

# Determining Writership of Historical Manuscripts using Computational Methods

Sargur N. Srihari  
CEDAR, University at Buffalo, The State University of New York  
Amherst, NY 14228, USA  
*srihari@buffalo.edu*

## Abstract

*The role of computational methods in the determination of authorship of historical manuscripts is considered. During the last few years the computational forensics community has developed automation tools for forensic document examination, in particular for determining whether a given handwriting specimen can be attributed to known writing. We describe how these methods can be used with historical manuscripts where authorship is obscure or disputed. As an example the task of determining writership of a manuscript known as the Hydrachos manuscript (the questioned document) is considered. Testing the hypothesis whether it is attributable to an important American writer (Herman Melville) is described. The questioned document is compared against handwriting samples of Herman Melville, a 19th century American author who has been hypothesized to be the writer as well as against samples crafted by several writers from the same time period. The comparison led to a high confidence result to the questioned document's writership, as well as gives evidence for the validity of the writer verification method in the context of historical documents. Such methodology can be applied to many such questioned historical documents, both in literary and legal fields.*

## 1. Introduction

Literary historians often study manuscripts whose authorships are either anonymous or disputed. In many of these cases, stylometry [6, 12]— which is the application of linguistic analysis— has been used. Modern stylometry draws heavily upon the use of computational methods of statistical analysis and large amounts of machine-readable corpora that are available.

The application of handwriting analysis in the humanities has been much more limited than stylometrics. They have been used in a limited number of still unresolved questions such as the alleged appearance of William Shakespeare's hand in the manuscript of Sir Thomas More (see Howard-Hill [9] for a survey).

In cases where both known and unknown samples are handwritten, and contain sufficient sample size to generate reproducible results, forensic handwriting analysis may provide a more reliable standard for identifying authorship, for literary as well as legal documents. Over the last century forensic document science has developed progressively and more recently versatile computational methodologies for ascertaining the authorship of disputed documents have been developed [14].

Here we summarize the standard and computational procedures determining writership and give as an example the task of determining writership of a historical manuscript of American writers [3, 2].

## 2 Human Procedure

Methods to examine handwritten items have been described in numerous books, guidelines and standards. For instance, the requirements, before handwriting comparison can be undertaken, are summarized as *CAT*: (i) known exemplars are *comparable* to the disputed text, (ii) they are *adequate* in amount and (iii) they are *timely* or contemporaneous.

### 2.1 Questioned Document (QD) terminology

While human procedures have a long history, there have been many computer algorithms and software have been developed over the last few years. The vocabulary of pattern recognition is quite different from that of human examiners since they have developed

almost independently. Thus the two distinct terminologies need to be integrated for a unified solution. We first describe some of the vocabulary.

*absent character*: present in one body of writing and not the other

*character*: language symbol: letter, numeral, punctuation

*characteristic*: a feature, quality, attribute or property

*class characteristics*: properties common to a group

*comparable*: same types, also contemporaneous, instruments

*distorted*: unnatural: disguise, simulation, involuntary

*handwritten item*; cursive, hand-print or signatures

*individualizing characteristics*: unique to individual

*item*: object or material on which observations are made

*known (K)*: of established origin in matter investigated.

*natural writing*: without attempt to control/alter execution

*questioned (Q)*: source of question, e.g., common with *K*

*range of variation*: deviations within a writer's repetitions

*significant difference*: individualizing charac. outside range

*significant similarity*: common individualizing characteristic

*sufficient quantity*: volume required to assess writers' range

*type of writing*: hand-print, cursive, numerals, signatures

*variation*: deviations introduced by internal (illness, medication) and external (writing conditions, instrument)

## 2.2 The Standard Procedure

The standard (human expert) procedure for examining handwritten items can be summarized as follows.

1. Determine if comparison is *Q v. Q*, *K v. K*, or *Q v. K*. The first when there are no suspects or to determine number of writers. The second to determine variation range. The third to confirm/repudiate writership.
2. Determine whether *Q* and *K* are original or copies. If not original, evaluate quality of best reproduction and check whether significant details are reproduced with sufficient clarity. If not discontinue procedure.
3. Determine whether *Q* and *K* are distorted.
4. Determine the type of writing. If more than one, separate into groups of single type.
5. Check for *internal inconsistencies* in groups. If inconsistencies suggest multiple writers, divide groups into consistent subgroups. For *K*, if there are unresolved inconsistencies, stop procedure and report accordingly.
6. Determine range of variation for each group/subgroup.
7. Detect presence/absence of *individualizing characteristics*.
8. Evaluate comparability of *Q* and *K*. If not comparable request new *K* and repeat.
9. Compare bodies of writing.
10. Compare and analyze differences and similarities to form conclusion. The recommended US terminology for expressing the conclusion is a nine-point scale: 1-Identified as same, 2-Highly probable same, 3- Probably did, 4-Indications did, 5- No conclusion, 6- Indications did not, 7- Probably did not, 8- Highly probable did not, 9-Identified as Elimination.

The standard procedure is a general guideline. In following the steps, the examiner needs to make several decisions, since every case has special needs. A *ransom note* could be written by multiple writers thus requiring comparison of document sub-parts, in which case each sub-part becomes a *Q* or a *K*. In a *historical manuscript* different writers may be more similar to each other than with contemporary writers thus requiring recalibration of individualizing characteristics [2].

Some details within the procedure are left to the experience of the examiner. For example, handwriting characteristics are determined based on years of training and power of recall of the examiner, although general guidelines are stated, e.g., the seven S's, viz., *size, slant, spacing, shading, system, speed, strokes*. Adequacy of the amount of handwritten material for a comparison is also based on experience.

In converting the standard procedure to a computational one, probabilistic analysis can be useful in articulating steps based on experience. In Step 7, the examiner selects individualizing characteristics. Such characteristics can be stated as those that are rare within the population, i.e., the letter/word formations have a low probability of occurrence. In Step 8, adequacy of material can be managed by using confidence intervals. In Step 10, the opinion can be formed by discretizing a probability of identification/exclusion.

### 3 The Computational Procedure

Some of the knowledge engineering necessary for the computational approach to FDE of handwritten items is already available as described in Section 2. We next summarize the vocabulary of the relevant pattern recognition literature which is also used in the algorithmic formulation of the standard procedure that follows.

#### 3.1 Pattern Recognition (PR) terminology

The following terms are commonly used in the pattern recognition community which is concerned with developing computerized methods.

*bigram*: a pair of characters, usually common ones, e.g., *th*

*cropping*: process of specifying boundary of a region

*distance*: measure of difference between two feature sets, reciprocal of similarity measure

*features*: characteristics of writing, e.g. macro (spacing, slant, etc), micro (character or bigram shape)

*image processing*: enhancement processes, e.g., noise removal, thresholding

*likelihood ratio (LR)*: ratio of probability of Q and K under hypotheses that they had same/different writership

*log-likelihood ratio (LLR)*: logarithm of the LR

*region of interest (ROI)*: region of document to be compared

*resolution*: number of pixels per inch, typically 300

*scanning*: conversion of item to digital image by specifying resolution and number of gray levels

*transcript mapping*: automatically associating text in a transcript with each word image

*truthing*: associating image of character/word with text

*word segmentation*: process of separating images of words

#### 3.2 Algorithm for Comparison of Manuscripts

The algorithmic formulation of the standard procedure is given in the pseudocode of Algorithm 1. The algorithm has four phases: determining the type of comparison, determining properties of each of the documents to be compared, determining handwriting characteristics for comparison and performing the comparison. In this section we describe implementations of the four phases: initialization, document analysis and recognition, determining handwriting characteristics and document comparison. Some of the implementations may be performed manually and others by using computer algorithms. Several of the steps can be effectively implemented using probabilistic formulations.

Initialization involves scanning to convert paper to digital form, determining the regions of writing to be compared and the type of comparison. In the document analysis and recognition phase each document is subjected to a variety of operations including noise removal and text recognition. In the next phase handwriting characteristics, particularly those useful for individualization, are determined. In the comparison phase document characteristics are compared.

**Scanning:** Here we are assuming that the handwriting is first represented in digital form by scanning handwriting on paper. Issues such as ink type and indentation on paper would have to be addressed as other types of evidence which could be combined with results of handwriting comparison.

---

**Algorithm 1** Comparison of Handwritten Manuscripts

---

```
1: Phase 1: Initialization
2:   Scan Documents
3:   Determine Comparison Type:
4:      $Q \vee Q$  (no suspect or determine no. of writers)
5:      $K \vee K$  (to determine variation range)
6:      $K \vee Q$  (to determine/repudiate writership)
7:   Determine Region of Interest
8: Phase 2: Document Analysis and Recognition
9: for each Q or K do
10:  Quality: determine whether the writing is of sufficiently high quality for comparison.
11:  Distortion: determine whether the writing is natural or distorted.
12:  Writing type: determine whether the writing is hand-printed or cursive.
13:  Segmentation and Recognition: segment and recognize individual characters and words of text.
14:  Internal consistency: whether writing within document is consistent or indicates multiple writers.
15:  Determine range of variation: compare parts of document to determine range.
16: end for
17: Phase 3: Handwriting Characteristics
18:  Identify individualizing characteristics: those that are unusual.
19: Phase 4: Handwriting Comparison
20: for each Comparison do
21:  Comparability: whether both handwritten items are of same type (Step 12).
22:  Adequacy: whether sufficient handwriting is present.
23:  Comparison: compare characteristics of Q and K to determine similarities and differences.
24:  Form Opinion: decision of identification/exclusion/no opinion using scale of opinions.
25: end for
```

---

First  $Q$  and  $K$  documents are scanned and digitized. Scanning resolution is one of the first issues to be addressed. Commonly used scanning resolution is 300 dpi. While systems are designed to operate at this resolution, to deal with higher resolution scanned images, they are internally converted to 300 dpi images before processing.

**Comparison Type:** In the first step of the Algorithm 1, problem type is determined by problem specification.  $Q$  and  $K$  need not be entire documents but image regions selected by cropping and removing extraneous information; with care so as to not affect writing where text overlaps seals and paper creases. Document level comparison can be refined by paragraph and word level results.

### 3.2.1 Document Analysis and Recognition

**Noise:** Images may be noisy (contain salt/pepper noise) or have unwanted elements, e.g., stamps and seals. Noise removal is a common problem in image processing. Several interactive tools are available for preparing the document for processing: the region of interest (ROI) can be isolated by cropping, noise reduction to remove speckles, adaptive thresholding to extract writing from background [11], eraser to remove unwanted artifacts, and automatic rule line removal.

**Text Line Segmentation:** Textual writing is necessarily sequential. Determining lines and words of text in the image is one of the first analysis tasks. This can be done automatically, using one of several methods – a very effective method uses statistical models to estimate line structure [1]. Since the results of automatic segmentation are not always perfect, the results can be corrected manually, e.g., by using a polygon or lasso tool.

**Recognition:** Prior to comparison of documents it is necessary to recognize characters and words of text. Domain-independent word recognition is still error prone but many software tools exist. A better approach is transcript mapping where text is associated with word images (truth) by automatic transcript mapping is done by providing a plain text transcript [16]. The slowest, but least

error prone method is manual truthing of hand-segmented words. Also errors in automatic recognition or transcript mapping can be manually corrected. Writing elements (characters and words) are provided with truth provided by automatic/manual recognition.

**Quality and Distortion:** Quality can be determined either visually or by evaluating results of automatic processing, e.g., evaluation of recognition results. Distortion can be determined using rarity measures.

**Internal Consistency:** Within document sub-parts, presence of multiple authors can be performed by repeatedly applying the comparison step as described below. Range of variations for a given writer can be accomplished by analyzing the range of *LLR* values for the same writer, with negative values indicating wide variability.

### 3.2.2 Handwriting Characteristics

Characteristics of handwriting are used to determine writership. Characteristics useful for determining writership can be quite different from those used for handwriting recognition, particularly because recognition involves determining the average way of writing whereas writer identification involves determining deviations from the norm. Such characteristics can either be determined automatically or manually and used in subsequent automated processes. Since several characteristics are subtle, e.g., detecting a retrace, manually determined characteristics have an advantage, at the expense of speed. The main difference between the results of the manual and computer approach lies in the choice of characteristics.

There are several methods for computing handwriting characteristics automatically. They are referred to as features in the pattern recognition literature. As one example, for each  $Q$  or  $K$ , a set of macro or global features are computed [15]; they refer to global attributes such as slant, spacing, thickness, pixel distribution, etc, which can be turned on/off individually. Micro features are computed for characters; these are a set of 512 bits known as GSC features [16]. Similar features are computed for bigrams, and words; for words a 1024 bit string is used. Two additional features are *th* features and lexeme features. The special treatment of *th* is because it is the most common bigram in English and also has the highest discriminating power among letters and bigrams; its features are computed by first skeletonizing the image and obtaining a small set of descriptive features.

### 3.2.3 Handwriting Comparison

Several computational methods for the comparison of forensic samples have been developed over the last two decades. The probability of the samples originating from the same individual can be determined by knowing the likelihood ratio whose numerator is the probability of the observed characteristics under the "same-individual" hypothesis and the denominator is the probability under the "different individuals" hypothesis [18].

**Distance Measure:** The distributions of dissimilarities conditioned on being from the same or different writer are used to compute likelihood functions for a given pair of samples. Such distributions can be learned from a training data set and represented either as probability tables or as probability density functions. Probability distributions for discrete valued distances are represented as probability tables. In the case of continuous-valued distances parametric methods are used to represent probability density functions. Parametric methods are more compact representations than nonparametric methods. For example, with the k-nearest-neighbor algorithm, a well-known nonparametric method, we need to store some or all training samples. This compactness results in a faster run-time algorithm. Also parametric methods are more robust in a sense that nonparametric methods are more likely to over-fit and therefore generalize less accurately. Of course the challenge is to find the right parametric model for the problem at hand. While the Gaussian density is appropriate for mean distance values that are high, the Gamma density is better for modeling distances since distance is never negative-valued.

**Independence Assumption:** When each document is characterized by more than one feature one can make the assumption that the writing elements or features are statistically independent, although strictly speaking this is incorrect. Complex models, such as a Bayesian neural network, give an improvement of 1-2% on overall accuracy, which is not significant. There are several other justifications for choosing naïve Bayes. First, its functioning is simple to explain and modify, e.g., since log-likelihood ratios are additive we can observe the effects of adding and removing features easily. Its simplicity goes back to the earliest QDE literature [1] where there is reference to multiplying the probabilities of handwriting elements. Second, as has been observed in other machine

learning tasks, more complex models tend to overfit to the data, which can lead to poorer performance on large amounts of unseen data.

Each of the two likelihoods that the given pair of documents were either written by the same individual or by different individuals can be expressed, assuming statistical independence of the features as follows. For each writing element  $e_i, i = 1, \dots, c$ , where  $c$  is the number of writing elements considered, we compute the distance  $d_i(j, k)$  between the  $j$ th occurrence of  $e_i$  in the first document and the  $k$ th occurrence of  $e_i$  in the second document for that writing element. The likelihoods are extracted as

$$L_s = \prod_{i=1}^c \prod_j \prod_k p_s(d_j(j, k))$$

$$L_d = \prod_{i=1}^c \prod_j \prod_k p_d(d_i(i, k))$$

The LLR in this case has the form

$$LLR = \sum_{i=1}^c \sum_j \sum_k [\ln p_s(d_i(j, k)) - \ln p_d(d_i(j, k))]$$

That is, the final *LLR* value is computed using all the features, considering each occurrence of each element in each document. When the likelihood ratio (*LR*) is above 1, the *LLR* value is positive and when the likelihood ratio is below 1, the *LLR* value is negative. Hence, if the final score obtained is positive, the system concludes that the two documents were written by the same writer. Similarly, if the final *LLR* score is negative, the system concludes that the two documents were written by different writers.

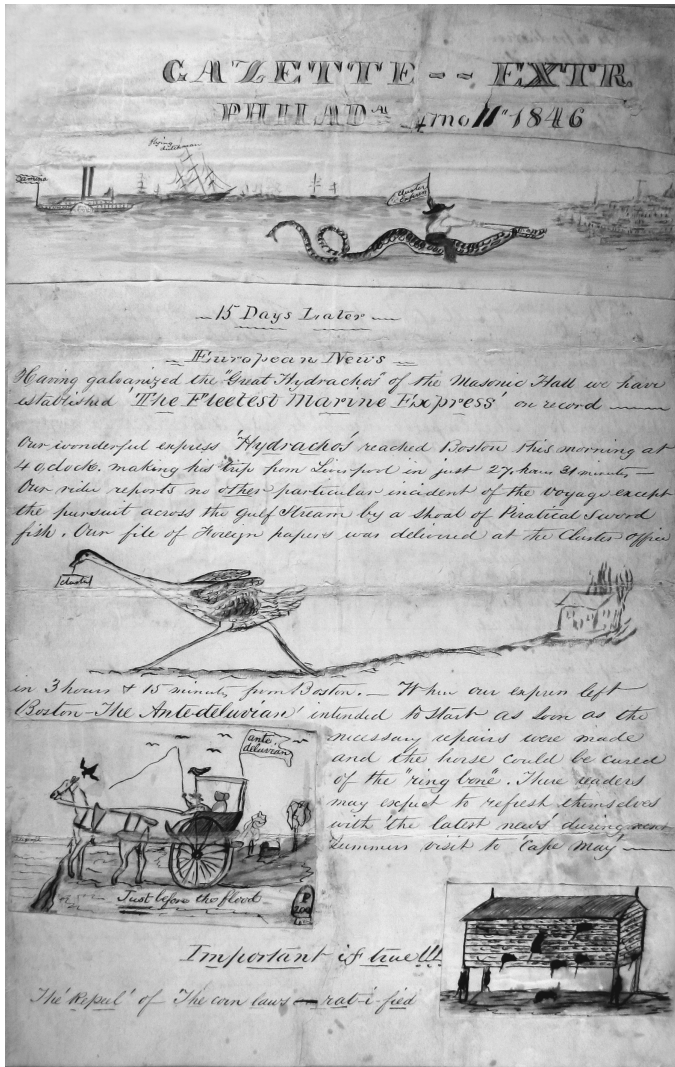
#### 4 The Hydrachos Manuscript and Herman Melville

A case for determining authorship of a historical manuscript is the Hydrachos manuscript (H.)<sup>1</sup>. It is an April 1846 satirical newspaper, The PHILADAL GAZETTE - EXTR<A>, with seven pen and ink drawings accompanying a 437 word handwritten commentary on U.S. and world news. “Hydrachos” is the document’s misspelling of the name of a notorious paleontological curiosity of the 1840s, the Hydrarchos Sillimani, also called *Basilosaurus* and eventually renamed *Zeuglodon cetoides*. As summarized in the words of a contemporary novelist it was the “skeleton of an extinct monster, found in the year 1842, on the plantation of Judge Creagh, in Alabama.”[10]. The Alabama doctors declared it a huge reptile, and bestowed upon it the name of Basilosaurus. But some specimen bones of it being taken across the sea to Owen, the English Anatomist, it turned out that this alleged reptile was a whale, though of a departed species. In 1846 the remains of the Hydrarchos, having previously appeared in 1845 in New York at the Apollo Saloon, were on display at the Philadelphia Natural Museum housed in the Masonic Hall on Chestnut Street[13, 115] which is so conspicuously alluded to in the H. document’s opening lines.

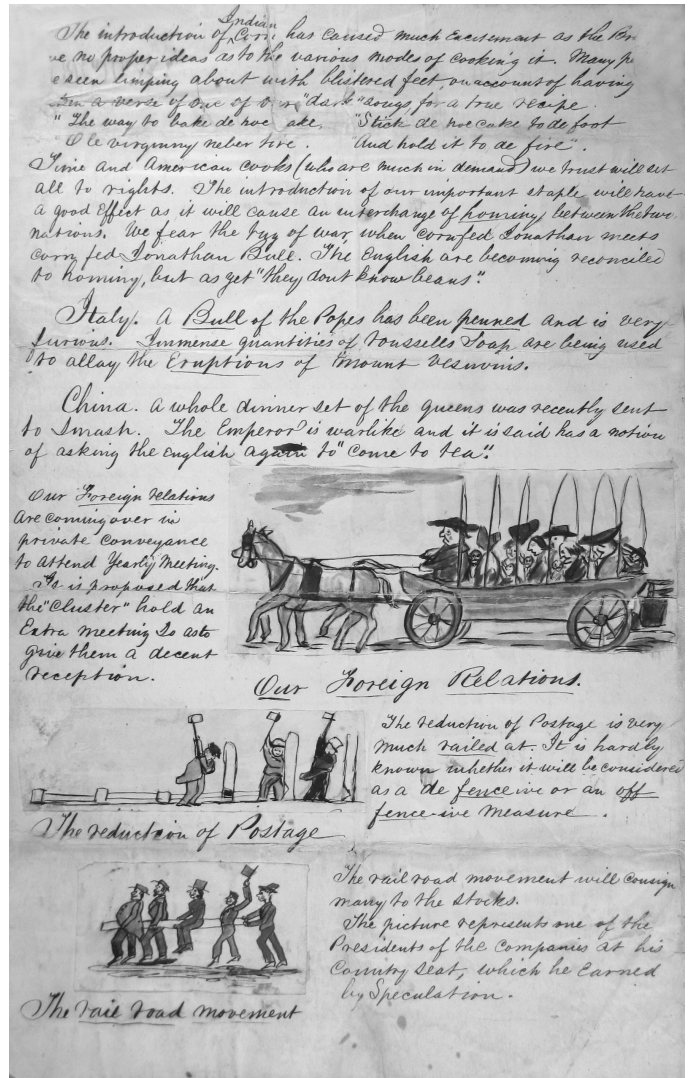
Measuring 40 x 25 cm, the document contains satirical news content, primarily from the United States, Great Britain, Italy, and China, on both the recto (Figure 1(b)) and verso sides (Figure 1(a)).

Lexical, grammatical, thematic, visual, content, and situational analysis[17] all support the hypothesis that the manuscript’s author is the New England author Herman Melville whose description of the controversy surrounding the “extinct monster” unearthed in Alabama from Moby Dick (Chapter 104, “The Fossil Whale”). This synoptic report, drawing a more complete unpublished analysis by Stritmatter [17], summarizes some reasons for hypothesizing this attribution, surprising as it might seem to contemporary Melville scholars. Although Melville does not use the word Hydra<r>chos, many other phrases and allusions in the manuscript can be traced to his published writings. A striking example is the document’s leading conceit of the sea monster capable of carrying mail between America and England. Melville’s fall 1847 satirical squib to Yankee Doodle, appearing in print eighteen months after the date on the Hydrachos manuscript, similarly views the sea monster not as an extinct pile of bones but as a living asset to the communications industry. The satire offers a reward of one thousand dollars to anyone able to “procure a private interview with the Sea-serpent, of Nahant notoriety,” for the purpose of concluding a negotiation with the Postmaster General to license the beast “for the transmission of European mails from Boston to Halifax” (Hayford 429). The satire also advertises for a “smart jockey” to “superintend” the shipment. The situation of Melville’s satire directly echoes the Hydrachos visual depiction of a sea monster mounted by a “rider,”

<sup>1</sup>The discussion on the Hydrachos manuscript and observed similarities to Herman Melville’s writing is due to Roger Stritmatter.



(a) Hydrachos manuscript, verso



(b) Hydrachos manuscript, recto

Figure 1: Questioned document—the Hydrachos manuscript

equipped like a horse jockey with a bridle and a riding hat, ferrying mail between Boston and Liverpool. Like so much else in Melville's writing, the motif of the sea monster transporting international mail from the Eastern United States to Europe can be directly traced to Melville's own circumstances. In April 1846 he was engaged in an intense transatlantic correspondence with his brother Gansevoort, the secretary to the American legation in London, who was concluding negotiations for the publication of Herman's first book, *Typee*. The book was published in England by John Murray and in the United States by Wiley & Putnam in March, 1846; throughout March Gansevoort was sending Herman British newspapers containing *Typee* reviews. The H. document, which has been folded five times vertically and once horizontally, to form an envelope-sized packet, 9x12 cm, preserves traces of evidence that it was at one time sent through the mail as a part of such a correspondence. More specifically, the April 11 date is of particular interest given that Gansevoort sent to Herman by the March 18 "Unicorn" a number of papers, "principally Examiners & Critics contg notices of Herman's *Marquesas Islands*" (Parker, "London Journal," 53). It is proposed that this is the shipment alluded to in the H. manuscripts statement that "our file of foreign papers was delivered at the cluster office" - the latter perhaps referring to Alan Melville's Wall Street law office in Manhattan, where international correspondence for the family was typically routed. Qualitative stylistic analysis supports the attribution. In the 437 word document, Stritmatter[17] was able to trace 59 words and phrases many of an apparently particularistic nature, found in the Melville canon, including:

- Flying Dutchman
- galvanized
- Masonic
- Antidiluvian (sic);
- telegraph
- the latest news
- Indian corn
- much excitement
- modes of
- hoe cake (used in an identical context)
- in demand
- set all to rights
- the introduction of
- John Bull and Jonathan (as sobriquets for England and America)
- become reconciled
- don't know
- eruptions
- Vesuvius
- China
- emperor
- warlike
- a notion; conveyance
- reception



- railed at.

Analysis of diction and content is also consistent with Melville’s authorship.

Melville’s penchant for topical satire and humorous drawings is well known, and the surviving direct evidence for his artistic style strong supports the attribution of the H. document. (Figures Four and Five). Although Melville’s typical handwriting, as known through his journals and his few extant literary manuscripts[8, 7, 4] is markedly less elegant than the Hydrachos hand, which manifests the writer’s graphic intent to produce an artistic document, survey of readily available samples yielded a number of suggestive qualitative comparisons supporting the attribution and suggesting that the observed differences constitute “natural” rather than “systematic” variation[5]. Curators of the NYPL Gansevoort collection subsequently supplied samples of two early letters written by the young Melville in a more careful hand (Davis Gilman 2 and 10; Horth, 7-10 27-30) for systematic quantitative comparison. A section of one of these (known) documents is shown in Figure 4.

In addition, while the utility and accuracy of the writer verification system used in CEDAR-FOX is well-established on a large data set, it has not been tested extensively on documents of historic origin. To help establish that the system is era-independent, or at least appropriate for comparison to Melville’s time period, letters written by two of Melville’s male relatives (Thomas Melville, Jr. and Alan Melville) were obtained, both produced in 1818.

## 4.1 Results of Comparison

To evaluate the documents according to the writer verification method described in the previous section, we followed the following sequence of steps to prepare the documents for analysis:

Step 1: Manual preprocessing: We modified the documents to remove the hand-drawn non-text images, and remove the more major noise.

Step 2: Pre-processing with CEDAR-FOX: Binarization, line and word segmentation, document wide feature calculation, automatic character recognition.

Step 3: Word segmentation correction: Some manual correction to the word segmentation was performed since there were some slight word segmentation errors.

Step 4: Transcript mapping, manual transcript corrections: Using the “transcript mapping” function and correct transcripts, the known and unknown documents were manually verified to be correctly word truthed.

Some of the processed documents used are shown in Figure 3.

## 4.2 Cohort Comparison Results

We treated the problem as a case of four known (letters) and two unknown (verso and recto) documents and generated an LLR comparison value for each of the pairs of documents. Comparisons of some select bigram features are shown in Figure 5.

The comparison method was validated on a test set of 1,648 test cases[16]. When performing on the test cases using all features on documents containing different content, the overall error rate for the general population was 0-2.4%. These test documents were also used to map the LLR score onto a nine point opinion scale, ranging in order as follows: identified as same, highly probable same, probably did, indications did, no conclusion, indications did not, probably did not, highly probable did not, identified as different. The overall LLR and opinion results for the documents are described in Table 1.

The LLR scores are compound scores consisting of several feature components: the global (macro), character (micro), bigram, word, and lexeme features.

The characters were automatically extracted based upon the word truth. All words were split into candidate characters and the automatically generated characters were validated by a character recognizer. Characters failing the automatic verification check were not used in the comparison so as to not introduce artifacts into the comparisons.

Each acceptable character was compared exhaustively to each acceptable corresponding character in the comparison document. When all LLR’s of all characters are combined, overall trends are observed. In this case, all characters led to positive results, adding information suggesting same authorship.

To more characterize the strength of confidence in the LLR scores, we modeled the 1,648 validation cases and evaluated which cases had higher or lower LLR scores – that is, we tabulated the LLR values of all known same and different writer cases (with full page documents consisting of different content) and determined the percentage of such validation cases that had higher or lower LLR values than our current study’s results. The results are presented in Table 2.

In short, none of the validation cases consisting of full page documents written by different writers with different content had LLR values that were higher than the H. Melville letters vs. Hydrachos sets presented, implying that the documents we compared are more

Pittsfield Dec 31st 1837

My Dear Uncle

At my departure from Albany last fall with Robert you expressed a desire that I should write you when my school should have gone into operation — but when in a few weeks it again returned, you did not repeat your request. Still, however, I considered my promise binding — & it is with pleasure that I now proceed to redeem it.

I should have taken up my pen at an earlier day had not the variety & multiplicity of the duties incident to my vocation been so numerous and pressing, that they absorbed a large portion of my time.

But now, having become somewhat acquainted with the routine of business, — having established a system in my mode of instruction, — and being familiar with the characters & dispositions of my scholars: in short, having brought my school under a proper organization — a few intervals of time are afforded me while I improve by occasional writing & reading.

My scholars are about thirty in number, of all ages, sizes, ranks, characters, & conditions; some of them who have attained the ages of eighteen can set up a sum in arithm, while others have paralleled through the Arithmetic: but with so great swiftness that they cannot recognize objects in the road on a second's journey; & are about as ignorant of them as though they had never passed that way before.

My school is situated in a remote & secluded part of the town about five miles from the village, and the house at which I am

(over)

(a) H. Melville Letter 1

New York, Jan 4 20th 1845

My Dear Sister: — What a charming name is yours — the most engaging I think in our whole family circle. — I don't know how it is precisely, but I have always been very partial to the particular appellation & can not avoid investing the person who bears it, with certain quite captivating attributes; so, when I hear of Kate Pook or one — whether it be Kate Smith or Kate Jones, or Kate My Body Else I invariably impute to the said Kate all manner of delightful characteristics. — Now, that terms of general admiration will do at all, when applied to the Clan-Kate — for the Kates, I you see, are a speculative race, & are distinguished by speculative attributes — Thus, the Kates as a general rule are decidedly handsome, but if we may not speak of their beauty in terms of magnificent admiration, they still will be found to incline towards good looks, and at any rate, they are never positively ugly. — But, "Some feathers don't make fine birds" & "Handsome is, that handsome does" & all that sort of thing; — & so if the Kates were only

(b) H. Melville Letter 2

Figure 2: Known Writings of Herman Melville.



Figure 3: Pre-processed and truthed documents: (a) known writing of H. Melville, and (b) Hydrachos verso manuscript.

Table 1: Results of comparisons, where the overall log-likelihood ratios are given.

	Total LLR	System Opinion
H. Melville Letter 1 vs. Hydrachos, verso	35.29	Identified As Same
H. Melville Letter 2 vs. Hydrachos, verso	129.42	Identified As Same
H. Melville Letter 1 vs. Hydrachos, recto	20.82	Highly Probable Same
H. Melville Letter 2 vs. Hydrachos, recto	60.12	Identified As Same
H. Melville Letter 1 vs. H. Melville Letter 2	333.52	Identified As Same
Hydrachos manuscript, verso vs. recto	42.87	Identified As Same
Full Hydrachos manuscript vs. H. Melville Letter 1	56.11	Identified As Same
Full Hydrachos manuscript vs. H. Melville Letter 2	189.54	Identified As Same
T. Melville Letter vs. H. Melville Letter 1	-14.84	Indications Did Not
T. Melville Letter vs. H. Melville Letter 2	12.38	Indications Did
T. Melville Letter vs. Hydrachos, recto	-68.18	Identified As Different
T. Melville Letter vs. Hydrachos, verso	-73.03	Identified As Different
A. Melville Letter vs. H. Melville Letter 1	-671.41	Identified As Different
A. Melville Letter vs. H. Melville Letter 2	-386.09	Identified As Different
A. Melville Letter vs. Hydrachos, recto	-318.96	Identified As Different
A. Melville Letter vs. Hydrachos, verso	-448.61	Identified As Different
A. Melville Letter vs. T. Melville Letter	-284.46	Identified As Different

similar than any of the validation pairings. In the case of Herman Melville Known Letter 2 vs. the average of the verso and recto sides of the questioned document, the LLR score implies similarity stronger than all but about 2% of the validation cases.

When comparing the T. Melville, Jr. letters to the Hydrachos manuscript, the system clearly identified the writers as different. When comparing the T. Melville, Jr. letters to the Herman Melville known letters, the results were essentially non-decisive, resulting in very low LLR scores when considering the large amount of text compared. One of the two scores was (incorrectly) slightly positive. When examining the incorrect result in more detail, one major factor for the score was that the letters being compared had very few

Dear Sir Boston May 30<sup>th</sup> 1818

I have this day put on board  
 the Sloop George Washington for Albany  
 one half bushel of Wine containing  
 5th Gallons. As it is below it is said  
 by our best Judges to be very good, I hope  
 that it will please Mr. Janswood, I  
 have also sent to your care two small  
 boxes add for my son at Pittsford,  
 At Pittsford will sail for later day, near  
 for Francis Mose to inform Maria  
 she will direct her letter under cover  
 to me, and I will see that it goes  
 I shall be happy at all times to render  
 you or any the family any service in  
 my power we are all in our usual  
 health, and are joined in Love to Maria  
 and affectionate remembrance to all  
 and yourself -

Yours friend  
 Thos. Melville

5th Gallons @ \$1.50 = \$7.50  
 Package = .13  
 Total = \$7.63  
 Shall wait by the Post

Peter Janswood Esq.  
 Albany -

(a) A. Melville Letter

Pittsford May 2<sup>nd</sup> 1818.

We are quite well, and thank the Lord for  
 my kind letter & which helped to make my  
 mind at ease - to hear that you had not yet  
 shown the letter of the good old lady, is the  
 more painful this year, a want of it, but I do not  
 forget - whether I shall find it, I do not  
 know, but I shall certainly not fail; in my  
 sister's, she says you for, had, with in the ground  
 excepted you in it, and in her - I do not know  
 she has the letter - and then her name, in the letter to you  
 and - this is on the letter, - and the letter, we  
 brought in of them, all the letters, which he has  
 the letter and has been all the former in Pittsford  
 right kind of letter, I am indeed to wish the favor  
 of you, if possible, to obtain from Mr. York, the favor of your  
 time, going by the stage Post, on Monday, the  
 and on Monday you, and on Monday will be glad  
 to see you, if it be possible, for Pittsford.

(b) T. Melville Letter

Figure 4: Writings of contemporaries of Herman Melville written in 1818: (a) Alan Melville and (b) Thomas Melville, Jr.

Figure 5: Comparison of “ea,” “ou,” “oo,” and “wh” between H. Melville’s writing and the Questioned document



Table 2: Confidence Score Breakdown

	% same writer having > LLR	% diff writer having > LLR
HM <sub>1</sub> vs. Qverso	14.32	0
HM <sub>2</sub> vs. Qverso	0	0
HM <sub>1</sub> vs. Qrecto	41.99	0
HM <sub>2</sub> vs. Qrecto	4.61	0
TM vs. HM <sub>1</sub>	0.97	74.51
TM vs. HM <sub>2</sub>	85.92	0.48
TM vs. Qrecto	0	1.21
TM vs. Qverso	0	8.73
AM vs. HM <sub>1</sub>	0	0
AM vs. HM <sub>2</sub>	0	0
AM vs. Qrecto	0	0
AM vs. Qverso	0	0
AM vs. TM	0	0

large words in common, which meant that the word comparison had to be based on very short words, which yield less strong results.

When comparing A. Melville’s letters to Herman Melville’s letters and to the Hydrachos document, the system reported very strongly that the writers were different. When comparing the T. Melville, Jr. and A. Melville letters to one another, again the system reported a strong result that they were different. In fact, in all five cases, 0% of the validation test cases were the results so strong in either the same or different writer groups.

### 4.3 Discussion

A computational procedure for the comparison of historical handwritten manuscripts has been described. Steps within the procedure are several automation tools such as preparing the documents to compare, e.g., extracting regions of interest, noise removal, determining type of writing, and comparison tools that extract features and compute likelihood ratios. These automation tools can be combined with manual procedures, e.g., using manually specified handwriting characteristics together with rarity computation. Quantitative results of comparison can be obtained using statistical methods that learn from data sets of representative handwriting.

In the comparison of the Hydrachos manuscript with the writing of Herman Melville, the LLR numbers indicated Melville likely likely penned the questioned document; in addition, comparisons between the questioned document and with known documents penned by various contemporary writers likely to have similar penmanship styles resulted in strong opinions that the documents were

written by different writers. Further comparisons between the known different writers resulted in reasonable results indicating that the documents were indeed recognized as being created by different writers, with a single exception which resulted essentially in a non-conclusive result. This helps to establish that the comparison method described returns valid results for the time period in question. Further results could be generated by comparing other known documents containing similar phrasing to the questioned document. Additional results could be gained by comparing these documents to other known documents.

## References

- [1] M. Arivazhagan and S. N. Srihari. A statistical approach to handwritten line segmentation,. In *Document Recognition and Retrieval*, volume XIV. SPIE, 2007.
- [2] G. R. Ball, S. N. Srihari, and R. Stritmatter. Writer identification of historical documents among cohort writers. In *Proc. Int. Conf. Frontiers Handwriting Recog. Kolkata, India*. IEEE Comp. Soc. Press, 2010.
- [3] G. R. Ball, R. Stittmeyer, and S. N. Srihari. Writer verification in historical documents. In *Proceedings Document Recognition and Retrieval XVII*, San Jose, CA, January 2010. SPIE.
- [4] J. Bryant. *Melville Unfolding: Sexuality, Politics, and the Versions of Typee*. University of Michigan, Anne Arbor, 2008.
- [5] W. R. Harrison. *Suspect Documents*. Fredrick A. Praeger, New York, 1958.
- [6] D. Holmes. *Journal of the Royal Statistical Society. Series A (General)*, 148(4):328–341, 1985.
- [7] H. C. Horsford and L. Horth, editors. *The Writings of Herman Melville: Journals*, volume 15. Northwestern-Newberry Edition, Chicago, 1989.
- [8] L. Horth. *The Writings of Herman Melville: Correspondence*. 14. Northwestern-Newberry, Chicago, 1993.
- [9] T. H. Howard-Hill, ed., editor. *Shakespeare and Sir Thomas More: Essays on the Play and its Shakespearean Interest*. The University Press, Cambridge, 1989.
- [10] H. Melville. *Moby Dick*. The Northwestern Newberry Library, Evanston, Ill, 1997.
- [11] P. Palumbo, P. S. Swaminathan, and S. N. Srihari. Document image binarization: Evaluation of algorithms. In *Applications of Digital Image Processing IX*, volume 697, pages 278–285. SPIE, 1986.
- [12] R. D. Peng and N. Hengartner. Quantitative analysis of literary styles. *The American Statistician*, 56(3):175–185, 2002.
- [13] D. R. Prothero and R. M. Schoch. *Horns, Tusks, and Flippers: The Evolution of Hoofed Mammals*. Johns Hopkins University Press, Baltimore, 2003.
- [14] S. Srihari and K. Singer. Role of automation in the examiation of handwritten items. *Pattern Recognition*, 2014.
- [15] S. N. Srihari, S.-H. Cha, H. Arora, and S. Lee. Individuality of handwriting. *Journal of Forensic Sciences*, 47(4), July 2002.
- [16] S. N. Srihari, C.Huang, and H. Srinivasan. On the discriminability of the handwriting of twins. *Journal of Forensic Sciences*, 53(2):430–446, 2008.
- [17] R. Stritmatter. The great hydrachos satire of April 11, 1846: A lost literary manuscript attributed to Herman Melville. June 2009.
- [18] Y. Tang and S. N. Srihari. Likelihood ratio estimation in forenisc identification using similarity and rarity. *Pattern Recognition*, 47(3):945–958, 2014.