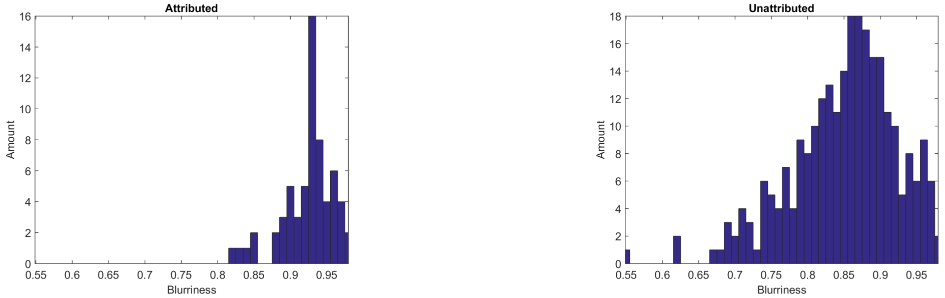


(a) (b)

Figure 3.7: Histogram of stroke width of attributed and unattributed marks, **a)** attributed mark, **b)** unattributed marks

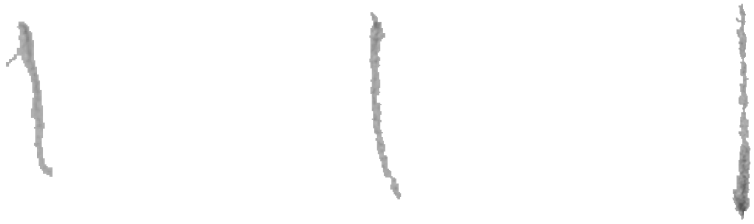
3.2.3 Blurriness

The histograms of Blurriness, Figure 3.8, showed some interesting patterns. Attributed marks had a maximum peak value of 0.95, while unattributed marks was 0.86 and 0.87. The lowest value of the attributed marks was 0.825, while unattributed marks had a low of 0.55, see Table 3.3. Approximately 31.8% of unattributed marks had a blurriness value less than 0.825, and 64% of unattributed marks were below the 90% value of the attributed marks. Figure 3.9 shows the unattributed marks that have the smallest value of blurriness. All three marks had dark regions surrounded by very light regions, with multiple pixels in-between the two extremes. This gave an idea of what the differences were between authors and what to look for in terms of other marks and features that make these stand out when compared to each other.



(a) (b)

Figure 3.8: Histogram of Blurriness of attributed and unattributed marks,
a) Attributed mark, **b)** Unattributed marks

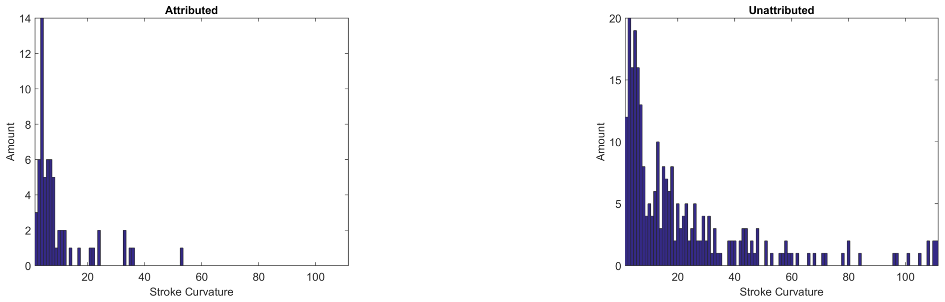


(a) (b) (c)

Figure 3.9: Unattributed marks with lowest Blurriness value,
a) Blurriness: 0.549, **b)** Blurriness: 0.618, **a)** Blurriness: 0.624

3.2.4 Stroke Curvature

The comparison of Stroke Curvature can be seen in Figure 3.10. The maximum value of attributed marks was 52.9, the unattributed marks had a maximum value of 111.5, see Table 3.3. Attributed marks had 74.6% of its marks with values less than 10, unattributed marks had only 42.5% of its marks below 10. There were 29.5% of unattributed marks that were above the 90% value of 23.7 of the attributed mark group. Three examples of these marks can be seen in Figure 3.11a. Unattributed marks had 6.9% of marks that are above the maximum value found in attributed marks. This could be an indicator that a high Stroke Curvature, greater than 20, could be a characteristic of an author other than Melville.



(a) (b)

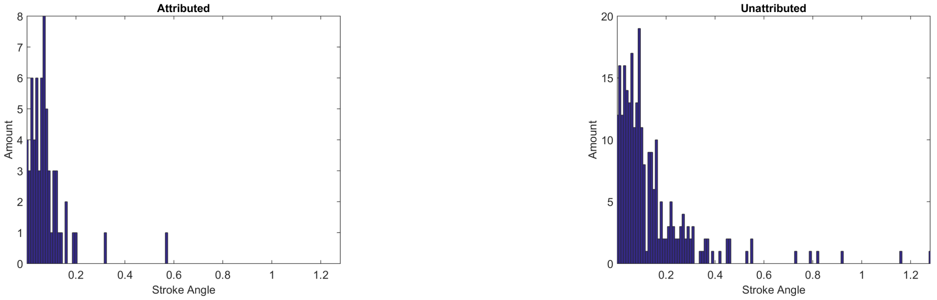
Figure 3.10: Histogram of Stroke Curvature of attributed and unattributed marks, **a)** Attributed mark, **b)** Unattributed marks

(a) (b) (c)

Figure 3.11: Example of Unattributed marks with large Stroke Curvature values, **a)** CN: 40.2872, **b)** CN: 84.1715, **c)** CN: 101.2632

3.2.5 Stroke Angle

The histograms of Stroke Angle can be seen in Figure 3.12. Attributed marks had a 90% value of 0.146, unattributed marks had 30.7% of marks above this value. Marks with an angle greater than 0.15 could be an indication of having been created by an author other than Melville. Figure 3.13 shows examples of the maximum Stroke Angle values from both the unattributed and attributed marks.



(a) (b)

Figure 3.12: Histogram of Stroke Angle of attributed and unattributed marks,
a) Attributed marks, **b)** Unattributed marks

(a) (b)

Figure 3.13: Attributed and unattributed marks max valued Stroke Angle,
a) Attributed marks, Stroke Angle: 0.571, **b)** Unattributed mark, Stroke Angle: 1.28

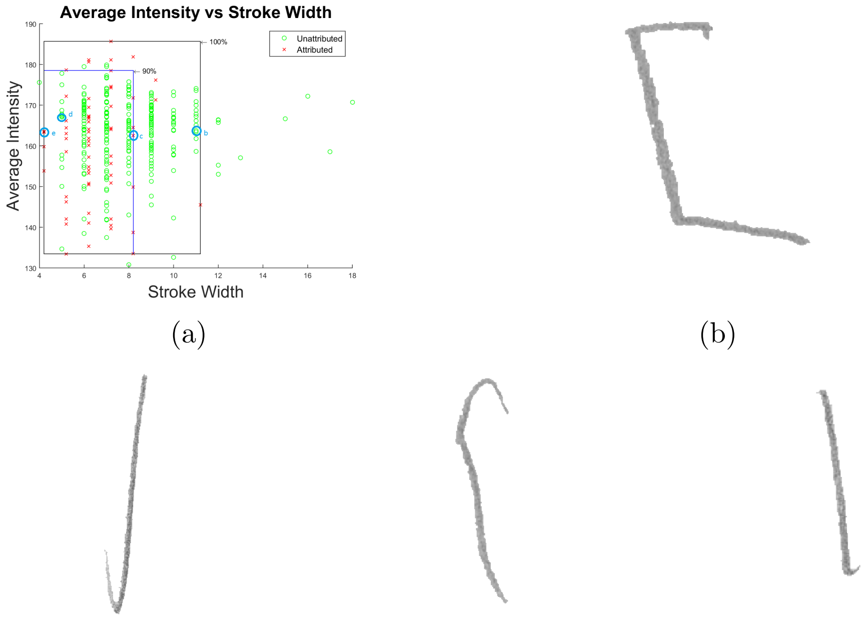
When the individual features of the unattributed and attributed marks were compared, some patterns emerged. The Unattributed marks have marks with values that are outside the normal range of those seen with the Attributed marks. These features with noticeable discrepancies are blurriness, stroke curvature, and stroke angle. Unattributed mark group that have small blurriness values, large stroke curvature, and large stroke angle values could be an indicator of an author other than Melville. We will next look at features in a 2D when feature are combined.

3.3 2D Feature Space: Comparing Attributed and Unattributed Marks

Here we look at pairs of features together as these could create correlating patterns between features. By plotting data from both unattributed marks and marks attributed to Melville on the same plot, we directly compare the features and their clustering to each other.

We have labeled the data points of the unattributed marks with green circles and the data points of the attributed marks with red x's to distinguish between unattributed and attributed marks in Figure 3.14a. Stroke Width is an integer valued feature. To make the plots more legible, the unattributed and attributed data points are displayed offset from each other in the feature space plots by incrementing each Stroke Width feature of the attributed mark by 0.2. Two boxes have been added around the attributed marks in each of the figures to highlight how the marks were clustered for the given features. One box contains 90% of the attributed marks, while the second box contains all attributed marks, see Table 3.3 for these values. The 10% of the marks contained in the difference could be outliers.

Average Intensity versus Stroke Width, Figure 3.14a, shows that average intensity was evenly spread for both unattributed and attributed marks, with no discernible difference between them. Figure 3.14b shows an example of an unattributed mark with a stroke width value of 11. Figure 3.14c shows an attributed mark with a stroke width value of 8. Figure 3.14d and Figure 3.14e show the minimum stroke width values for the unattributed and attributed marks respectively. Marks that had a larger stroke width than the average could be an indicator that those marks belong to an author other than Melville.



(a)

(b)

(c)

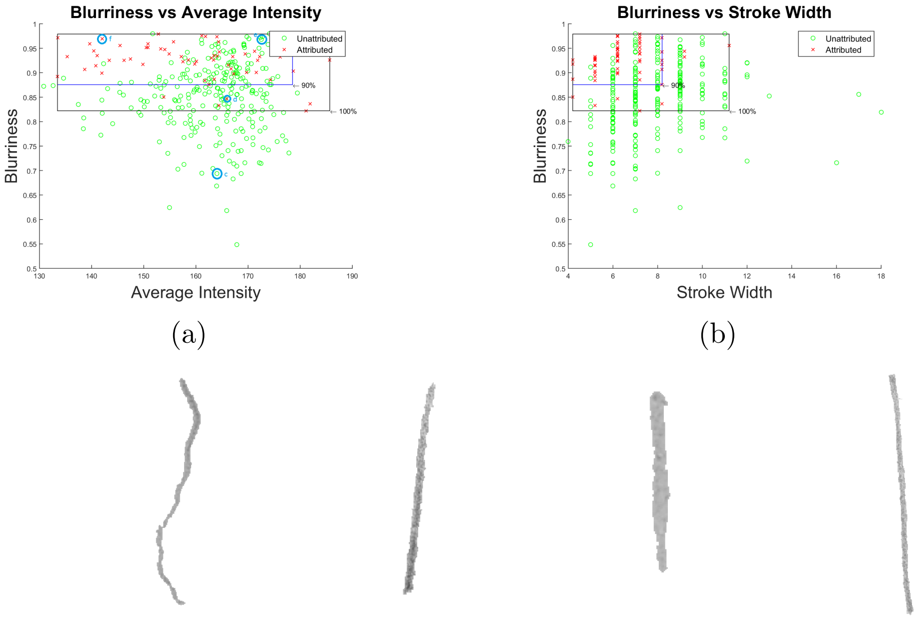
(d)

(e)

Figure 3.14: Comparing Unattributed and Attributed marks, Average Intensity vs. Stroke Width **a)** 2D plot of Average Intensity vs. Stroke Width, **b)** Unattributed mark, stroke width: 11, **c)** Attributed mark, stroke width: 8, **d)** Unattributed mark, stroke width: 5, **e)** Attributed mark, stroke width: 4

While the Blurriness versus Average Intensity plot, Figure 3.15a, confirms that average intensity has a similar distribution between the unattributed marks and the attributed marks, attributed marks have large values of blurriness and are clustered toward the top of the graph, and unattributed marks have small values of blurriness and are more distributed throughout the graph. A large portion of unattributed marks are not contained within the 100% zone of attributed mark values. The figure for Blurriness and Stroke Width, Figure 3.15b, shows the same trend in the Blurriness feature. Figure 3.15c and 3.15d are examples of the smallest value of blurriness of each respective unattributed and attributed marks. The maximum value of the unattributed and attributed marks were around 0.95, as seen in Figure 3.15e and

3.15f. This highlights that low values of blurriness could be a strong indicator of marks that are not Melville's.

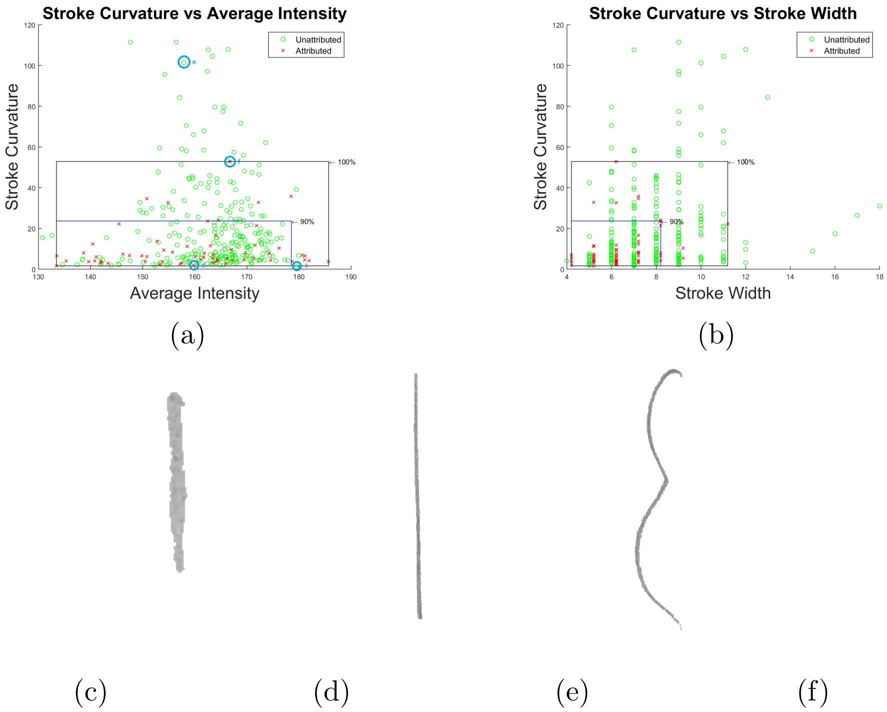


(a) (b)

(c) (d) (e) (f)

Figure 3.15: Comparing Unattributed and Attributed marks, Blurriness vs. Average Intensity **a)** 2D plot of Blurriness vs. Average Intensity, **b)** Unattributed mark, Blurriness: 0.694, **c)** Attributed mark, Blurriness: 0.847, **d)** Unattributed mark, Blurriness: 0.953, **e)** Attributed mark, Blurriness: 0.963

The plot for Average Intensity versus Stroke Curvature, Figure 3.16a, shows that unattributed marks have a large spread of stroke curvature values, while attributed marks are tightly clustered with small values. The plot of Stroke Curvature versus Stroke Width, Figure 3.16b, shows the same trend. Figure 3.16c and 3.16d show unattributed and attributed marks that had a small stroke curvature value. Figure 3.16f shows the largest stroke curvature value that the attributed marks had, showing a slightly curved line. Figure 3.16e was an unattributed mark that has a curvature value of 101, this mark was in the shape of a bracket. Large values of stroke curvature could be an indicator they were created by an author other than Melville.



(a) (b)

(c) (d) (e) (f)

Figure 3.16: Comparing Unattributed and Attributed marks, Average Intensity vs. Stroke Curvature **a)** 2D plot of Average Intensity vs. Stroke Curvature, **b)** 2D plot of Stroke Curvature vs. Stroke Width, **c)** Unattributed mark, Stroke Curvature: 1.55, Average Intensity: 180, **d)** Attributed mark, Stroke Curvature: 1.87, Average Intensity: 160, **e)** Unattributed mark, Stroke Curvature: 101, Average Intensity: 158, **f)** Attributed mark, Stroke Curvature: 52.9, Average Intensity: 167

Figure 3.17 shows four other examples of unattributed marks that have curvature numbers larger than 100. Each of these examples had shapes that indicated that they were used as brackets. This might be an indication that marks in the shape of brackets, or marks that have a high Stroke Curvature, stand out from other marks as an indication of a different writer.

Figure 3.18 shows the Stroke Curvature versus Blurriness plot. This group was clustered at the bottom right of the graph, with attributed marks containing large value curve numbers and small blurriness values. Unattributed marks have much larger extremes in values of stroke curvature and blurriness than is seen in the attributed mark group. There are attributed marks that have larger values of stroke

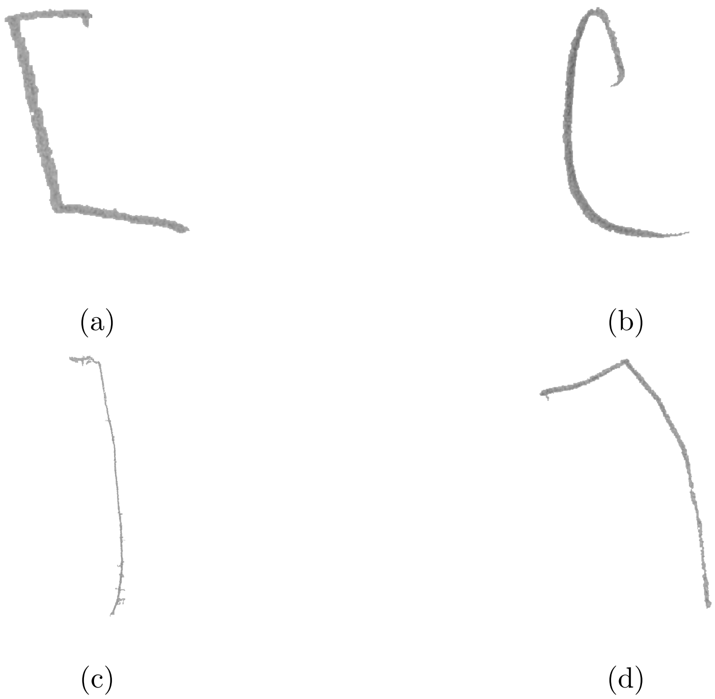
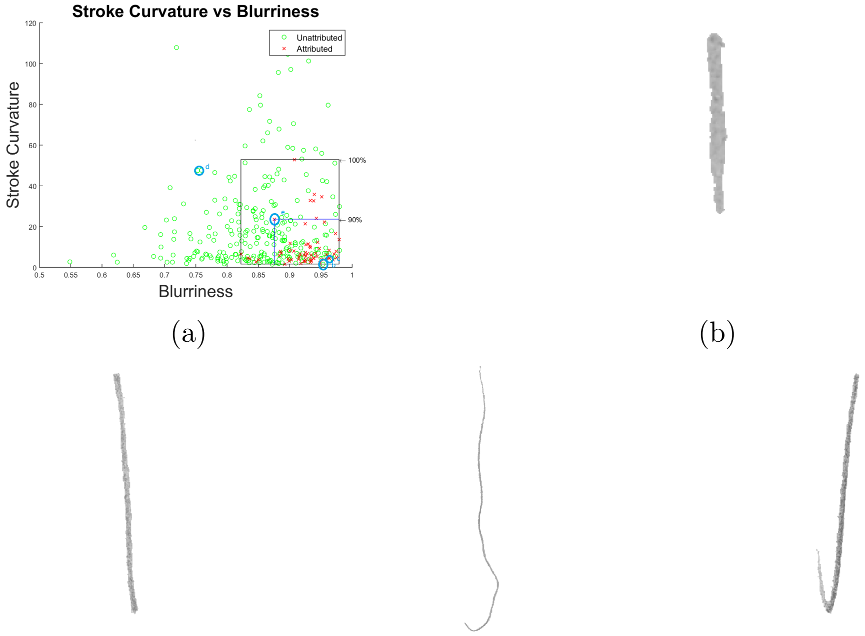


Figure 3.17: Unattributed marks with large Stroke Curvature
a) Stroke Curvature: 105, **b)** Stroke Curvature: 111, **c)** Stroke Curvature: 108, **d)** Stroke Curvature: 108

curvature or smaller values of blurriness, but these are not seen in the same mark. There are many unattributed marks that have both features at these extreme values. While unattributed marks are clustered in the same patterns as found in attributed marks, there are many marks that are outside this pattern.

Figure 3.19a shows a plot for Stroke Angle versus Average Intensity. There are six unattributed marks that have a larger value than the largest attributed mark, with two of those unattributed marks having twice the stroke angle value. If we exclude the largest 10% of the attributed marks, the largest value drops by two thirds. While the majority of unattributed marks are clustered the same way, many have a stroke angle that is larger than the average attributed mark. Both Figure 3.19b and 3.19c show examples of small stroke angle values from unattributed and attributed marks,



(a)

(b)

(c)

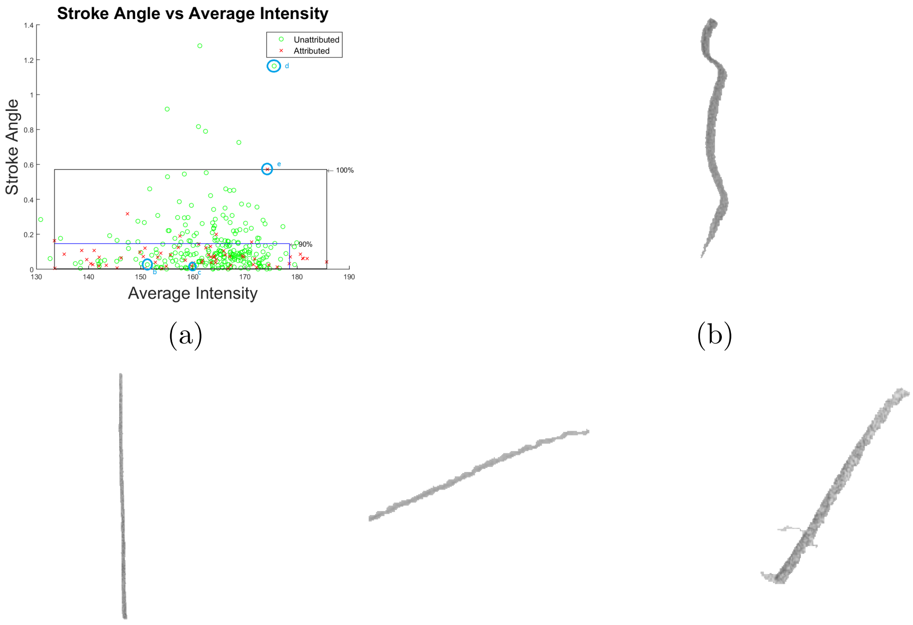
(d)

(e)

Figure 3.18: Comparing Unattributed and Attributed marks, Stroke Curvature vs. Blurriness. **a)** 2D plot of Stroke Curvature vs. Blurriness, **b)** Unattributed mark, Stroke Curvature: 1.55, Blurriness: 0.953, **c)** Attributed mark, Stroke Curvature: 4.54, Blurriness: 0.963, **d)** Unattributed mark, Stroke Curvature: 47.6, Blurriness: 0.753, **e)** Attributed mark, Stroke Curvature: 23.6, Blurriness: 0.876

these marks were almost perfectly vertical. Figures 3.19d and 3.19e shows examples of the maximum stroke angle value that each group of marks contain. The majority of the attributed marks had a stroke angle less than 0.2, the unattributed mark group had many marks that were above this value. Marks with a stroke angle greater than 0.2 could be an indication of having been created by an author other than Melville.

Figure 3.20a shows Stroke Curvature versus Stroke Angle. Both unattributed and attributed marks showed the trend of low values of both stroke curvature and stroke angle. The attributed mark group did have marks that had larger values than the main group, but those marks either had a larger value of stroke curvature or stroke angle, but not both in the same mark. Both Figure 3.20b and Figure 3.20c show



(a)

(b)

(c)

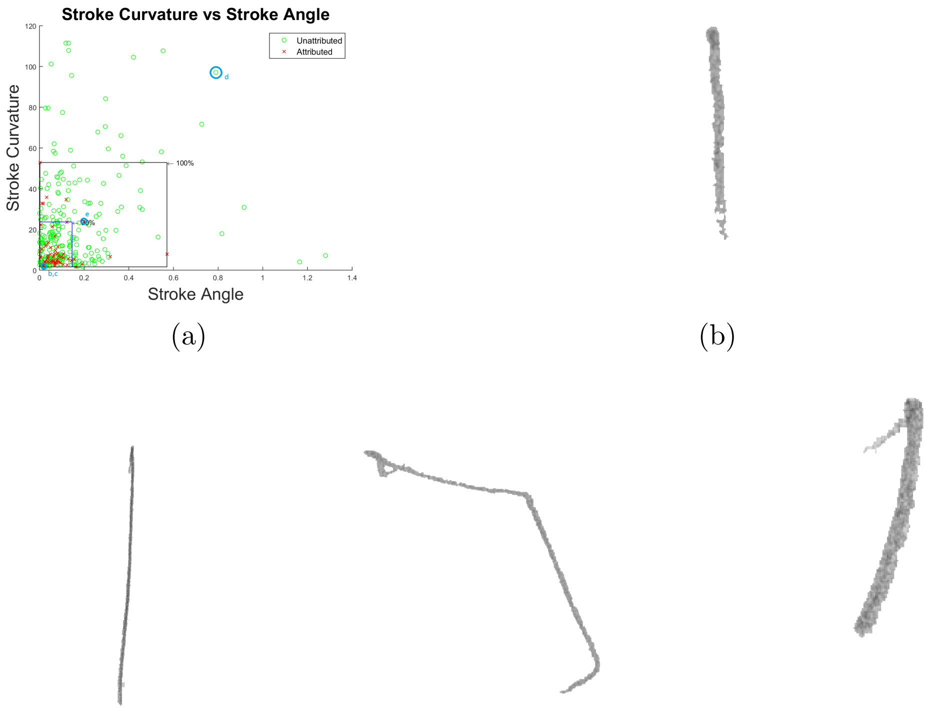
(d)

(e)

Figure 3.19: Comparing Unattributed and Attributed marks, Stroke Angle vs. Average Intensity. **a)** 2D plot of Stroke Angle vs. Average Intensity, **b)** Unattributed mark, Stroke Angle: 1.55, Average Intensity: 0.953, **c)** Attributed mark, Stroke Angle: 4.54, Average Intensity: 0.963, **d)** Unattributed mark, Stroke Angle: 47.6, Average Intensity: 0.753, **e)** Attributed mark, Stroke Angle: 23.6, Average Intensity: 0.876

example marks that had small values for both features, they show vertical marks that have little curvature. Both Figure 3.20d and Figure 3.20e show example marks showing an attributed and unattributed mark that has both a large stroke angle value and a large curvature value. The unattributed mark was in the shape of a bracket that was tilted. Marks that had both a large stroke angle value and a large stroke curvature value did not follow the pattern established by the attributed marks. These marks could belong to a different author than Melville.

Figure 3.21a shows the plot of Stroke Angle versus Blurriness. Figure 3.21 shows example images highlighting marks with similar stroke angle and blurriness values. With Figure 3.21b and 3.21c showing examples of marks that have similar values of



(a) (b)

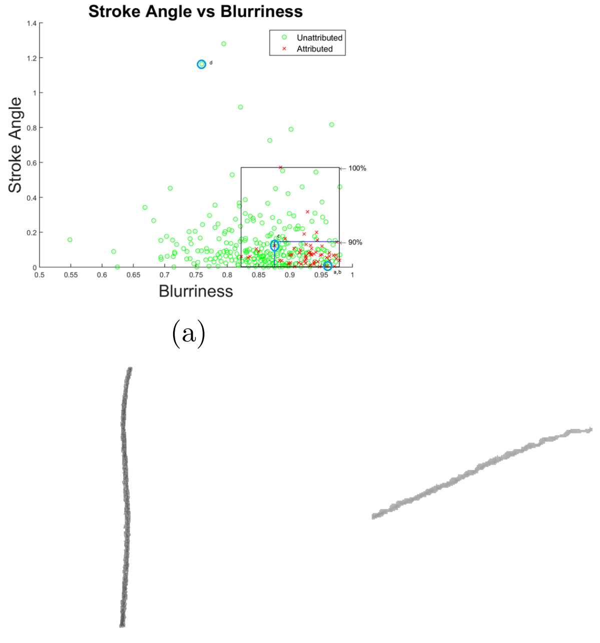
(c) (d) (e)

Figure 3.20: Comparing Unattributed and Attributed marks, Stroke Curvature vs. Stroke Angle. **a)** 2D plot of Stroke Curvature vs. Stroke Angle, **b)** Unattributed mark, Stroke Curvature: 2.46, Stroke Angle: 0.042, **c)** Attributed mark, Stroke Curvature: 3.74, Stroke Angle: 0.055, **d)** Unattributed mark, Stroke Curvature: 97.2, Stroke Angle: 0.789, **e)** Attributed mark, Stroke Curvature: 24, Stroke Angle: 0.199

stroke angle and blurriness values, these marks could be from Melville. Figure 3.21d and 3.21e show examples of marks that have small values for blurriness with larger stroke angle values. The unattributed mark stood out by having a noticeable tilt to the mark.

The plot of Stroke Angle versus Stroke Width is shown in Figure 3.22a. Both stroke angle and stroke width showed minor differences between unattributed and attributed marks. Figure 3.22 shows examples from the Stroke Angle vs. Stroke Width plot. There were some unattributed marks that had high stroke angle values, or had large stroke width values, but very few marks had both of these characteristics. Marks that have both a large value of stroke angle and a relatively large value of stroke width could be an indicator of a mark created by someone other than Melville.

Some of the discrepancy between the data sets could be from the difference in the number of collected data points. We only obtained 63 attributed marks, this is about one forth the number of unattributed marks, it is possible that with a similar number of attributed marks the data points could display similar patterns and distribution to that seen in the unattributed mark group. C-means clustering was run on the features collected and it will be discussed next.



(a)

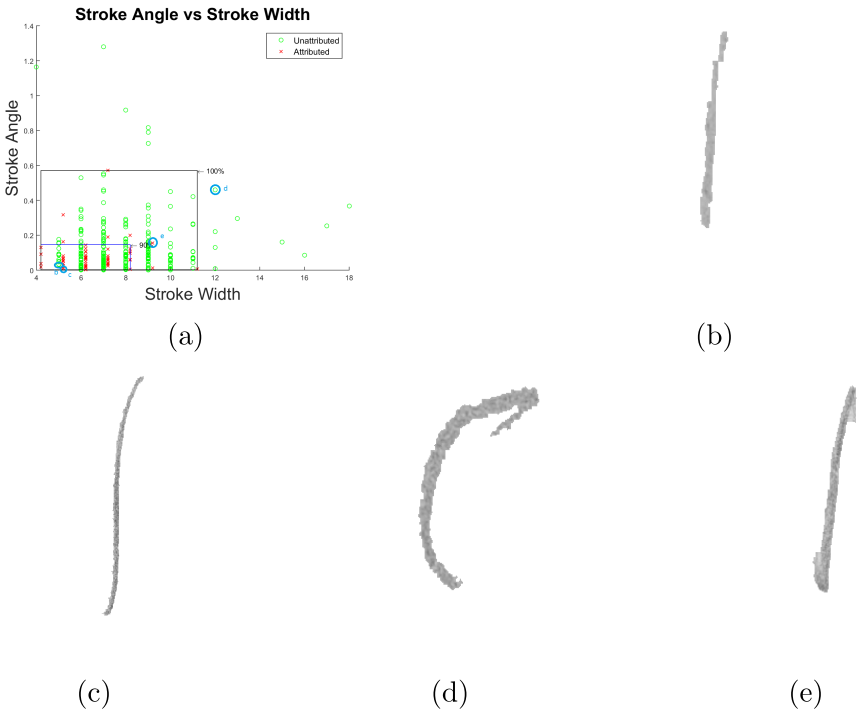
(b)

(c)

(d)

(e)

Figure 3.21: Comparing Unattributed and Attributed marks, Stroke Angle vs Blurriness. **a)** 2D plot of Stroke Angle vs Blurriness, **b)** Unattributed mark, Stroke Angle: 0.063, Blurriness: 0.974, **c)** Attributed mark, Stroke Angle: 0.004, Blurriness: 0.971, **d)** Unattributed mark, Stroke Angle: 1.16, Blurriness: 0.76, **e)** Attributed mark, Stroke Angle: 0.106, Blurriness: 0.847



(a)

(b)

(c)

(d)

(e)

Figure 3.22: Comparing Unattributed and Attributed marks, Stroke Angle vs. Stroke Width. **a)** 2D plot of Stroke Angle vs. Stroke Width, **b)** Unattributed mark, Stroke Angle: 0.089, Stroke Width: 5, **c)** Attributed mark, Stroke Angle: 0.082, Stroke Width: 5, **d)** Unattributed mark, Stroke Angle: 0.46, Stroke Width: 12, **e)** Attributed mark, Stroke Angle: 0.156, Stroke Width: 9

3.4 C-means Clustering Results

C-means clustering was an algorithmically simple method to group marks in a multidimensional feature space. The Davies-Bouldin and Silhouette validation methods were used to validate the clusters that were created by the C-means clustering algorithm. A low value from Davies-Bouldin and a high value from Silhouette indicate stronger confidence in the cluster validity. Table 3.4 shows that as the number of clusters increases, the Davies-Bouldin value increases, and similarly the Silhouette results decrease, indicating that two clusters is more likely the correct number.

Table 3.4: Clustering validation results

Mark type	Validation method	2 clusters	3 clusters	4 clusters
Unattributed	Davies-Bouldin	0.6425	0.6849	0.7696
Unattributed	Silhouette	0.8053	0.6967	0.6051
Attributed	Davies-Bouldin	0.8872	0.7185	0.6600
Attributed	Silhouette	0.5605	0.6780	0.6327

For comparison, clustering was performed on attributed marks, with the assumption that all the marks were created by a single author. Table 3.4 shows the results; as the number of clusters increases, the confidence value of each method improves. If all the data falls into a single cluster, then as the number of clusters increases the confidence value will increase as small clusters are formed.

3.5 Semi-Supervised Learning

We have looked at C-means clustering on both the attributed and unattributed marks and now we will use this information to perform semi-supervised learning.

Each attributed mark was paired with an unattributed mark, then labeled as belonging to the cluster to which the unattributed mark belongs. This pairing was determined by smallest euclidean distance between the marks.

Row one of Table 3.5 shows the number of unattributed marks that were placed into each of the two clusters. Of the 261 total number of marks, 214 were placed into Cluster 1, and 47 were placed into Cluster 2.

Row two of Table 3.5 shows that when attributed marks were paired with unattributed, 61 marks were placed into Cluster 1, and two marks were placed into Cluster 2. The two pairs of marks were shown in Figure 3.23. The two marks from the unattributed mark look similar to each other, and the two marks from the attributed mark also look like each other. But the two marks do not visually look similar to their paired mark.

The attributed and unattributed mark's features sets were merged and C-means clustering was performed. The new set had 322 marks; of those marks, 293 were placed into Cluster 1 and 29 into Cluster 2 (Table 3.5 row three), with 62 of the 63 attributed marks having been placed into Cluster 1. Of the two attributed marks that were paired into Cluster 2, one was placed into Cluster 2 by the C-means clustering algorithm and this mark can be seen in Figure 3.23c.

Table 3.5: Attributed and unattributed mark clustering

	Total Marks	Cluster 1	Cluster 2
Unattributed marks	261	214	47
Attributed marks	63	61	2
Attributed and Unattributed marks	324	276	48
Attributed marks	63	62	1

Table 3.6 shows the Davies-Bouldin and Silhouette validation results on the combined unattributed and attributed mark C-means clustering. The Davies-Bouldin validation showed the smallest value with two clusters, and then increased as the number of clusters increased. There was a dip in the value when three clusters were chosen, but at four clusters the values increased again, indicating that that choosing two clusters had the highest confidence. The Silhouette validation method started with the highest value and steadily decreased as the number of clusters was increased. The two validation methods showed trends that mirrored the trends seen when clustering validation was performed on unattributed marks.

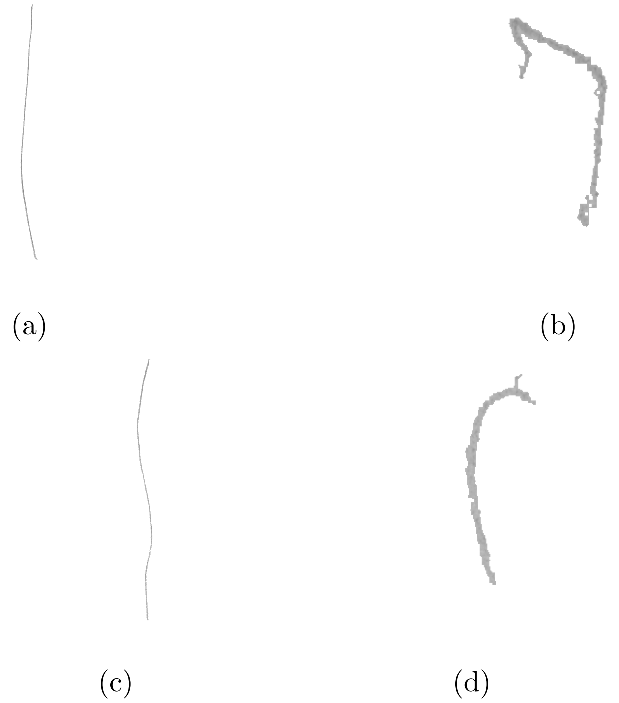


Figure 3.23: The two attributed marks that were placed into cluster 2 and their matched unattributed marks **a)** Attributed, **b)** Unattributed, **c)** Attributed, **d)** Unattributed

Table 3.6: Clustering of attributed and unattributed marks validation table

Validation type	2 clusters	3 clusters	4 clusters	5 clusters
Davies-Bouldin	0.6447	0.7205	0.6881	0.7248
Silhouette	0.8118	0.6719	0.6427	0.6233

When C-means clustering was performed on the unattributed marks, the confidence value was the highest for when it separated the marks into two groups. The majority of marks were placed into a single cluster by the clustering algorithm. The smaller of the two clusters contained marks with high stroke curvature values. Semi-supervised learning placed all but two of the attributed marks into Cluster 1. When the C-means clustering was performed on the combined feature space of both unattributed and attributed marks all but one of the attributed marks were placed into Cluster 1. This could indicate that marks that are placed into Cluster 2 belong to an author other than Melville.

3.6 Melville's Marginalia Online Attributed Labels

The staff at Melville's Marginalia Online has visually inspected and assigned labels to the marks on each page. They have attributed marks to Melville based on close proximity to Melville's handwriting and their experience looking at the marks. Marks that they can not attribute to Melville they label as unattributed. Throughout this thesis we have been using attributed and unattributed labels to represent marks from specific books. In this section we will refer to the labels given on Melville's Marginalia Online as assigned and unassigned respectively. *The Vision* (Dante) contained marks that Melville's Marginalia Online had labeled both as assigned and unassigned. We will compare the distribution and spread of these labeled marks to the other experiments that we conducted in this section. From our results we ended up with two groups, a larger mark group that had similar characteristics to those found in the Shakespeare anthology group, and a smaller group that did not conform to those characteristics. We would like to see a similar pattern appear with Melville's Marginalia Online labels as to what we have observed in our own analysis and clustering.

There are two groupings that we will look at: a grouping based on book section, and a grouping of marks labeled as assigned or unassigned. We will look at histograms of measured feature values with each of these labels applied and then we will look at 2D scatter plots with each of these labels applied. We will also look for similarities between Melville's Marginalia Online labeled marks and the patterns created by the marks from the Shakespeare anthology.

3.6.1 Book Sections

The book *The Vision* (Dante) contains three sections: Hell, Purgatory, and Paradise. Dr. Olsen-Smith has observed some commonalities of the marks within each of the sections. He has also noted that Melville showed more interested in specific sections than in others and has annotated more concerning some topics than others. We will be looking for patterns inherent in the marks within a book section, and comparing them to the marks from the Shakespeare anthology.

Paradise had the largest collection of marks, with 209 marks coming from that section, Hell only had 39 marks and Purgatory was the smallest with 13, see Table 3.7. Of those marks Purgatory had 12 assigned marks, Hell had 29, and Paradise had only 16 marks that were assigned to Melville.

Table 3.7: Unattributed marks by section and assigned or unassigned labels

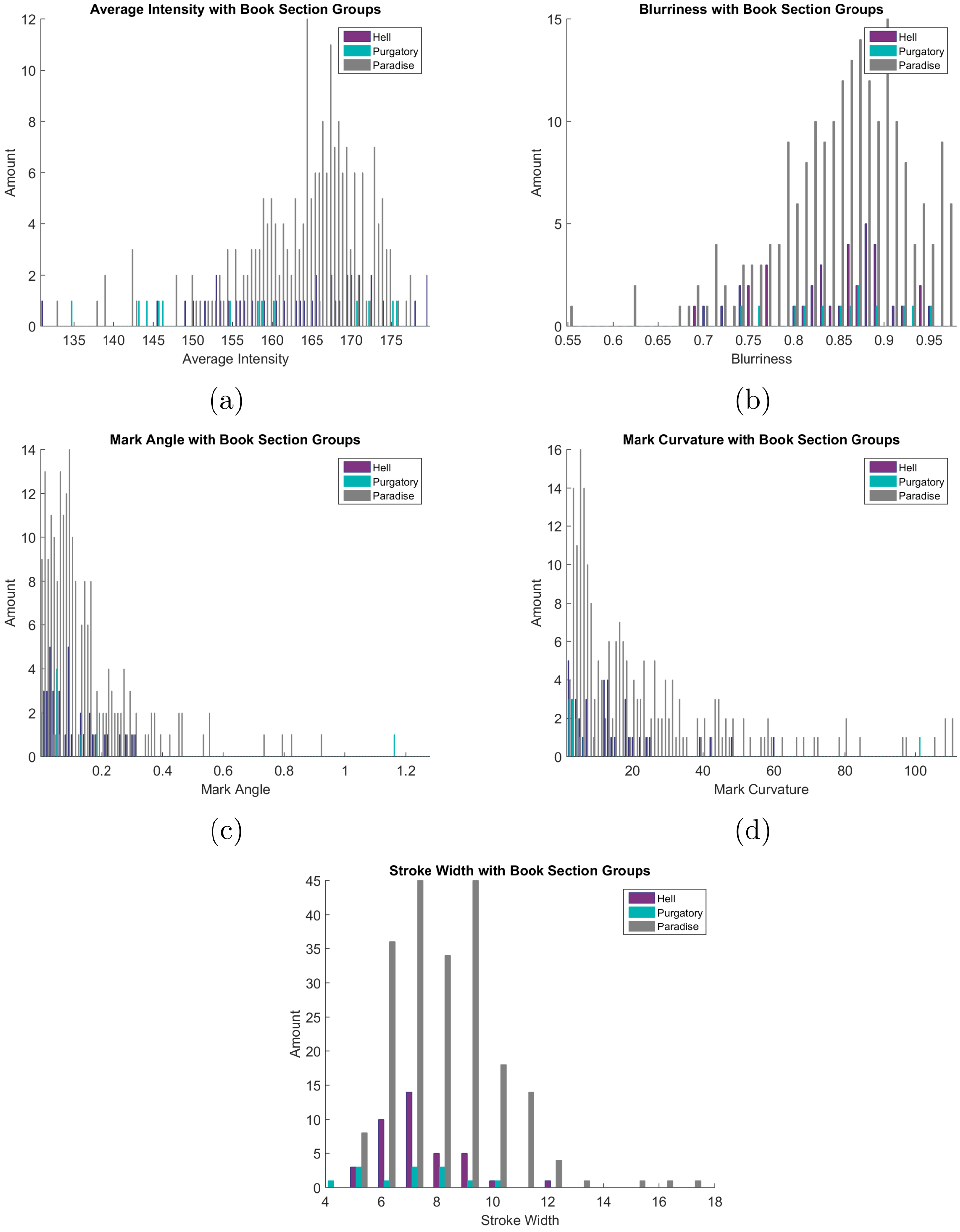
Marks	Assigned	Unassigned	Total
Hell	29	10	39
Purgatory	12	1	13
Paradise	16	193	209
Total	57	204	261

Histograms were created for each feature with these labels applied, see Figure 3.24. The histogram of Average Intensity, see Figure 3.24a, shows that marks in the Paradise section retained the Gaussian shape, see Figure 3.1e with all sections combined. The histograms for the other two sections showed a uniform distribution. Figure 3.24b, histogram of Blurriness, shows that marks from the Purgatory section have a uniform distribution while marks from the Hell and Paradise sections show similar patterns and spread to each other. The histogram of Mark Angle, see Figure 3.24c, showed no discernible difference in the spread or distribution of marks from different sections.

In the histogram of Mark Curvature, Figure 3.24d, the notable characteristic is that there was only a single Purgatory mark that had a value larger than 20, while the other two book sections had multiple marks above the Mark Curvature value of 20. The histogram of Stroke Width, Figure 3.24e, shows that marks from Purgatory were uniformly distributed, while the marks from both Hell and Paradise show a Gaussian distribution. The marks from Paradise have a larger spread and a mean value that is larger than the mean of marks from the Hell section. It is possible that marks from Paradise have a different Average Intensity characteristic than those seen in other sections, but this perceived difference could also be due to the large discrepancy in the number of marks found in Paradise versus the other two sections.

Figure 3.25 and Figure 3.26 show the 2D scatter plots, again separated by the section of *The Vision* (Dante) in which they appeared. We were looking for patterns or similarities in the distributions and spread of the marks in the scatter plots. Marks from the Purgatory section are the marks that had the more extreme values and were outside the values seen in the marks from the Shakespeare anthology. No other patterns appear in these plots.

We were looking for common patterns between the book sections and/or the attributed marks. Due to the low number of marks in the Hell and Purgatory sections it was difficult to judge the type of distribution they would have. As such we could not see distinct patterns created by the marks from specific book sections. We did not see any similarities between marks from specific groups and the attributed marks.



(a)

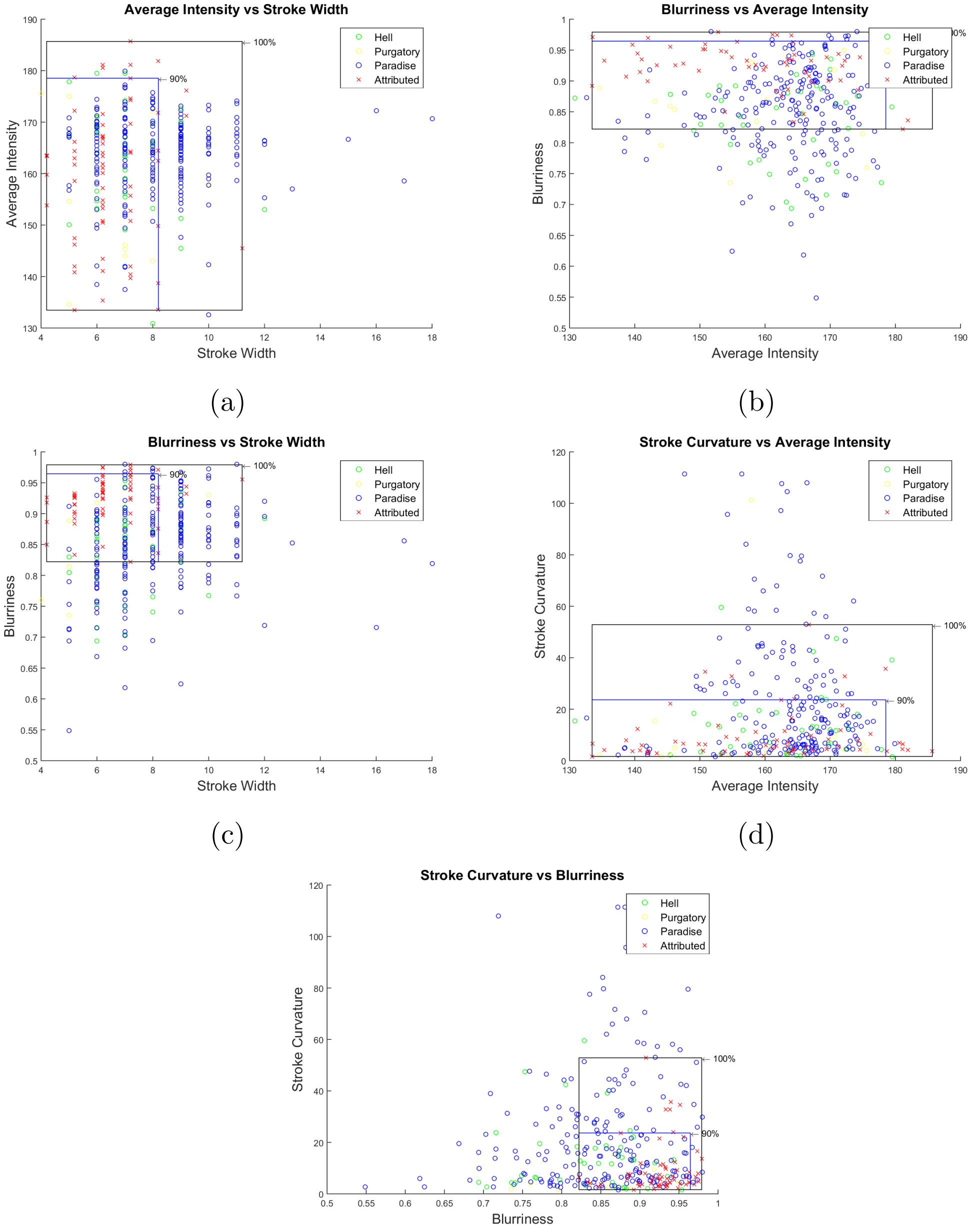
(b)

(c)

(d)

(e)

Figure 3.24: Histogram of features with book section labels applied
a) Average Intensity, **b)** Blurriness, **c)** Mark Angle, **d)** Mark Curvature, **e)** Stroke Width



(a)

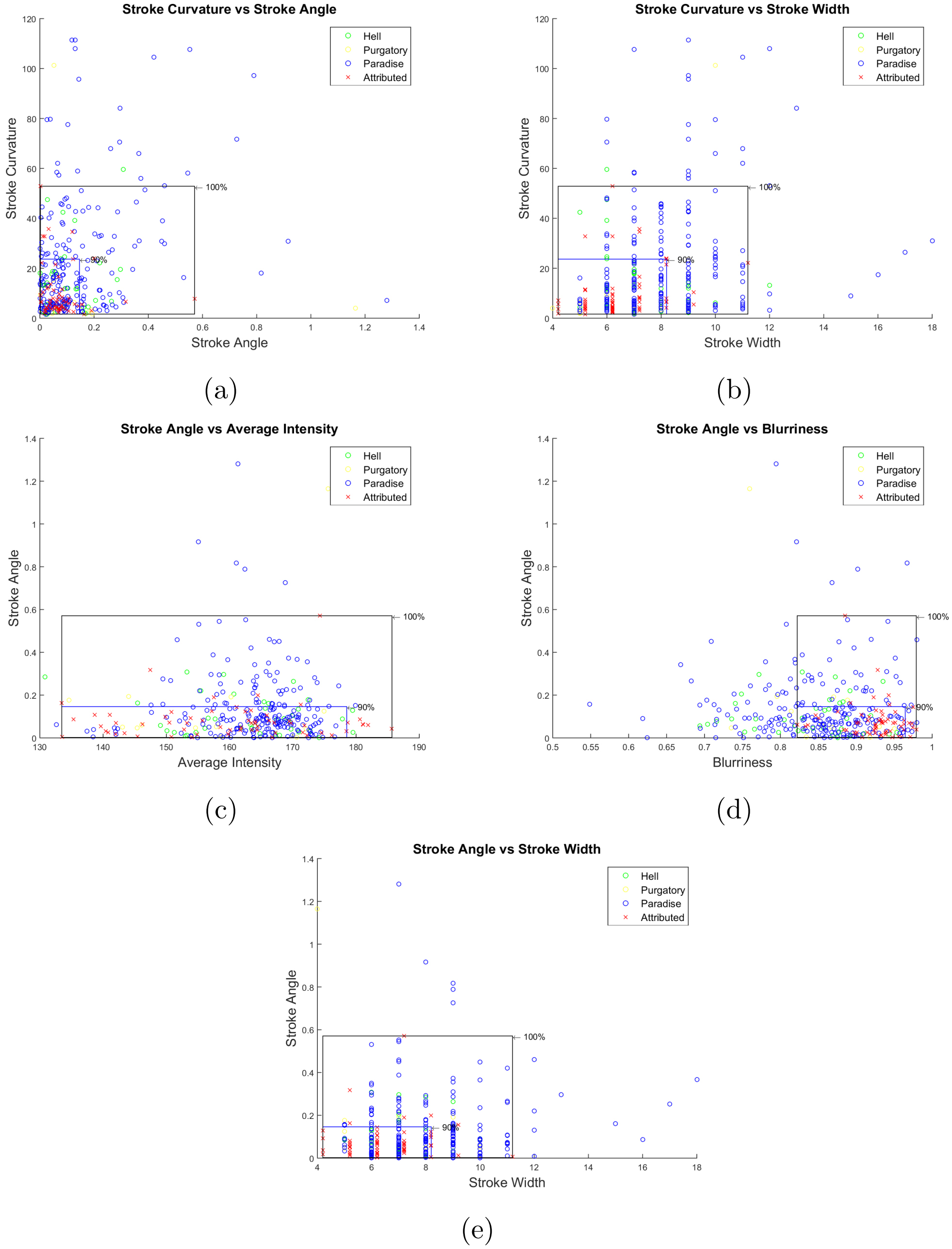
(b)

(c)

(d)

(e)

Figure 3.25: 2D plots with book section labels applied
a) Average Intensity vs. Stroke Width, **b)** Blurriness vs. Average Intensity, **c)** Blurriness vs. Stroke Width, **d)** Mark Curvature vs. Average Intensity, **e)** Mark Curvature vs. Blurriness



(a) (b) (c) (d) (e)

Figure 3.26: 2D plots with book section labels applied, continued
a) Mark Curvature vs. Mark Angle, **b)** Mark Curvature vs. Stroke Width, **c)** Mark Angle vs. Average Intensity, **d)** Mark Angle vs. Blurriness, **e)** Mark Angle vs. Stroke Width

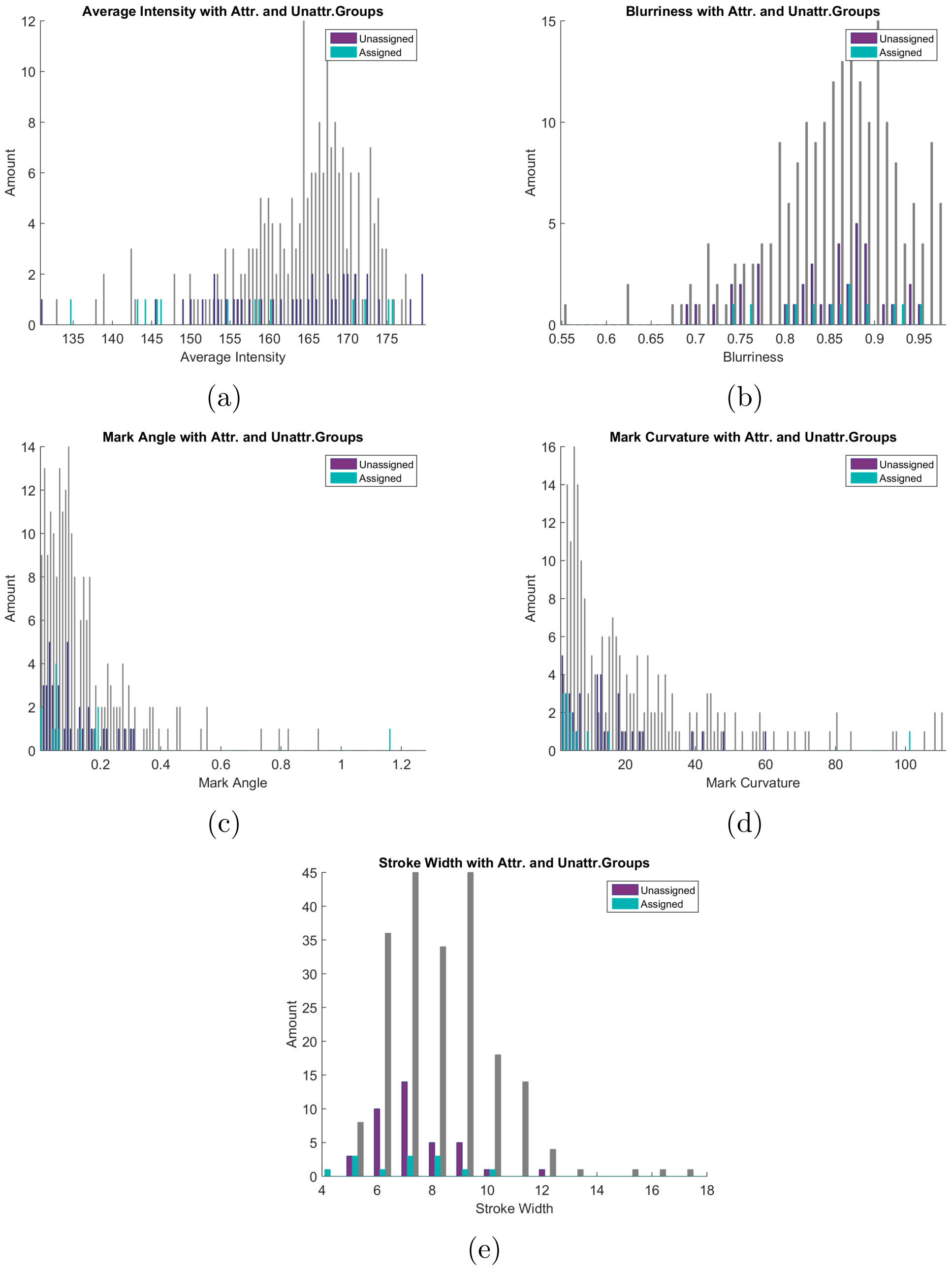
3.6.2 Assigned and Unassigned

Melville's Marginalia Online has assigned 57 marks to Melville and left 204 marks as unassigned in *The Vision* (Dante), Table 3.7. We want to see if the marks that were left unassigned by Melville's Marginalia Online are the marks that did not follow the set patterns seen from the Shakespeare anthology marks. First we will inspect the histograms looking for patterns in the distribution and spread, and how these label groups compare to the pattern and spread created from the marks from the Shakespeare anthology book. We will then look at the 2D scatter plots with these labels applied.

The histogram of Average Intensity, Figure 3.27a, shows a Gaussian distribution for unassigned marks while the assigned marks have a flat distribution. Assigned marks have lower values of average intensity, and have a larger spread than that seen in the unassigned marks. In the Blurriness histogram, Figure 3.27b, the two groups have similar distribution and spread to each other. With Mark Angle, Figure 3.27c, the unassigned and assigned marks show a similar pattern at low values of mark angle. Only unassigned marks have marks with values greater than 0.3. The histogram of Mark Curvature, Figure 3.27e, has two marks from the assigned group that has a value greater than 30, while the unassigned group has a large number of marks with a Curvature value greater than 30. The histogram of Stroke Width shows that the unassigned mark group has larger values of Stroke Width than what is seen in the assigned group.

Lastly we will look at the 2D plots when assigned and unassigned labels are applied. The plot of Blurriness versus Average Intensity, Figure 3.28b, shows that both the assigned and the attributed groups have marks with low values of average intensity. The other interesting plot is of Stroke Curvature versus Stroke Angle, Figure 3.29a. In this plot we can see that the majority of the assigned marks are in the bottom left corner: while not as tightly grouped as those seen in the Shakespeare anthology mark group, most of them are still within that 90th percentile.

While labeling the marks based on their books sections did not reveal much information, labeling the marks based on Melville's Marginalia Online assigned and unassigned labels did provide some information. Melville's Marginalia Online was only able to assign a small number of marks to Melville from *The Vision* (Dante), while experiments that we have performed indicate that there are possibly many more Melville marks contained in *The Vision* (Dante). The marks from the Shakespeare anthology book showed tight grouping of feature values while Melville's Marginalia Online assigned marks have a much larger spread and distribution and do not follow the same pattern. It is possibility is that the writing pencil changed dramatically between the creation of the different marks, or were made under different writing conditions. The marks could have also been created at different times in Melville's life. A larger collection of assigned Melville marks would create a better picture of their characteristics.



(a)

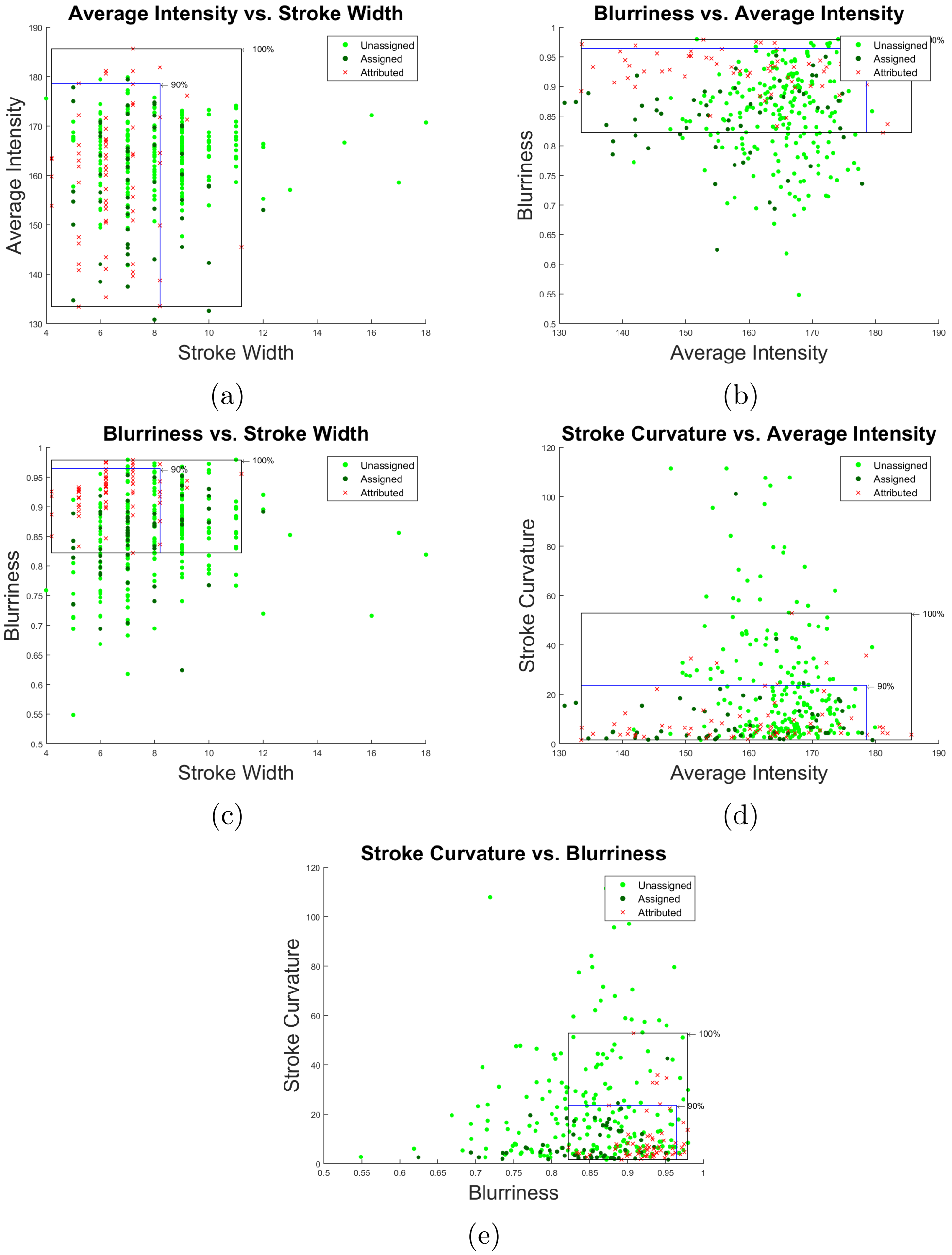
(b)

(c)

(d)

(e)

Figure 3.27: Histogram of features with assigned and unassigned labels applied
a) Average Intensity, **b)** Blurriness, **c)** Mark Angle, **d)** Mark Curvature, **e)** Stroke Width



(a)

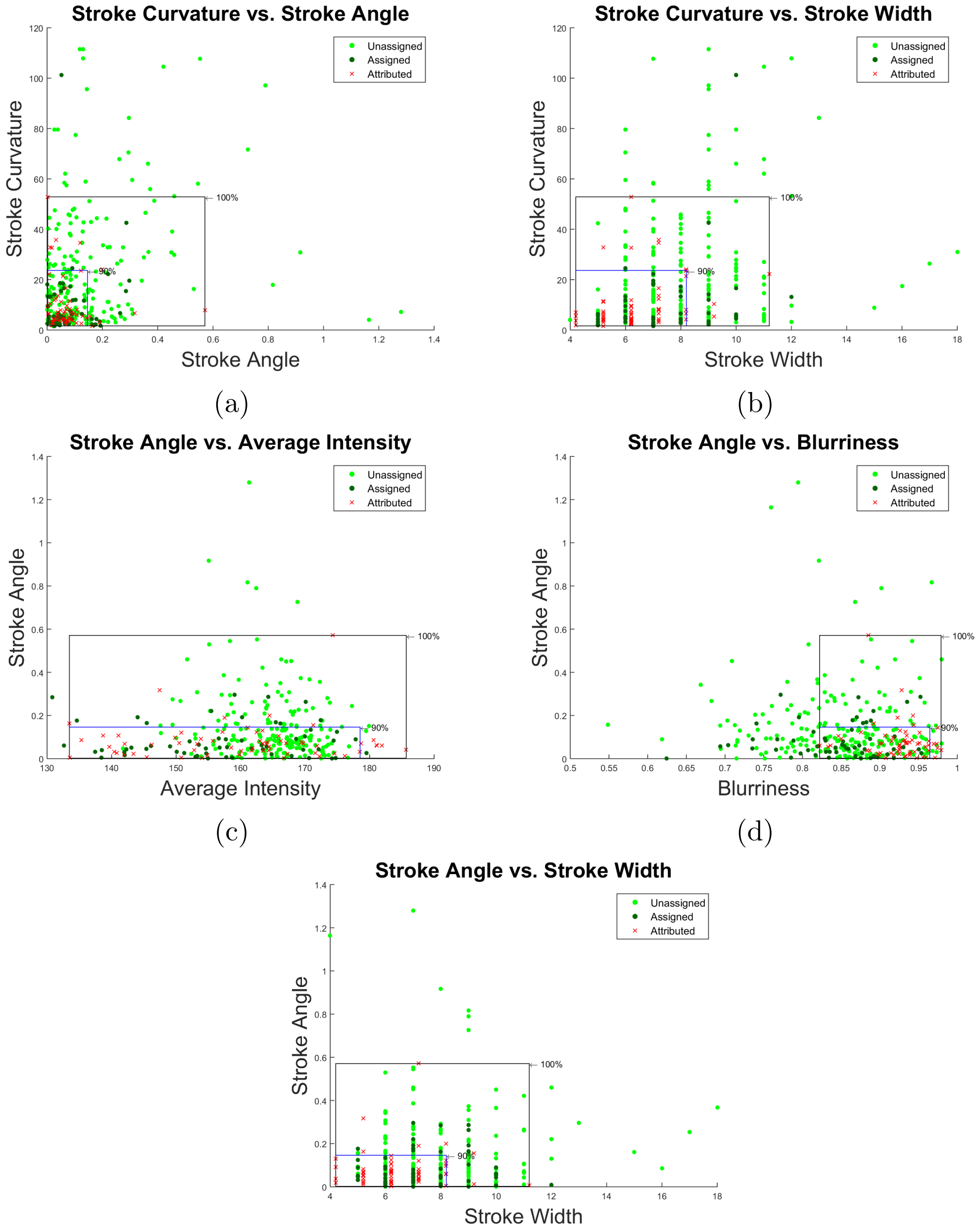
(b)

(c)

(d)

(e)

Figure 3.28: 2D plots with assigned and unassigned labels applied
a) Average Intensity vs. Stroke Width, **b)** Blurriness vs. Average Intensity, **c)** Blurriness vs. Stroke Width, **d)** Mark Curvature vs. Average Intensity, **e)** Mark Curvature vs. Blurriness



(a)

(b)

(c)

(d)

(e)

Figure 3.29: 2D plots with assigned and unassigned labels applied, continued
a) Mark Curvature vs. Mark Angle, **b)** Mark Curvature vs. Stroke Width, **c)** Mark Angle vs. Average Intensity, **d)** Mark Angle vs. Blurriness, **e)** Mark Angle vs. Stroke Width

CHAPTER 4

CONCLUSION

The purpose of this thesis was to explore the feasibility of using image processing algorithms in determining authorship of hand-written pencil marks. We were given images from two books; one book had marks attributed to Melville, the other book contain marks attributed to Melville and unattributed marks from unknown authors. The marks were compared and determined if any marks in the unattributed mark group matched the profile created by the attributed mark group. We will cover a summary of the work performed, what was learned from the experiments, and what could be done in future work.

4.1 Summary of Work

The first step was to segment the marks from the pages, and separate them into different types of marks. We looked at the vertical marks from both unattributed and attributed groups of marks. Five features were extracted from these marks; Average Intensity, Blurriness, Stroke Width, Stroke Angle, and Stroke Curvature.

Histograms of the extracted features were analyzed, we looked for bimodal and multimodal distributions. Out of the five features only two showed signs of a bimodal or multimodal distribution; average intensity and stroke curvature. Average intensity showed a possible multimodal distribution with four groups. These group labels were applied to the other features and then their respective histograms were evaluated for patterns and how the marks were distributed between the groups. Stroke curvature showed a possible bimodal distribution. The two group labels were applied to the other features and their histograms were analyzed.

The individual feature histograms from the unattributed and attributed marks were compared to each other. The features from the attributed and unattributed marks were compared using 2D scatter plots. We looked for groupings in the features and how the attributed and unattributed features were distributed when compared to each other.

C-means clustering algorithm was used to split up the marks into different clusters. Two forms of validation were performed on the clustering: Davies-Bouldin index and Silhouette. Using the clustering information obtained with C-means clustering, a semi-supervised algorithm was performed. This combined the two groups of marks, unattributed and attributed mark, and attempted to predict what clustering group the attributed mark would be placed in.

Lastly we looked at the marks from *The Vision* (Dante) under two types of labels. First we labeled the marks based on the book section from which they came. We then looked at them with Melvilles Marginalia's attributed and unattributed labels.

4.2 Conclusion

The work performed indicates that features extracted provide enough information from simple marks to distinguish between authors, instead of handwriting in the form of words.

We faced a number of challenges when comparing the two groups of marks. Some assumptions were made on the source of the marks, in that the same author created marks in both groups. The unattributed marks could have been created by the same author but at different times in their life, using different writing instruments, different locations and conditions, or the materials the book was created with. Approximately three times as many marks were unattributed than we collected of attributed marks. With more attributed marks it is possible it would show similar distribution to that found in the unattributed mark group.

We observed that there are marks in the unattributed mark group that do not follow the same pattern as laid out by the attributed mark group. These marks had common characteristics that we will discuss. The histogram of stroke curvature showed a bimodal distribution with one group having values larger than 38, and the other group with values less than 38. This grouping showed two Gaussian peaks in the stroke width histogram. From the 2D plot comparison we could see that attributed marks had a trend of smaller values of curvature. Running the clustering algorithm also grouped marks with a large curvature number together. Most marks with large curvature values were in the shapes of brackets and were used to bracket passages in the text. These type of marks were not present in the attributed marks group. Marks that have a stroke curvature value greater than 60 could be marks from another author. Blurriness is another feature that had unattributed marks that did not follow the pattern created by the attributed marks. Specifically marks with a blurriness value less than 0.8 could be from a different author. Another possible indicator that a mark is from a different author than Melville is if the stroke angle has a value greater than

0.6. Another indicator that a mark does not follow the pattern set by our attributed marks is if the stroke angle is greater than 0.4 coupled with the stroke curvature value greater than 40. The pattern that was created does not contain marks that have large values of stroke curvature coupled with large values of stroke angle. The last feature to look at was marks with a very large value of stroke width. While we could not directly compare the two stroke widths, we can look at the trend; the majority of the attributed marks have a stroke width value that is within a value of 5 of the peak. The unattributed mark has three marks that have values that are 7 greater than the peak.

In summary, marks that fall outside the attributed mark pattern have the following characteristics and could have been created by a different author: blurriness value less than 0.8, stroke angle greater than 0.6, stroke curvature greater than 60, stroke width greater than mean plus 7, and marks that have a stroke angle greater than 0.4 coupled with a stroke curvature greater than 40.

4.3 Future Work

Considerations for the future should consist of capturing images using a method that has a uniform light source and minimizes data outside the boundaries of the book's page. Images should then be saved using a lossless compression method. Other light spectrum could be utilized in order to obtain different information about the composition of the marks.

Only a single source of attributed marks was used; by utilizing marks from multiple sources the current profile could be expanded and refined. To do this we would extract more marks from books with marks that are attributed to Melville by Melville's

Marginalia Online. Doing this would expand the profile of Melville's marks, giving a better idea of the variance of how Melville created marks.

More features can be extracted from the marks and the methods described in this thesis can be used to analyze them. These features would look at different ways to differentiate marks by looking at how straight a mark is, whether the mark is a straight line or if it is wavy, and to what degree the mark is wavy. Many marks have small ticks at the head or tail. Analyzing the ends of the marks and how they were started and ended could further differentiate marks from each other. We would like to create a feature that looks at how these are formed, their length, width, and scale based on the size of the mark. We could look at how far away the mark is from the printed text. There were a number of instances when multiple marks were placed next to each other to form groups. These groupings could be used as a baseline of known marks created by the same author. Information could be generated on how these marks differ from each other as well as their spacing between each mark within the group. Many marks were created with an initial curve that proceeds into a straight line, analysis could be performed on this curved section. We could look at its proportion to the full size of the mark, angle of the arc created, and other features that make up this curve. We could look at how the pixel intensity of the mark is distributed along a mark's width, looking for patterns, whether the left or right side of a mark is predominantly different in intensity than the other side. A closer look at how the brackets were created and the variation within a bracket could reveal further information.

It could be advantageous to add the other types of marks as data points. Connecting the horizontal mark fragments to their wholes would give more information about how the author creates marks. We could also look into the symbols that were created, how they were placed and where they were placed. The last type of mark, multi-stroke marks, could be looked at in how they are formed and their inherent features.

REFERENCES

[1] ISO/IEC 13660:2001(E). Information technology - office equipment - measurement of image quality attributes for hardcopy output - binary monochrome text and graphic images. Standard, International Organization for Standardization, March 2000.

[2] Gregory R. Ball, Sargur N. Srihari, and Roger Stritmatter. Writer verification of historical documents among cohort writers. *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*, pages 314 – 319, November 2010.

[3] Sung-Hyuk Cha and Sargur N. Srihari. Assessing the authorship confidense of handwritten items. *Applications of Computer Vision, 2000, Fifth IEEE Workshop on*, pages 42 – 47, December 2000.

[4] David L. Davies and Donald W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227, April 1974.

[5] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Pearson Prentice Hall, third edition edition, 2009.

[6] R. A. Huber and A. M. Headrick. *Handwriting Identification: Facts and fundamentals*. CRC Press, 1999.

[7] Laurence Likforman-Sulem and Elisa Barney Smith. *Reconnaissance des Formes, Théorie et Pratique sous MATLAB*. Ellipses Marketing, 2013. In French. Trans. by Elisa Barney Smith.

[8] Mohammed Lutf, Xinge You, and Hong Li. Offline arabic handwriting identification using language diacritics. pages 1912–1915, 2010.

[9] Herman Melville. Melville's marginalia in dante's the vision. *Melville's Marginalia Online*, 2016.

[10] Herman Melville. Melville's marginalia in shakespeare's dramatic works. *Melville's Marginalia Online*, 2016.

[11] Steven Olsen-Smith. Herman melville's copy of thomas beale's the natural history of the sperm whale and the composition of moby-dick. *Harvard Library Bulletin*, 21(3):1–77, Fall 2010.

[12] Steven Olsen-Smith, Peter Norberg, and Dennis C. Marnon. Documentary note on melville's marginalia in dante's the vision. *Melville's Marginalia Online*.

[13] Steven Olsen-Smith, Peter Norberg, and Dennis C. Marnon. Melville's marginalia online. *http://melvillesmarginalia.org/m.php?p=policies*, 2016.

[14] Steven Olsen-Smith and Joshua Preminger. Newly deciphered erased and faded inscriptions in melville's copy of the commedia. *Leviathan: A Journal of Melville Studies*, 17(2):41–58, June 2015.

[15] Albert Sherman Osborn. *Questioned Document*. Boyd Printing, 2 edition, 1929.

[16] Antonio Parziale, Adolfo Santoro, and Angelo Marcelli. Writer verification in forensic handwriting examination: a pilot study. pages 447–452, 2016.

[17] Peter J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics*, 1987.

[18] S. N. Srihari, S.-H. Cha, H. A., and S. Lee. Individuality of handwriting. *Journal Forensic Science*, 10(3):1–12, September 2002.

[19] Jun Tan, Jian-Huang Lai, Chang-Dong Wang, and Ming-Shuai Feng. Off-line chinese handwriting identification based on stroke shape and structure. 2010.

[20] Guanglei Xiong. Local adaptive thresholding. *http://www.mathworks.com/matlabcentral/fileexchange/8647-local-adaptive-thresholding*, 2006. Accessed: 2016.