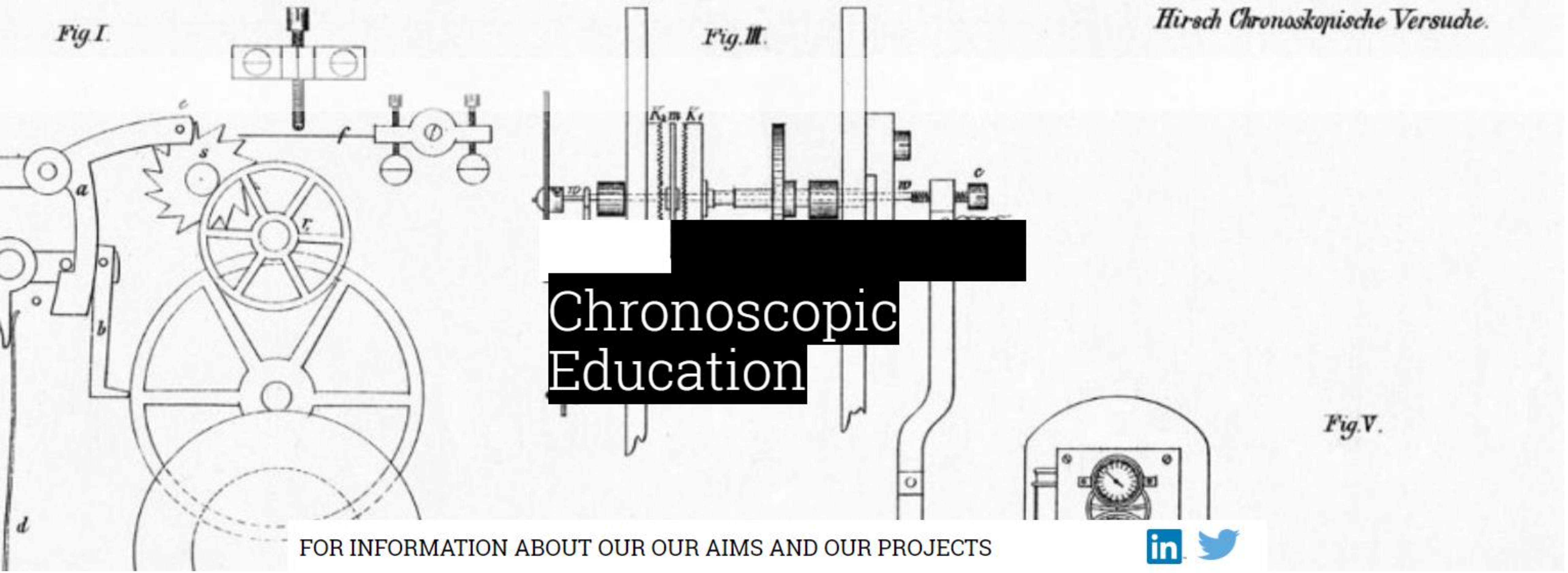
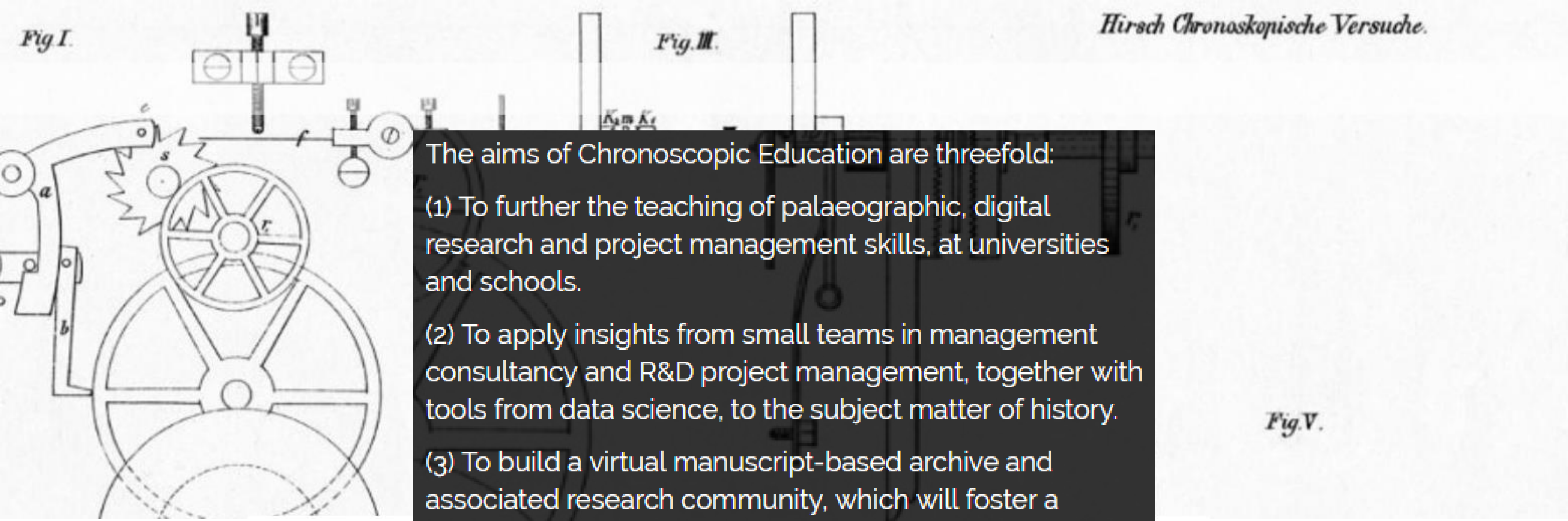


Signs of Literacy
Kaggle Research Competition

Colin Greenstreet

Thursday, June 28th, 2018

Our social aims



The aims of Chronoscopic Education are threefold:

(1) To further the teaching of palaeographic, digital research and project management skills, at universities and schools.

(2) To apply insights from small teams in management consultancy and R&D project management, together with tools from data science, to the subject matter of history.

(3) To build a virtual manuscript-based archive and associated research community, which will foster a culture of collaborative scholarship.

http://www.chronoscopic.org

Project portfolio

MarineLives

Signs of Literacy

Maphackathon

EM Textiles, Garments & Dyestuffs Glossary

EM Maritime & Mercantile Gazetteer

http://www.chronoscopic.org

Signs of Literacy Kaggle Research Competition, Nov 2018 – Jan 2019

Signs of Literacy Kaggle Research Competition, 2018
Colin Greenstreet on LinkedIn
April 30, 2018

Google owned Kaggle has selected us as one of a small number of pro bono competitions they support each year on the merits of our proposal, and the potential impact on the research field and community of the competition.

Kaggle will cover the running costs of the competition. We will provide the prize pool, and are now seeking to raise US $30,000 from potential sponsors and partners.

The Proof of Concept will contain two parts:

(1) Algorithmic identification of markes, initials and signatures.

(2) Algorithmic discrimination between degrees of "sophistication" within the three categories of "marke"; "initial(s)", and "signature".

Having proven the concept, we will seek out an image or vision oriented computational laboratory with which to develop a grant funded collaboration to take the work further in 2019 and beyond.

https://www.linkedin.com/pulse/proposed-signs-literacy-kaggle-research-competition-2018-greenstreet/

Technical vision & role of the Kaggle competition

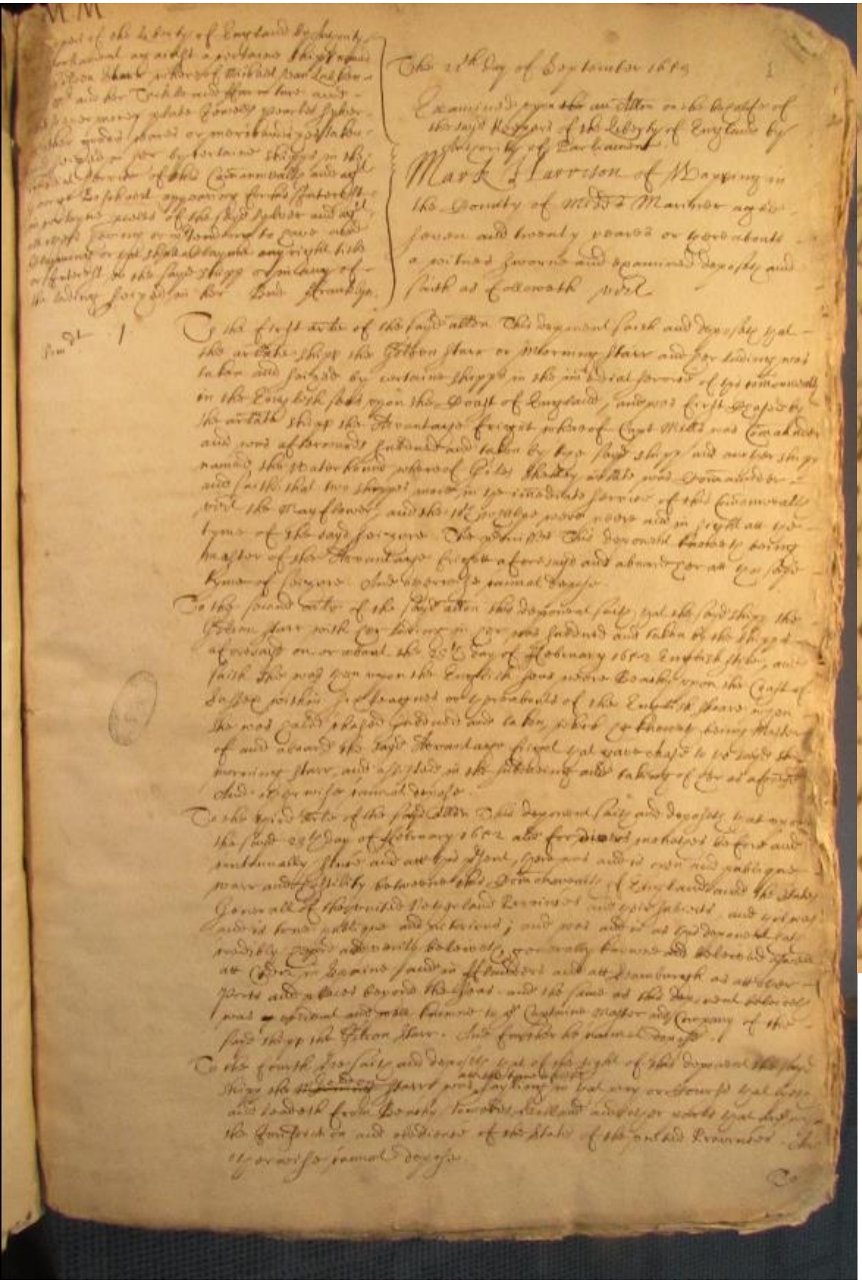
- Automatic identification of manuscript pages containing signoffs
- Markup of manuscript pages to isolate signoffs (markes, initials, signatures)
  - Hand markup of manuscript pages, but ideally automated markup
  - Signoffs can be single or multiple, for single or multiple depositions
  - Deponent signoffs; interpreter signoffs
- Automatic differentiation between classes of markes, initials and signatures
- Automatic differentiation within each class as to sophistication of execution & other parameters as a surrogate for literacy
- Automatic identification of manuscript pages containing deponent metadata (name; age; occupation; place of residence; date of deposition)
- Markup of manuscript pages to isolate deponent metadata
- Hand writing text recognition of deponent metadata and associate metadata with correct signoff

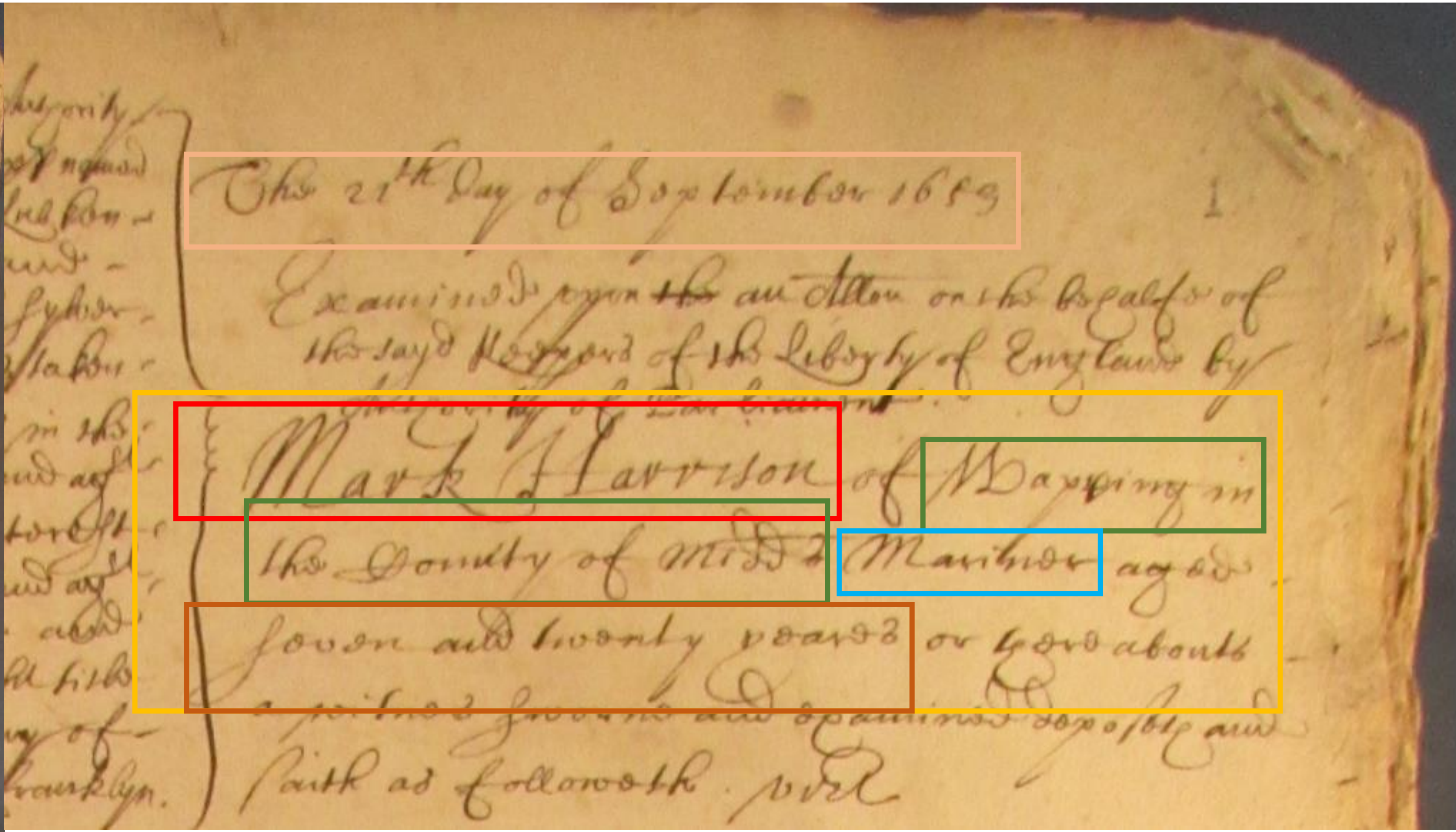
Legal deposition

Deposition of Mark Harrison; mariner and master; resident in Wapping, Middlesex; age 27;
Dated September 21st 1659 (TNA, HCA 13/68, ff. 1r-3r)

http://www.marinelives.org/wiki/HCA_13/68_f.1r_Annotate; http://www.marinelives.org/wiki/HCA_13/68_f.1v_Annotate; http://www.marinelives.org/wiki/HCA_13/68_f.2r_Annotate ; http://www.marinelives.org/wiki/HCA_13/68_f.2v_Annotate; http://www.marinelives.org/wiki/HCA_13/68_f.3r_Annotate

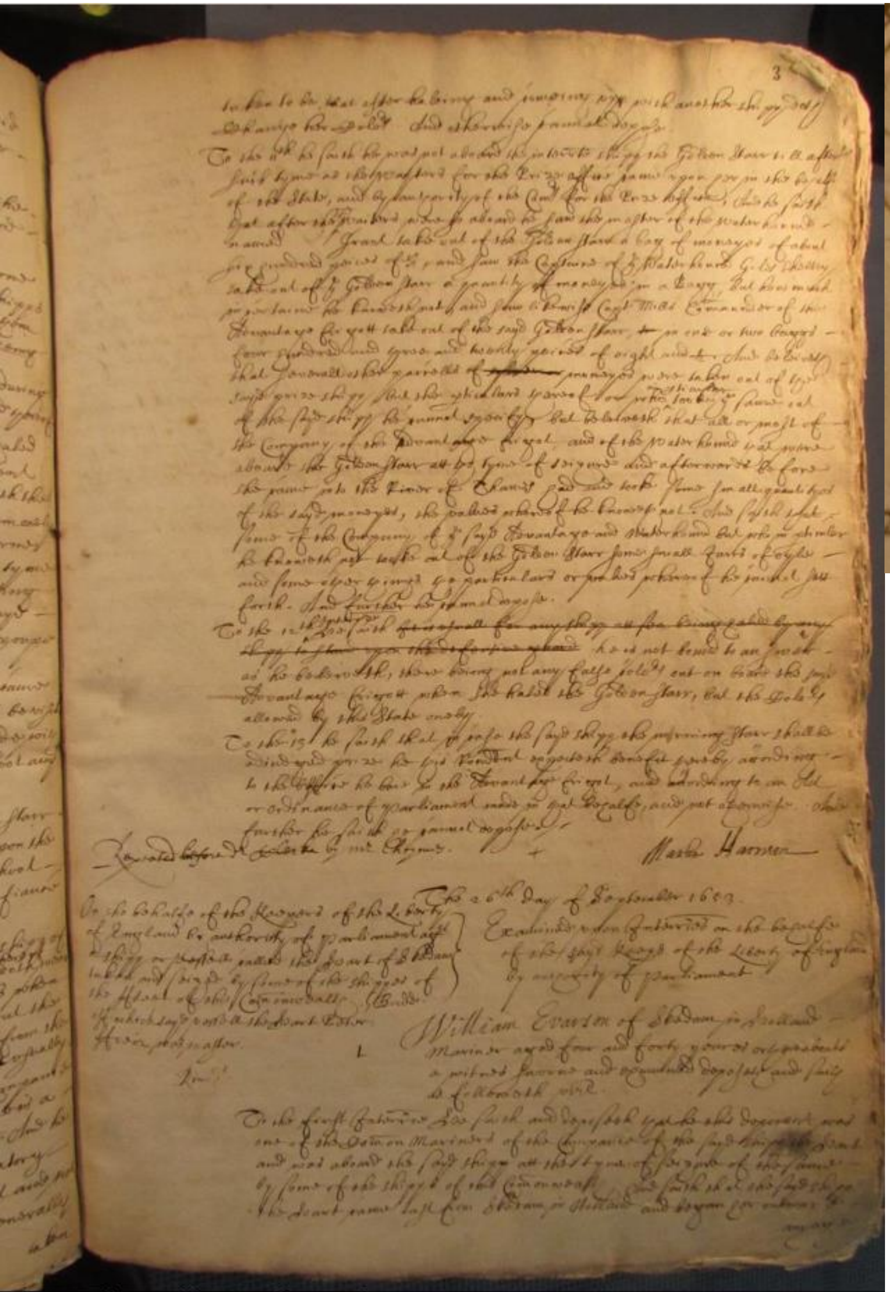
Metadata

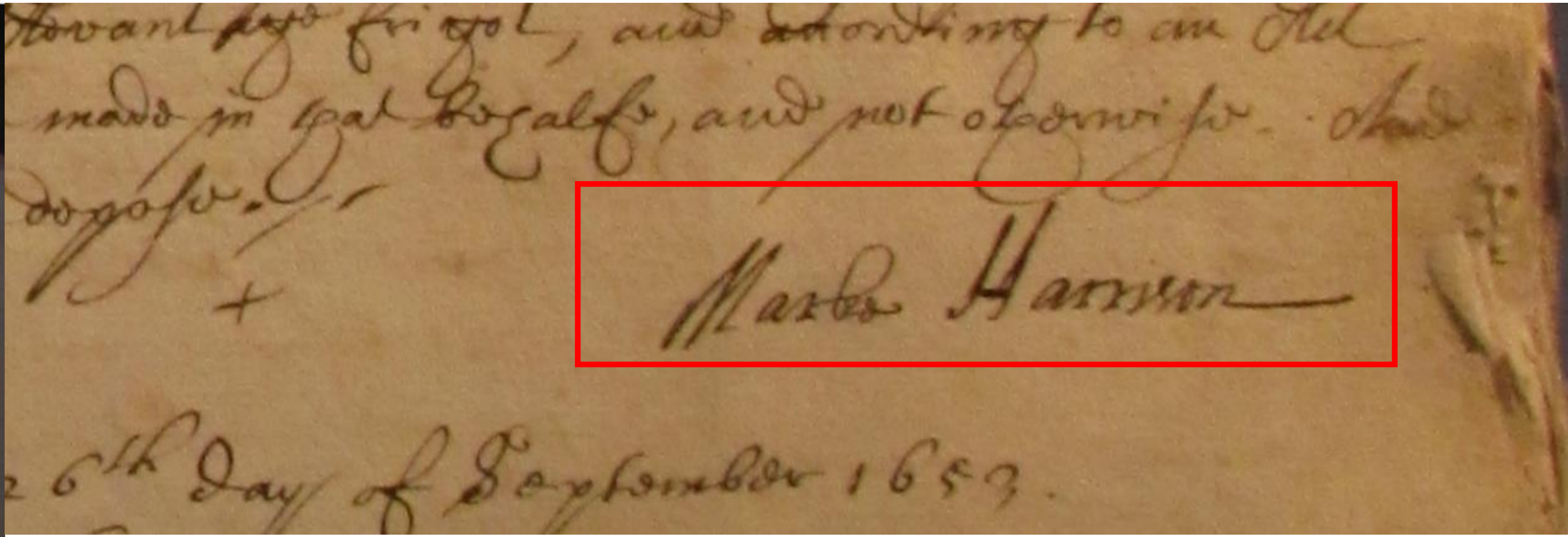




http://www.marinelives.org/wiki/HCA_13/68_f.1r_Annotate; http://www.marinelives.org/wiki/Mark_Harrison

Signoff



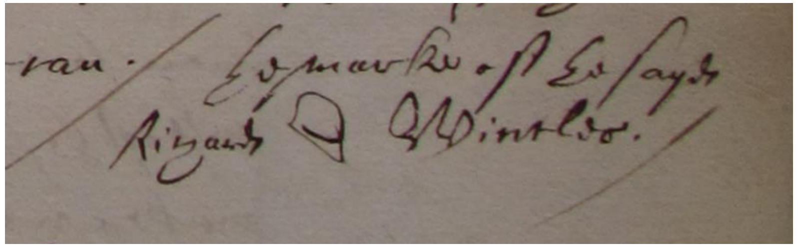


Servant frigot, and according to an Act
made in that behalfe, and not otherwise. And
deposeth.
+ Marke Harrison
6th day of September 1653.

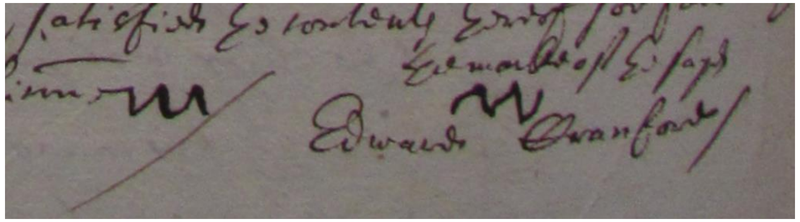
http://www.marinelives.org/wiki/HCA_13/68_f.3r_Annotate

Porters handling coals, whale oil, ginger & corn

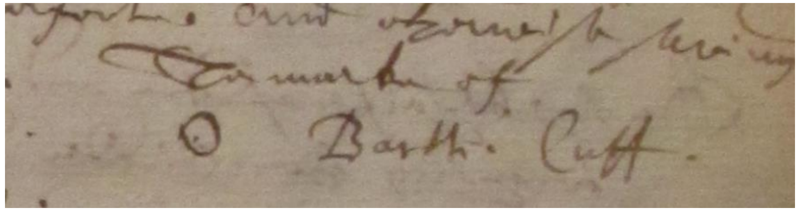
Richard Wincles, thirty-three year old porter, of the parish of Stepney, Middlesex, Dec. 15, 1656; employed as a labourer with fifteen other men to unload coals from the *Implyment* moored near Execution Dock, Wapping, into lighters for fixed rate of 12 *s* per man (HCA 13/70 f.554r)



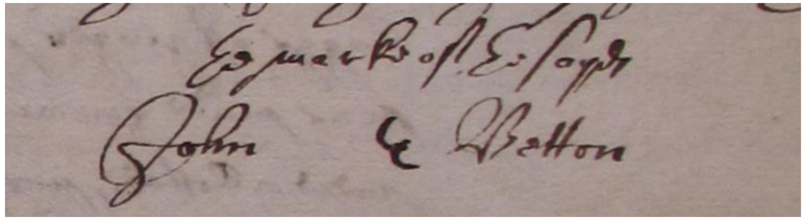
Edward Cranford, forty-four year old coale heaver or porter, of the parish of Stepney, Middlesex, Dec. 15, 1656; employed as a labourer with fifteen other men to unload coals from the *Implyment* moored near Execution Dock, Wapping, into lighters for fixed rate of 12 *s* per man (HCA 13/70 f.555v)



Bartholomew Cuff, sixty year old porter of the Stillyard, of the parish of Allhallowes the Greate, London, May 15, 1658; assisted in the landing of whale oil from lighters at the Stillyard Key and loading them away into a warehouse (HCA 13/70 f.555v)

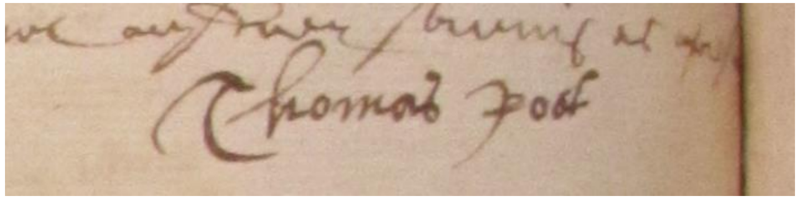


John Betton, fifty-four year old citizen and white baker of London, of the parish of Saint Buttolph Algate, London, Jul. 31, 1655; self-described as a porter employed by the Commissioners for Prize Goods to deliver ginger from a warehouse at Ralphes Key (HCA 13/70 f.449r)

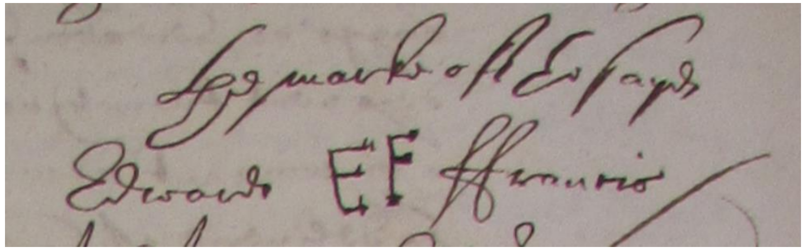


Edward Sherwin, fifty-six year old cittizen and leatherseller, of the parish of Little Allhallowes, London, Jul. 31, 1655; self-described as a porter employed by the Commissioners for Prize Goods to deliver ginger from a warehouse at Ralphes Key (HCA 13/70 f.449v)

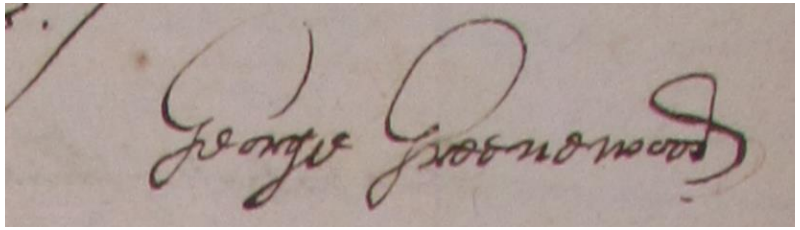
Thomas Roots, twenty-nine year old porter, of the parish of Greate Allhallowes, London, May 15, 1658; assisted in the landing of whale oil from lighters at the Stillyard Key, as one of the Stillyard porters, and loading them away into a warehouse (HCA 13/72 f.330v)



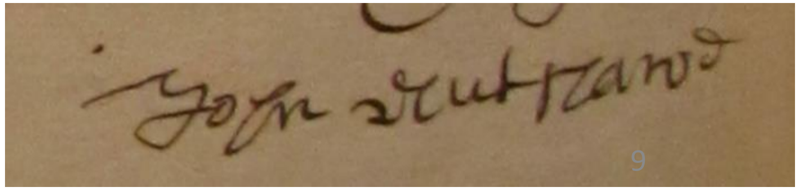
Edward ffrancis, citizen and merchant taylor of London, of the parish of Saint Olave in Southwarke, Jul. 31, 1655; self-described as a porter employed by the Commissioners for Prize Goods to deliver ginger from a warehouse at Ralphes Key (HCA 13/70 f.450v)



George Greenwood, thirty year old citizen and vintner of London, of the parish of Saint Buttolph Bishopsgate, London, Jul. 31, 1655; self-described as a porter employed by the Commissioners for Prize Goods to deliver ginger from a warehouse at Ralphes Key (HCA 13/70 f.454r)

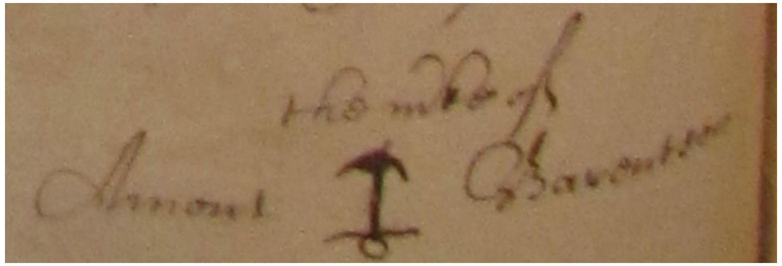


John Nutshall, fifty-five year old corne porter, of the parish of Saint Saviours Southwarke, Nov. 19, 1653; employed with a barber chyrurgeon/corne meter, an additional corne-meter, and other labourers to unlade a cargo of what in the *ffortune* of Stettin, moored against Limehouse; eight years of experience as a corne porter (HCA 13/70 f.352v)

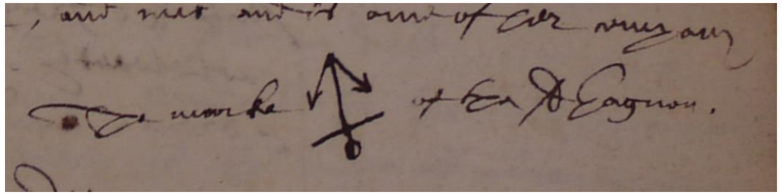


Anchors

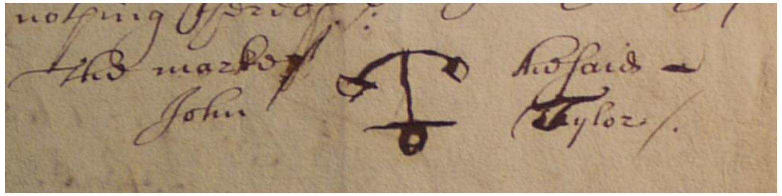
Amons Barentsen, thirty-five year old mariner, of Copenhagen, Denmark, October 13th, 1653; self-described as an “ordinary mariner”, hired to sail from the Sound to Hamburg on the *Golden Hawke* of Stockholm ([HCA 13/68 f.81v]())



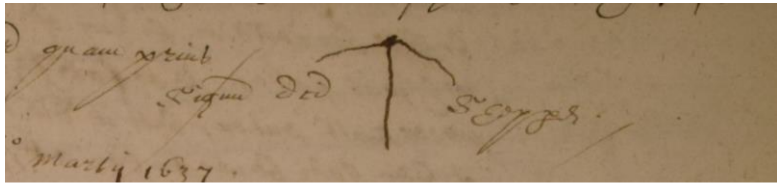
Claude de Gagnon, twenty-five year old mariner, of Melon, near Brest in Britanny, May 22nd, 1656 ([HCA 13/71 f.225r]())



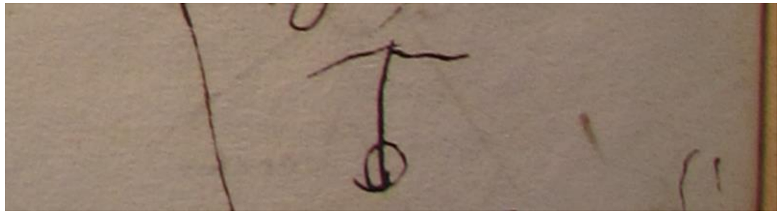
John Tylor, forty-two year old shipwright, of Lower Shadwell, in the parish of Stepney, Middlesex, February 14th, 1659 ([HCA 13/73 f.36r]())



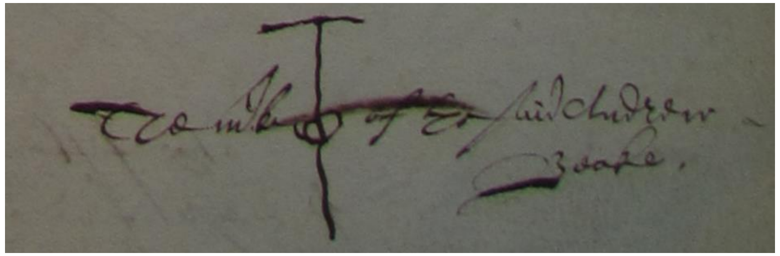
Richard Shepperd, fifty-eight year old cooke, of Brixton, Devon, March 29th, 1637; self-described cooke of the *Hope* of Ipswich ([HCA 13/53 f.87r]())



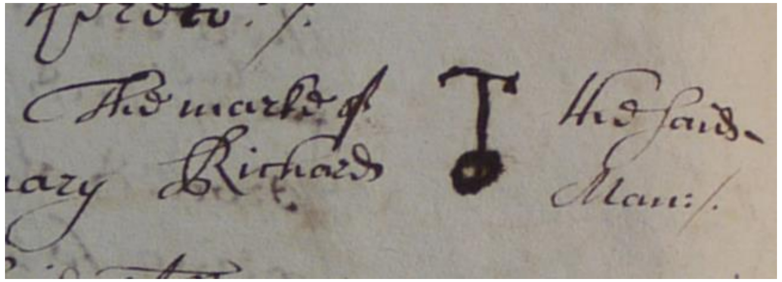
Andrew Beake, thirty-six year old lookeinglassemaker and formerly seaman, of Rose alley without Bishopsgate, London, January 21st, 1655 ([HCA 13/70 f.252v]())



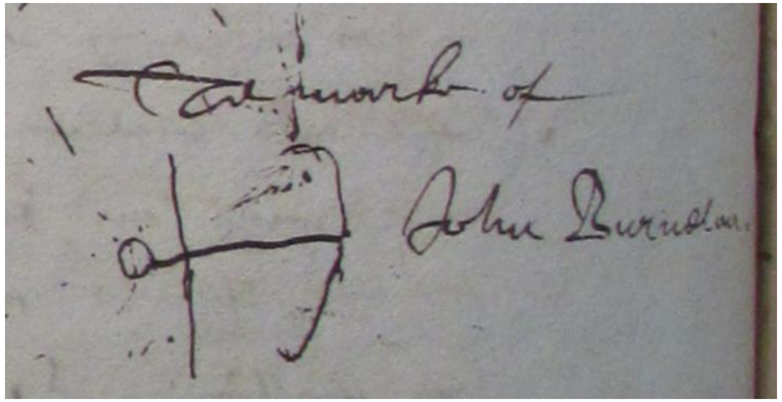
Andrew Beake, thirty-six year old looking-glasse maker, of Rose-Alley in Bishopsgate streete, London, February 13th, 1655 ([HCA 13/70 f.252v]())



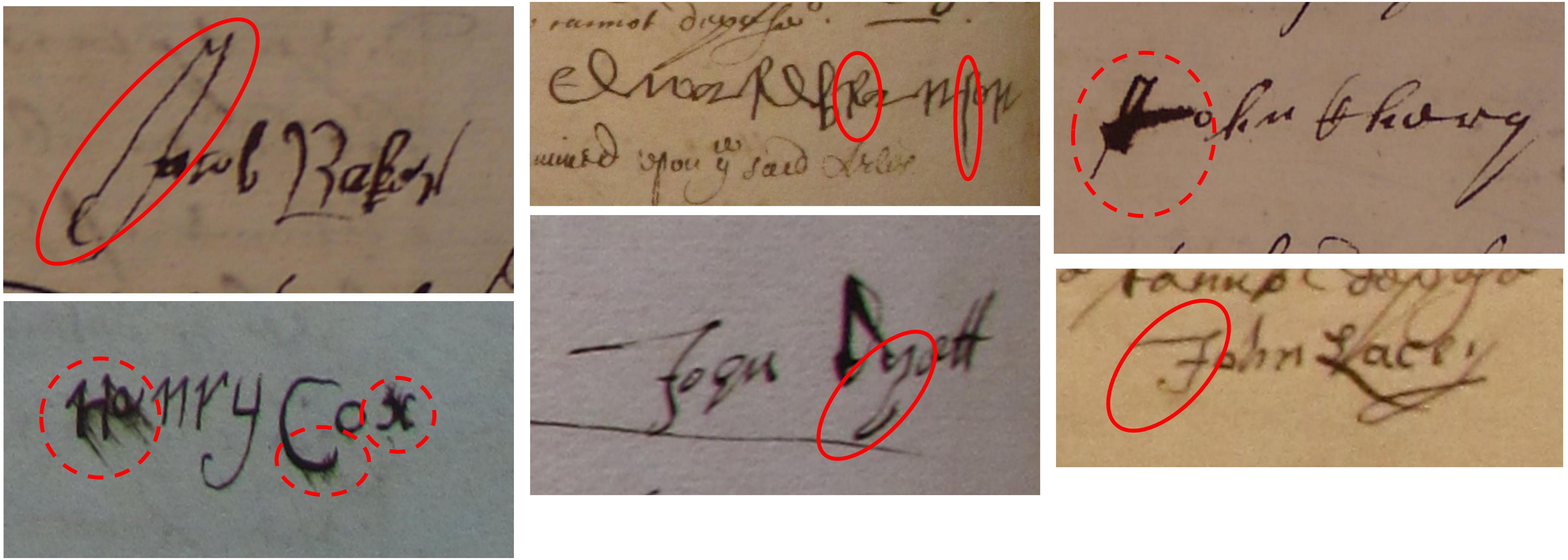
Richard Man, thirty-one year old mariner, of Southampton, January 8th, 1659; self-described common man of the *Lisbone ffrigott* on voyage to Oratava ([HCA 13/73 f.26v]())



John Burnelau, twenty-eight year old sailor, of Mornar, France,March 30th, 1661 ([HCA 13/73 f.486v]())



Physical characteristics of poorly executed signatures for machine detection – Part 1

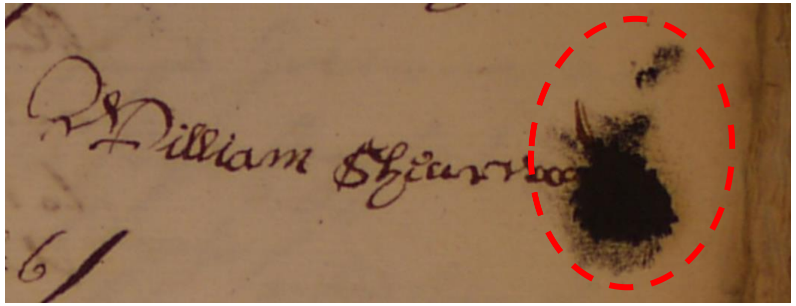


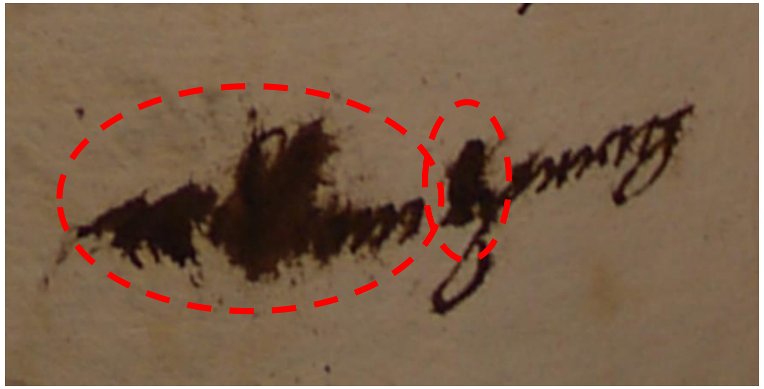
Shaky straight lines and/or loops

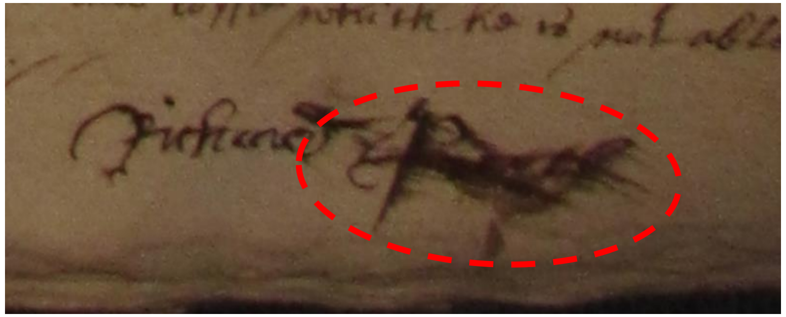
Ink blots or smudges

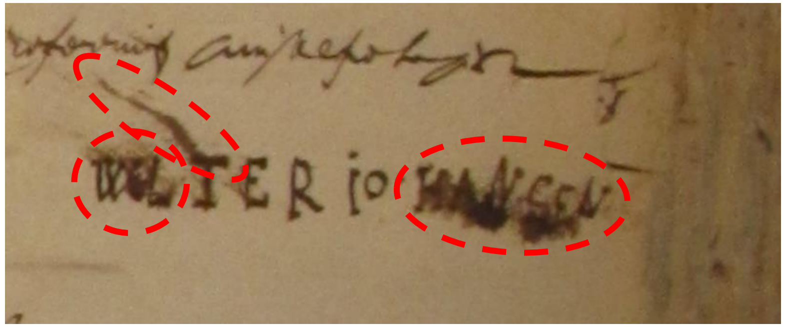
Source: Clockwise from top LH side: KaggleTestSnippet_HCA_1371_f.263v.PNG, KaggleTestSnippet_HCA_1368_f.483v.PNG, KaggleTestSnippet_HCA_1371_f.456r.PNG, KaggleTestSnippet_HCA_1368_f.51v.PNG, KaggleTestSnippet_HCA_1370_f.168v.PNG, KaggleTestSnippet_HCA_1370_f.167r.PNG

Physical characteristics of poorly executed signatures for machine detection – Part 2





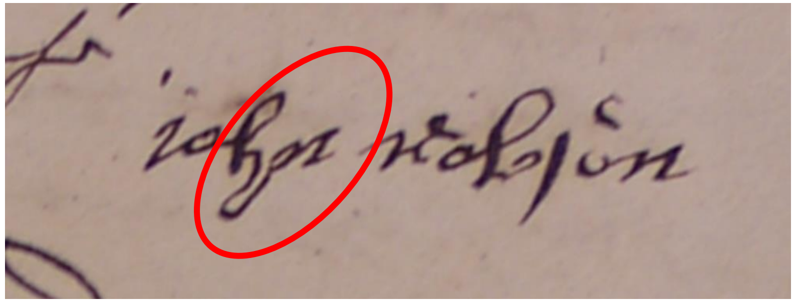


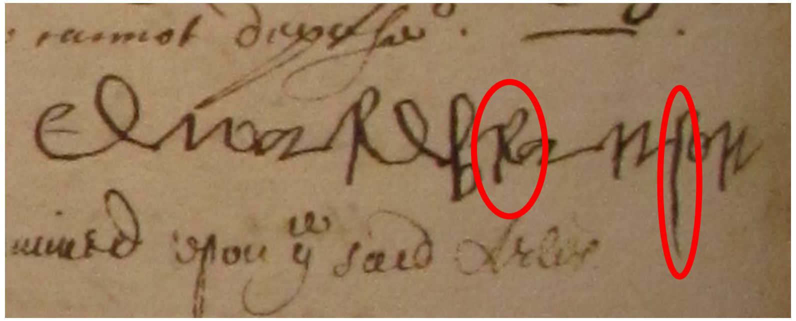


Ink blots or smudges

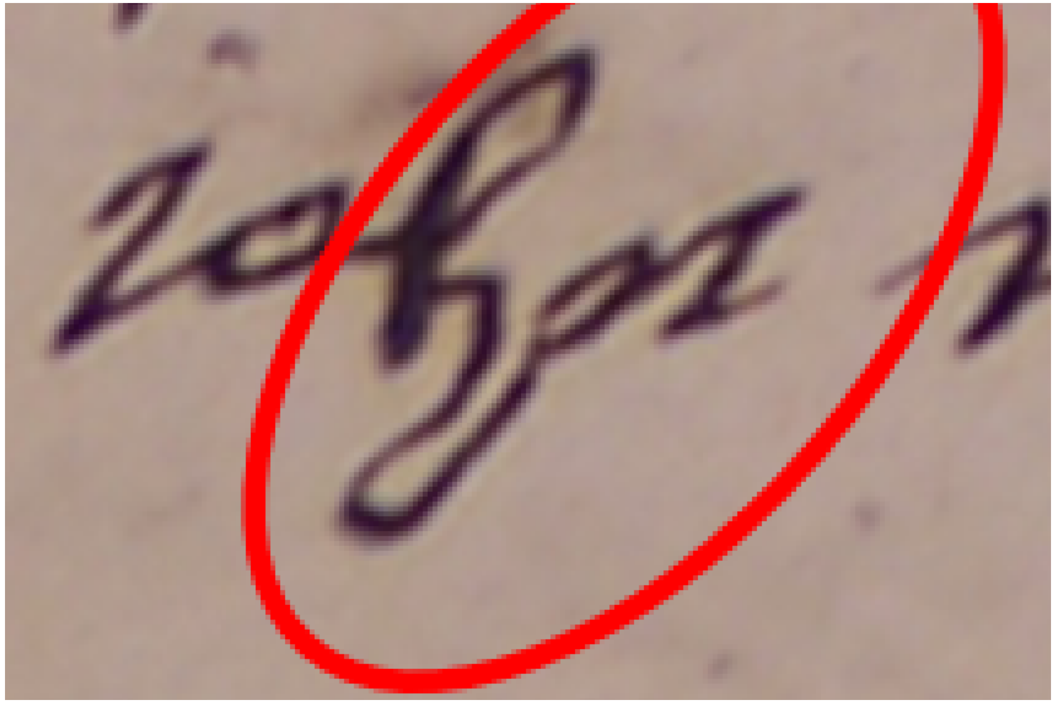
Source: Clockwise from top LH side: KaggleTestSnippet_HCA_1371_f.503r.PNG, KaggleTestSnippet_HCA_1373_f.498v.PNG, KaggleTestSnippet_HCA_1368_f.59r.PNG, KaggleTestSnippet_HCA_1368_f.231r.PNG

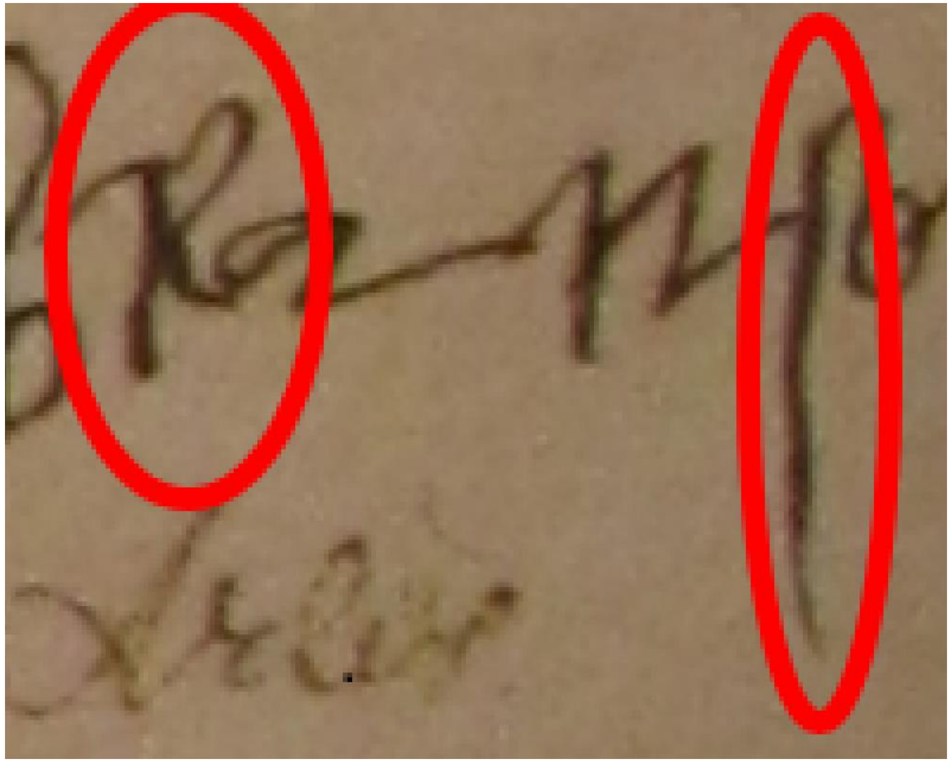
We are looking for algorithms to detect “shake” in straight and curved lines





HYPOTHESIS: Shaky lines are a sign of poor signature execution (and by inference, poor handwriting execution) suggesting lower level of literacy than smooth executed lines





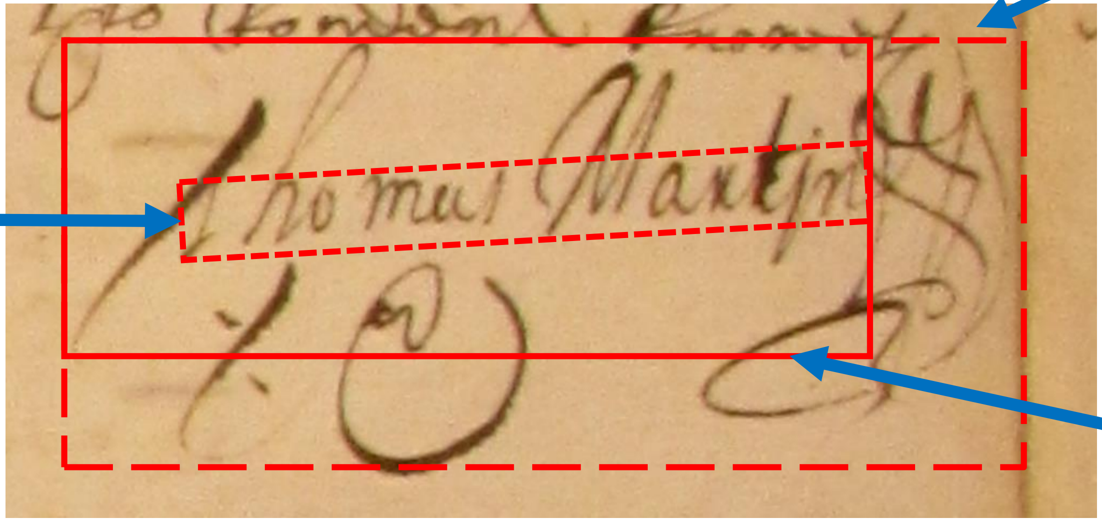
Shaky straight lines and/or loops

Clockwise, from top LH: KaggleTestSnippet_HCA_1371_f.435v.PNG, KaggleTestSnippet_HCA_1368_f.483v.PNG, KaggleTestSnippet_HCA_1368_f.483v_PIXELS.PNG, KaggleTestSnippet_HCA_1371_f.435v.PNG_PIXELS.PNG

Putting boundary boxes on C17th signatures

Boundary boxes marking the visual geometry of a signature

Inside boundary box, excluding uppers and downers

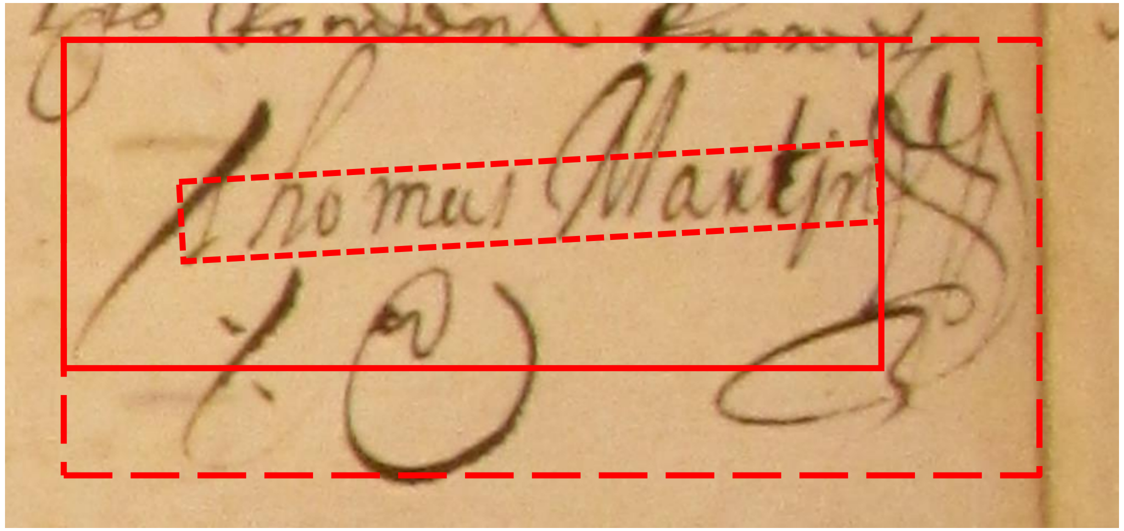


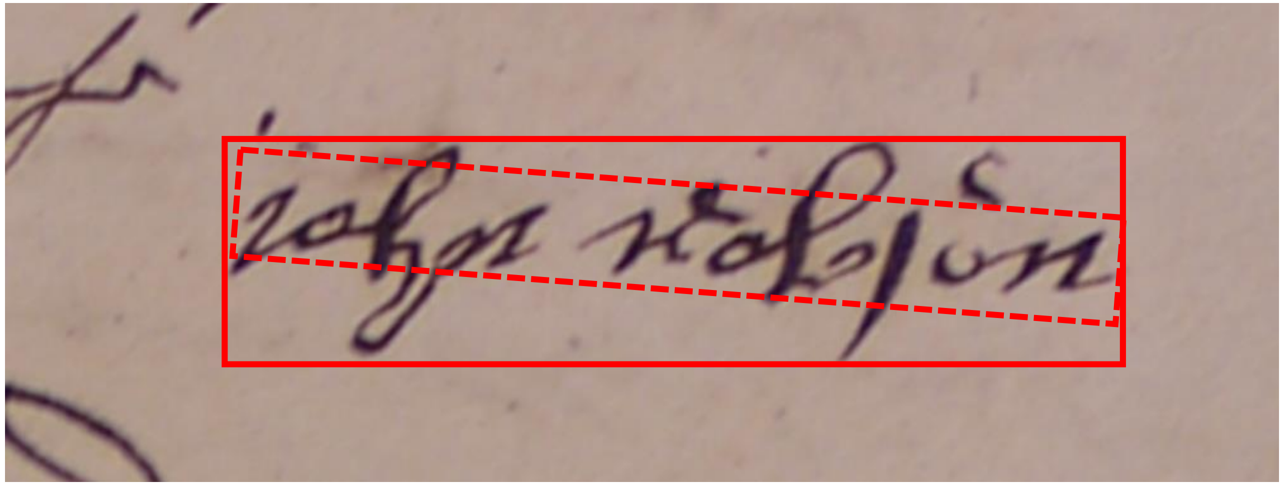
Outside boundary box, including flourish

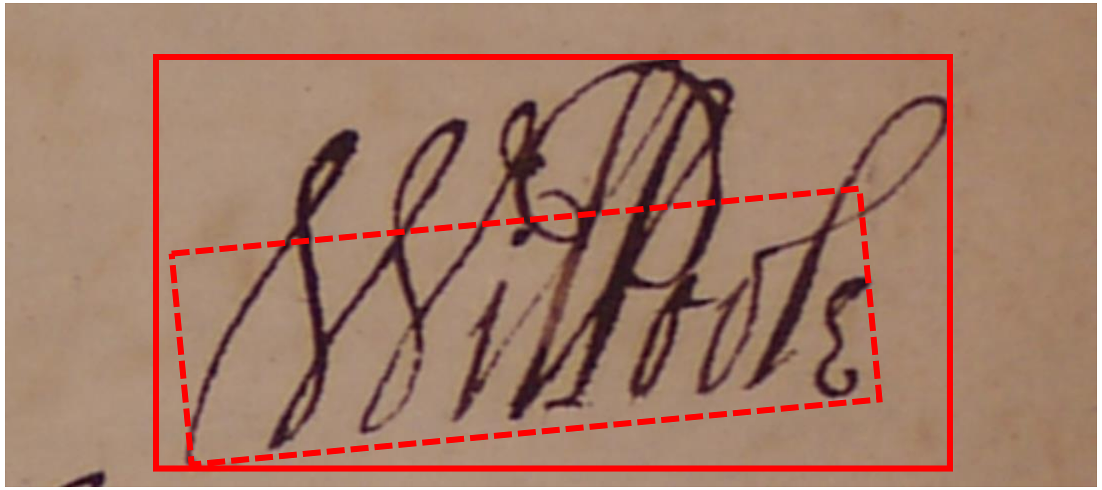
Middle boundary box, including all letters, but excluding flourish

Statistics
Inside boundary box: 9.0 x 1.1
Middle boundary box: 9.75 x 4.25
Outside boundary box: 12.75 x 5.75
Rotation from horizontal: ca. 340 degrees

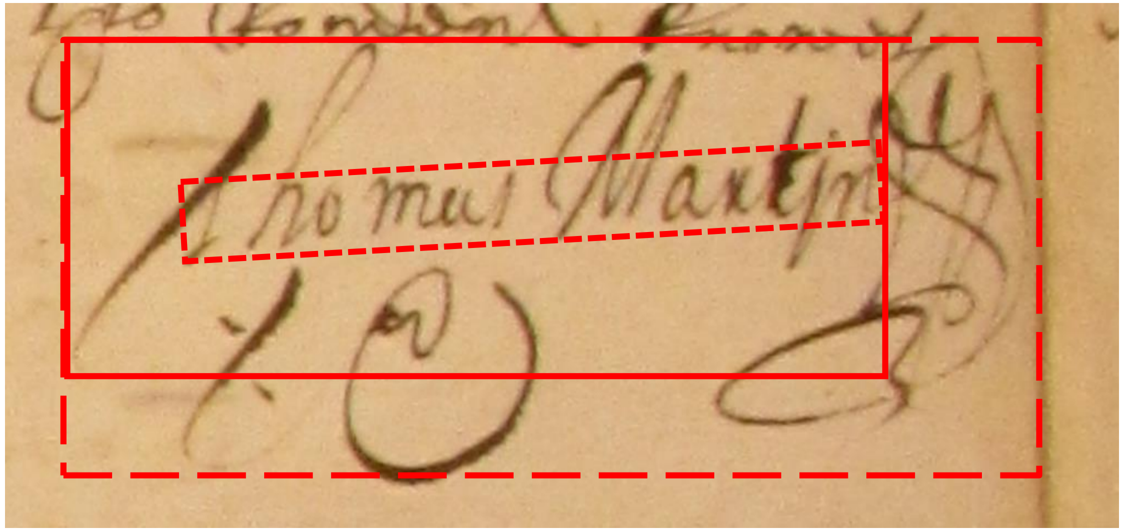
Different visual geometries of signatures

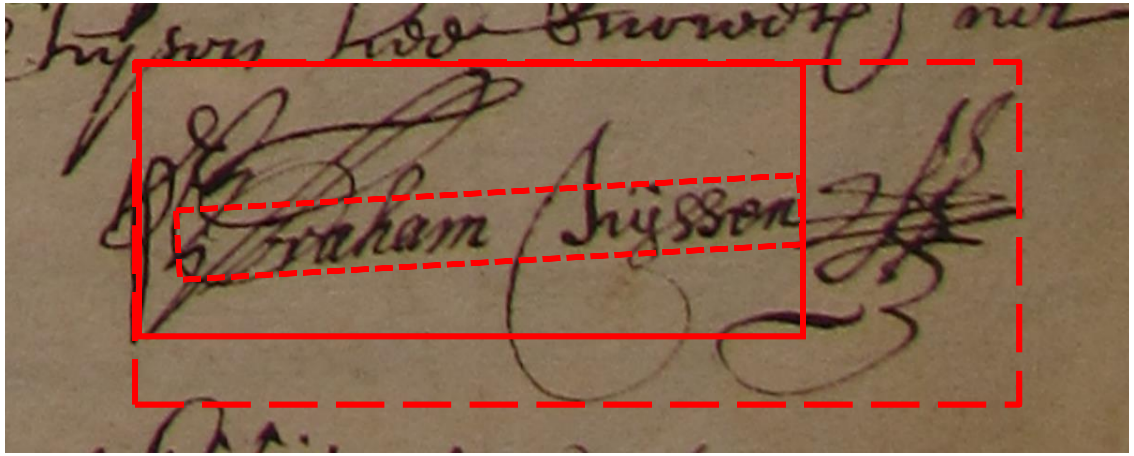


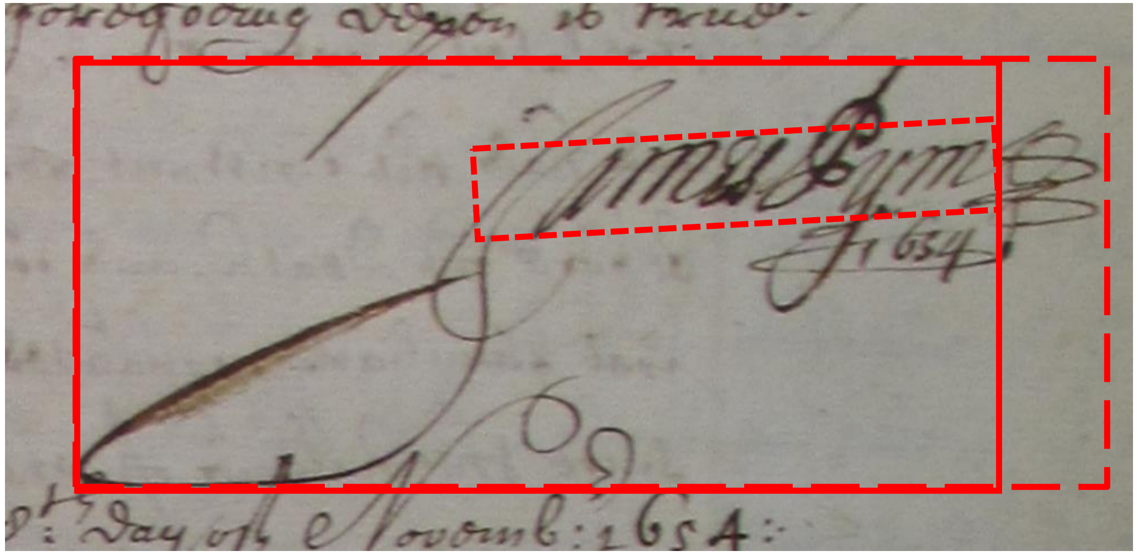


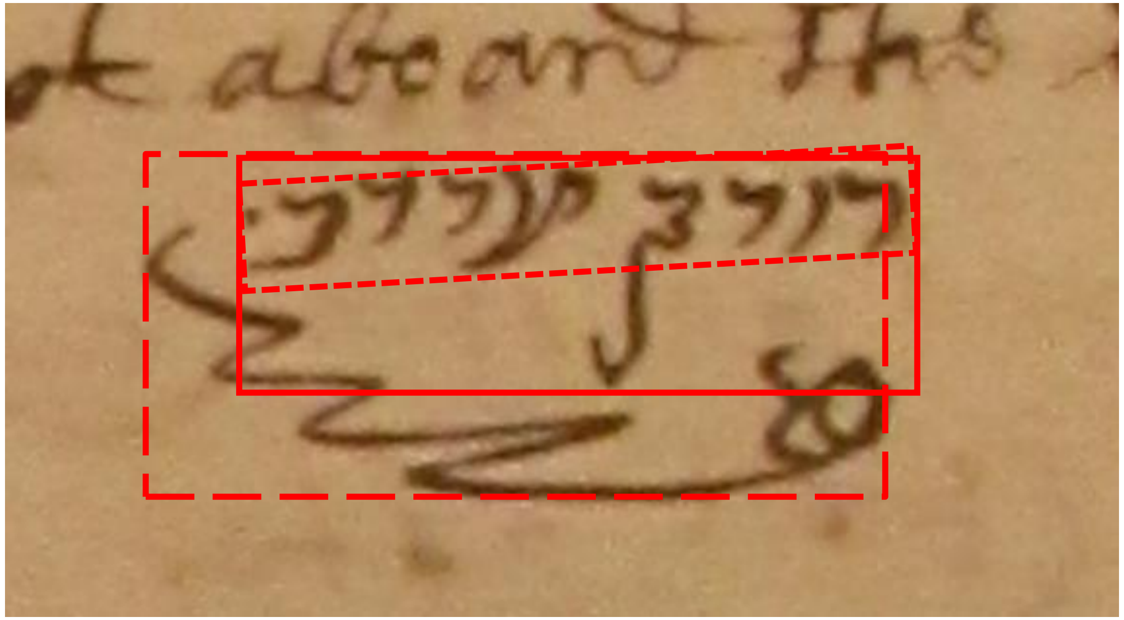


Visual geometries of flourishes – C17th Irish, Dutch, English & Moroccan merchants



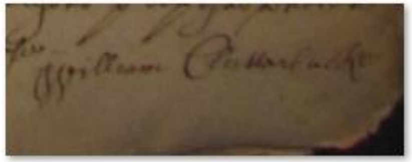




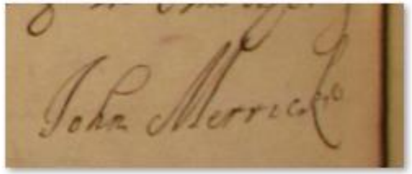


Source: Clockwise from top LH side: KaggleTestSnippet_HCA_1368_f.34v.PNG, KaggleTestSnippet_HCA_1370_f.366r.PNG, KaggleTestSnippet_HCA_1370_f.134r.PNG, KaggleTestSnippet_HCA_1368_f.58r.PNG

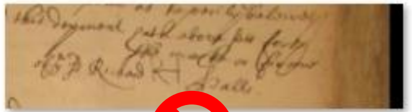
Challenge One: Identify the 11 image snippets on this page which contain signatures and highlight the 2 image snippets which contain signatures belonging to the same person



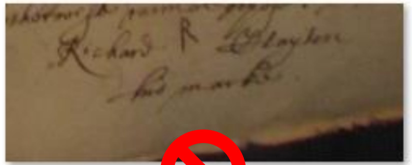
KaggleTestSnippet_HCA_1368_f.140r.PNG



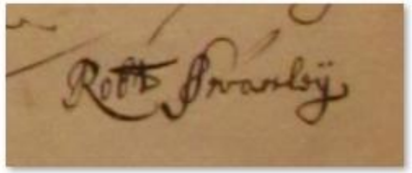
KaggleTestSnippet_HCA_1368_f.142r.PNG



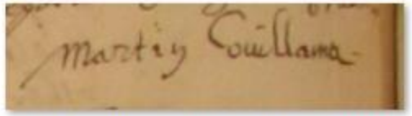
KaggleTestSnippet_HCA_1368_f.143r.PNG



KaggleTestSnippet_HCA_1368_f.145r.PNG



KaggleTestSnippet_HCA_1368_f.148r.PNG

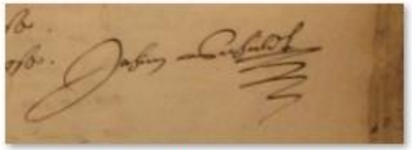


KaggleTestSnippet_HCA_1368_f.150v.PNG

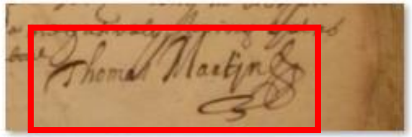
KaggleTestSnippet_HCA_1368_f.151v.PNG

KaggleTestSnippet_HCA_1368_f.153v.PNG

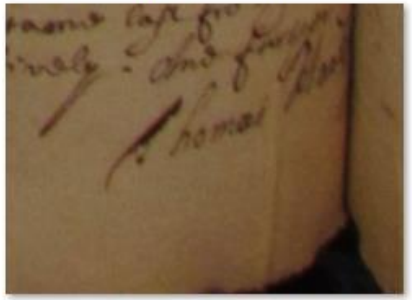
KaggleTestSnippet_HCA_1368_f.155r.PNG



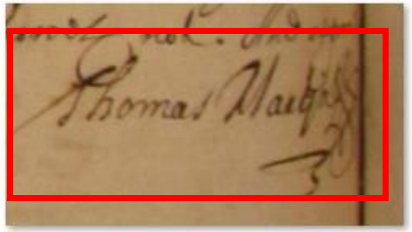
KaggleTestSnippet_HCA_1368_f.156r.PNG



KaggleTestSnippet_HCA_1368_f.158r.PNG



KaggleTestSnippet_HCA_1368_f.159v.PNG



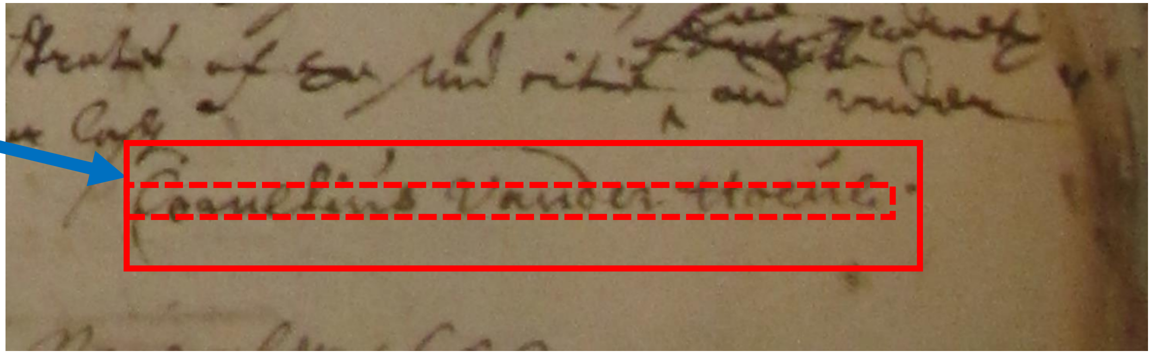
KaggleTestSnippet_HCA_1368_f.161v.PNG

Source: KaggleTestSnippets data set

Challenge Two: Detect a London based merchant, who has Dutch origin, from physical characteristics of signature, rather than spelling of name

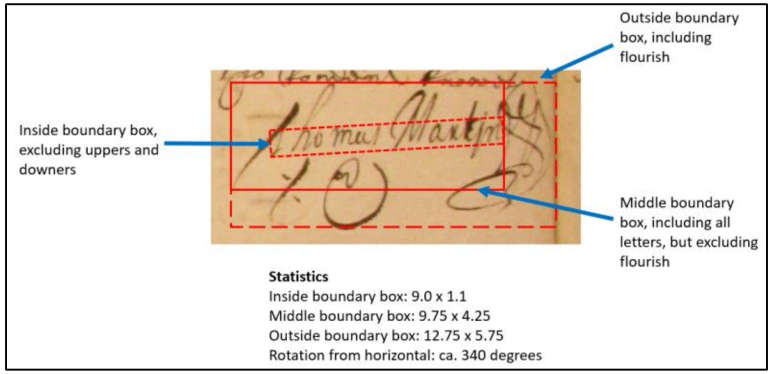
Semantic giveaways

- 3 names, not 2
- Specific names ["Cornelius"; "Vander"; "Hoene"]



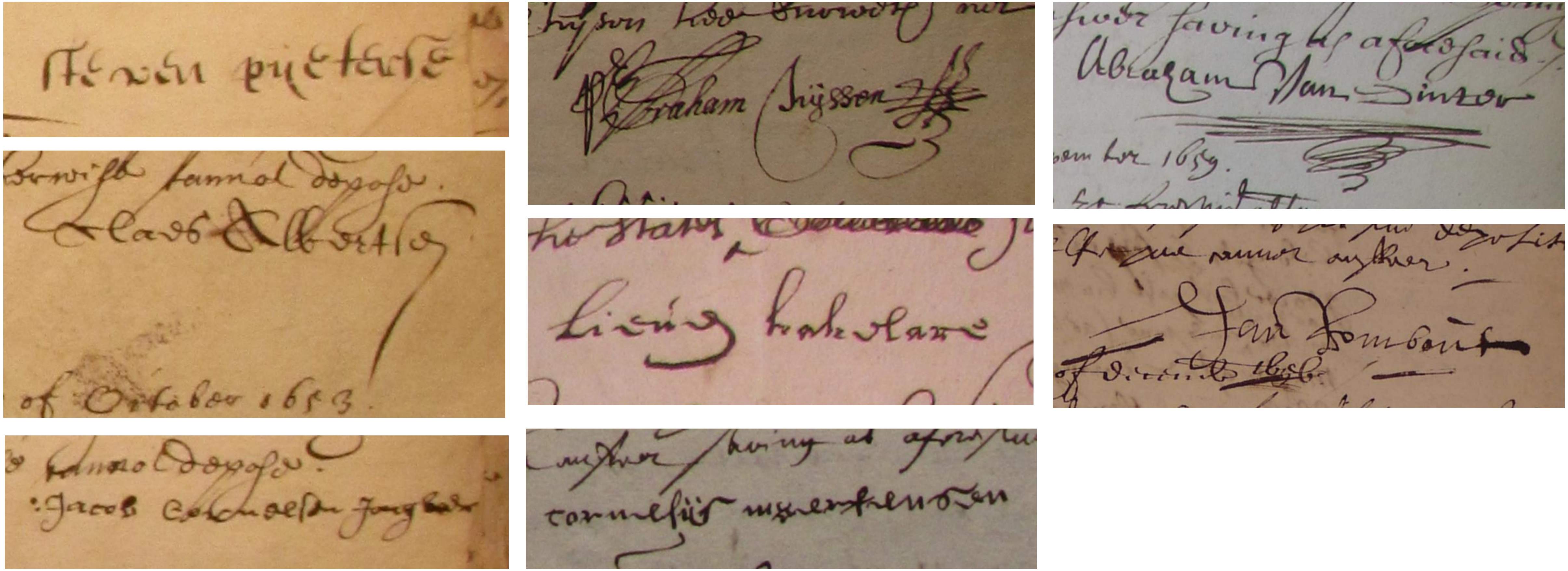
Physical giveaways

- Long, narrow signature
- No flourish
- Limited capitalisation
- Middle bounding box close to inner bounding box, with no outside bounding box



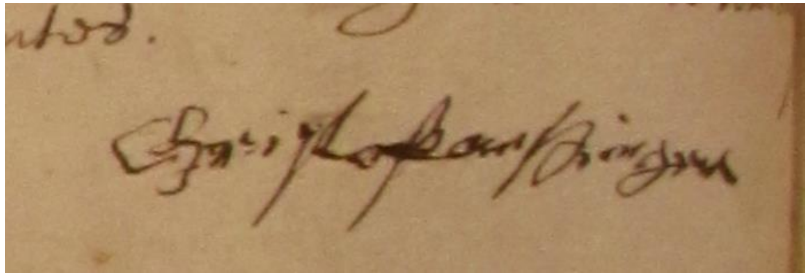
Source: KaggleTestSnippet_HCA_1368_f.233r_Two.PNG

Ilustrative research question: Was it less common to use capital letters in Dutch rather than English language signatures in the C17th

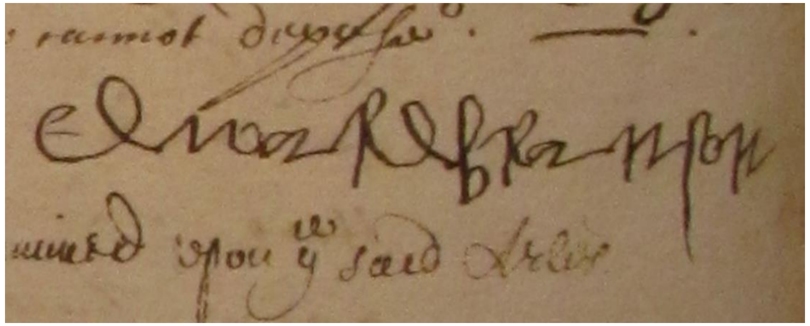


Source: Clockwise from top LH side: XXXXX

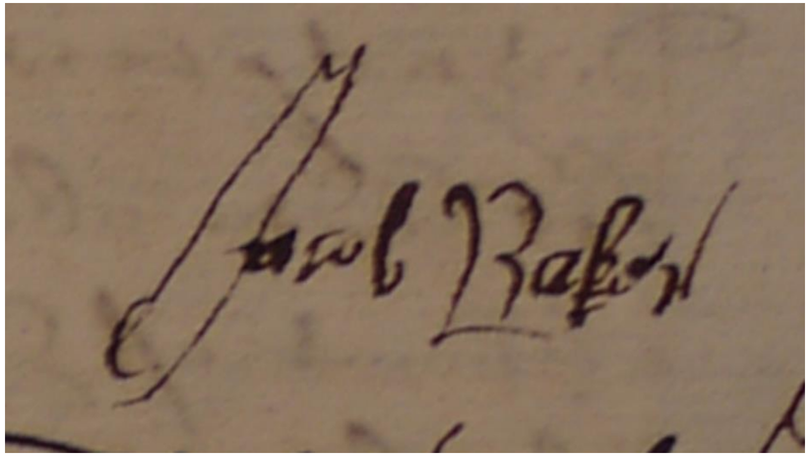
Ilustrative research question : Is there an age effect in terms of physical control of pen, independent of any age-independent physical infirmity, and independent of a tendency for lower literacy (and possibly lower control of pen) amongst earlier age cohorts in our samples of signatures from the mid-C17th?



Christopher Drake, 86 year old sugar refiner, of Saint Mary Street, London, born 1567, signature dated Nov 2, 1653



Edward Branston, 47 year old rope maker, of Saint Mary Matsellon alias Whitechapel, born 1607, signature dated Jan 25, 1654



James Baker, 21 year old mariner, of Wapping, one of the company of the ship the *Plaine Dealeing*, born 1635, signature dated Jun 21, 1656

Source: Top to bottom, KaggleTestSnippet_HCA_1368_f.60v.PNG, KaggleTestSnippet_HCA_1368_f.483v.PNG, KaggleTestSnippet_HCA_1371_f.263v.PNG

SUPPLEMENTARY MATERIAL

Issues

Pre-processing

- Scaling images
- Converting to grayscale
  - Some image processing & computer vision algorithms use grayscale images not colour images
  - Grayscale processing 3x faster than colour processing
- Normalising an image
  - Avoid larger feature values dominating smaller feature values [THINK ABOUT THIS]

Image processing packages

- OpenCV: reads & plots an image in BGR format. Reads PNG & JPG on 0 to 255 range
- Matplotlib: reads & plots an image in RGB format. Reads JPEG in 0 to 255 and PNG on 0 to 1 range

Do images need to be square for typical neural network models? Or at least have identical aspect ratios? Do we need to scale all images to a standard number of pixels for height & width? Do snippets need to be cropped to isolate the signoff? What does the mean image of a marke, an initial and a signature look like? [mean value of each pixel across all training examples]. Can also look at the standard deviation of the pixels for a group of images.

Reading

Colin Greenstreet, Pattern recognition of signatures and markes in historical manuscripts as the basis for sub-population recognition, March 2018 [available Signsofliteracy Github repository: Signsoliteracy/Signoff]

Colin Greenstreet, C17th alphabet of initials, 4th edn., April 4th, 2018 [available Signsofliteracy Github repository: Signsoliteracy/Signoff]

Mark Hailwood, 'The Rabble that Cannot Read', Ordinary Peoples Literacy in Seventeenth-Century England, October 13th, 2014

David Cressy, Literacy and the Social Order: Reading & Writing in Tudor and Stuart England, 1980

Reading

Cornell University Library

We gratefully acknowledge support from the Simons Foundation and member institutions

arXiv.org > cs > arXiv:1604.04004

Search or Article ID | All fields

(Help | Advanced search)

Computer Science > Computer Vision and Pattern Recognition

Understanding How Image Quality Affects Deep Neural Networks

Samuel Dodge, Lina Karam

(Submitted on 14 Apr 2016 (v1), last revised 21 Apr 2016 (this version, v2))

Image quality is an important practical challenge that is often overlooked in the design of machine vision systems. Commonly, machine vision systems are trained and tested on high quality image datasets, yet in practical applications the input images can not be assumed to be of high quality. Recently, deep neural networks have obtained state-of-the-art performance on many machine vision tasks. In this paper we provide an evaluation of 4 state-of-the-art deep neural network models for image classification under quality distortions. We consider five types of quality distortions: blur, noise, contrast, JPEG, and JPEG2000 compression. We show that the existing networks are susceptible to these quality distortions, particularly to blur and noise. These results enable future work in developing deep neural networks that are more invariant to quality distortions.

Comments: Final version will appear in IEEE Xplore in the Proceedings of the Conference on the Quality of Multimedia Experience (QoMEX), June 6-8, 2016

Subjects: **Computer Vision and Pattern Recognition (cs.CV)**

Cite as: **arXiv:1604.04004 [cs.CV]**

(or **arXiv:1604.04004v2 [cs.CV]** for this version)

Submission history

From: Samuel Dodge [view email]

[v1] Thu, 14 Apr 2016 00:47:50 GMT (2833kb,D)

[v2] Thu, 21 Apr 2016 20:44:52 GMT (2833kb,D)

Download:

- PDF
- Other formats

(license)

Current browse context:

cs.CV

< prev | next >

new | recent | 1604

Change to browse by:

cs

References & Citations

- NASA ADS

DBLP - CS Bibliography

listing | bibtex

Samuel F. Dodge
Lina J. Karam

Bookmark (what is this?)

Reading

Labeled Faces in the Wild

Menu
- LFW Home
  - Mailing
  - Explore
  - Download
  - Train/Test
  - Results
  - Information
  - Errata
  - Reference
  - Resources
  - Contact
  - Support
  - Changes
- Part Labels
- UMass Vision

Labeled Faces in the Wild Home

NEW SURVEY PAPER:

Erik Learned-Miller, Gary B. Huang, Aruni RoyChowdhury, Haoxiang Li, and Gang Hua.
Labeled Faces in the Wild: A Survey.
In *Advances in Face Detection and Facial Image Analysis*, edited by Michal Kawulok, M. Emre Celebi, and Bogdan Smolka, Springer, pages 189-248, 2016.
[Springer Page] [Draft pdf]

NEW RESULTS PAGE:

WE HAVE RECENTLY UPDATED AND CHANGED THE FORMAT AND CONTENT OF OUR RESULTS PAGE. PLEASE REFER TO THE NEW TECHNICAL REPORT FOR DETAILS OF THE CHANGES.

Welcome to Labeled Faces in the Wild, a database of face photographs designed for studying the problem of unconstrained face recognition. The data set contains more than 13,000 images of faces collected from the web. Each face has been labeled with the name of the person pictured. 1680 of the people pictured have two or more distinct photos in the data set. The only constraint on these faces is that they were detected by the Viola-Jones face detector. More details can be found in the technical report below.

There are now four different sets of LFW images including the original and three different types of "aligned" images. The aligned images include "funneled images" (ICCV 2007), LFW-a, which uses an unpublished method of alignment, and "deep funneled" images (NIPS 2012). Among these, LFW-a and the deep funneled images produce superior results for most face verification algorithms over the original images and over the funneled images (ICCV 2007).

Related:

[new] Collected resources related to LFW - updated 2017/05/09.

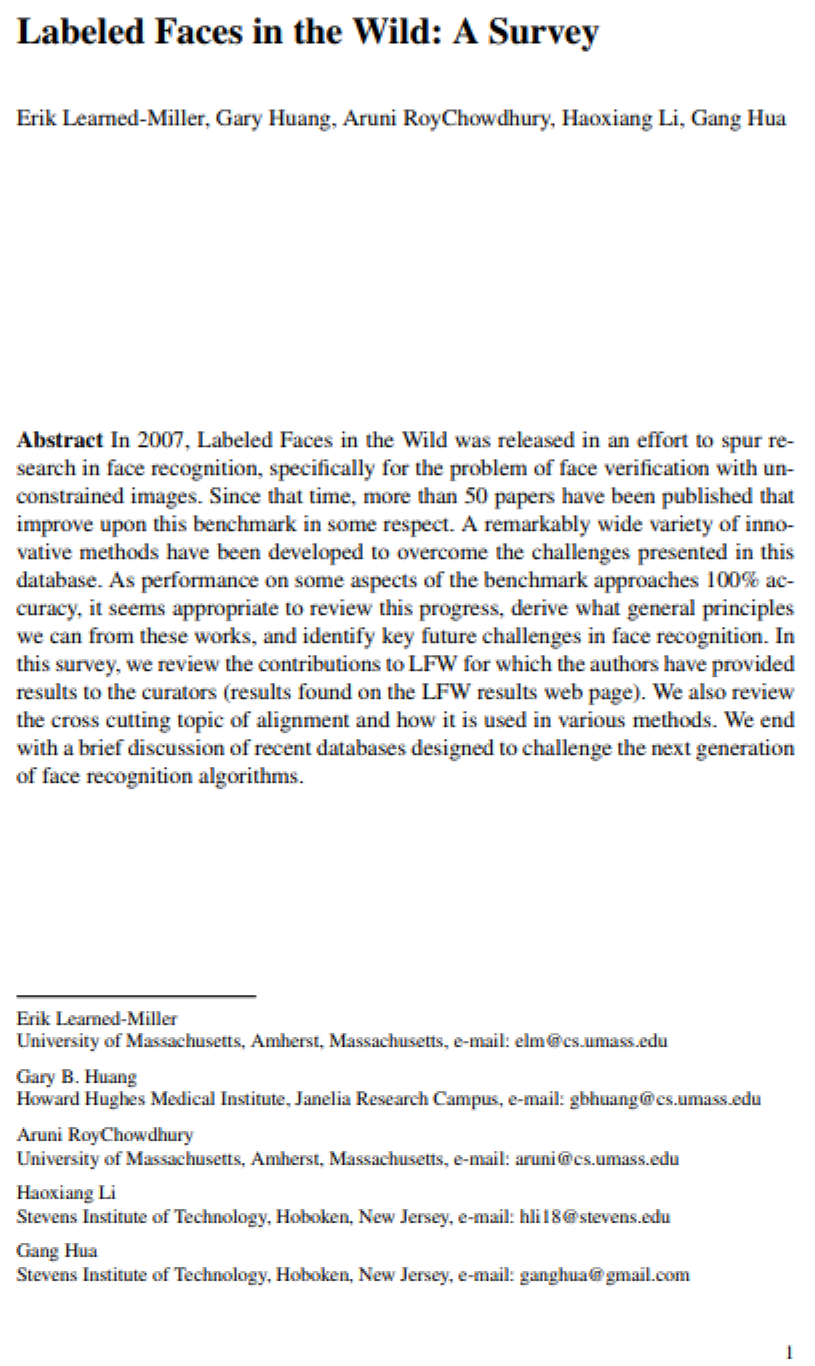
LFW Deep Funneled Images.

LFW attributes file (see Attribute and Simile Classifiers for Face Verification, Kumar et al.).

Face Detection Data set and Benchmark (FDDB), our new database for face detection research.

Faces in Real-Life Images workshop at the European Conference on Computer Vision 2008, run by Erik Learned-Miller, Andras Ferencz, and Frederic Jurie.

Reading



Labeled Faces in the Wild: A Survey

Erik Learned-Miller, Gary Huang, Aruni RoyChowdhury, Haoxiang Li, Gang Hua

Abstract In 2007, Labeled Faces in the Wild was released in an effort to spur research in face recognition, specifically for the problem of face verification with unconstrained images. Since that time, more than 50 papers have been published that improve upon this benchmark in some respect. A remarkably wide variety of innovative methods have been developed to overcome the challenges presented in this database. As performance on some aspects of the benchmark approaches 100% accuracy, it seems appropriate to review this progress, derive what general principles we can from these works, and identify key future challenges in face recognition. In this survey, we review the contributions to LFW for which the authors have provided results to the curators (results found on the LFW results web page). We also review the cross cutting topic of alignment and how it is used in various methods. We end with a brief discussion of recent databases designed to challenge the next generation of face recognition algorithms.

Erik Learned-Miller
University of Massachusetts, Amherst, Massachusetts, e-mail: elm@cs.umass.edu

Gary B. Huang
Howard Hughes Medical Institute, Janelia Research Campus, e-mail: gbhuang@cs.umass.edu

Aruni RoyChowdhury
University of Massachusetts, Amherst, Massachusetts, e-mail: aruni@cs.umass.edu

Haoxiang Li
Stevens Institute of Technology, Hoboken, New Jersey, e-mail: hli18@stevens.edu

Gang Hua
Stevens Institute of Technology, Hoboken, New Jersey, e-mail: ganghua@gmail.com

1

Reading

Siamese Convolutional Neural Networks for Authorship Verification

William Du
Stanford University
450 Serra Mall, Stanford, CA 94305
willadu@stanford.edu

Michael Fang
Stanford University
450 Serra Mall, Stanford, CA 94305
mjfang@stanford.edu

Margaret Shen
Stanford University
450 Serra Mall, Stanford, CA 94305
marshen@stanford.edu

Abstract

Determining handwriting authorship of a written text has practical significance in the realm of forensics, signature verification, and literary history. While there have been studies in signature verification and handwriting classification, a vast literature review reveals that very little work has been done in handwriting verification. Recent advances in convolutional architectures, particularly those involving facial verification, suggest that the task can be tackled effectively. In this study, we build a Siamese convolutional neural network to determine whether two pieces of handwriting are written by the same author. We examine questions such as whether long pieces of handwriting must be present to achieve good results, how many samples are needed, what features are important, and how different architectures perform on this task. We explore different convolutional architectures like VGG, GoogLeNet and ResNet, to determine which architecture produces the best encoding of each sample. We note that our best performing single model, TinyResNet, achieves a 92.08% accuracy on the held out test set.

1. Introduction

Determining the authorship of a written text has practical significance in the realm of forensics, signature verification, and literary history [3]. In manuscript analysis, for instance, historians frequently ask questions regarding the number of authors for a text, whether an anonymous work can be confidently attributed to a historical figure, and what time period a text might be from. These kinds of analyses are all based upon comparisons between different writing samples [1]. Techniques in the field have remained largely subjective, however, making the transition to automatic tools difficult.

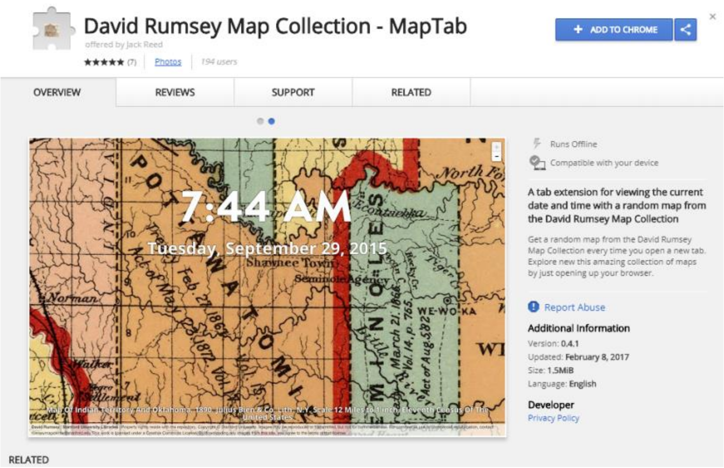
In addition, handwriting analysis is an established area of study in forensics, but there has not yet been any formal experiments measuring the accuracy of such analysis. As a result, the field is surrounded by much skepticism because of how subjective the process is (compared to, say, DNA testing) [5]. In addition, forensic handwriting analysis is time-intensive and requires two years of training for a person to obtain proper qualifications. The primary objective of this project is to develop an automatic, high-accuracy system which can determine if any two writing samples are written by the same person. In addition, our system should be able to handle authors it has never encountered before.

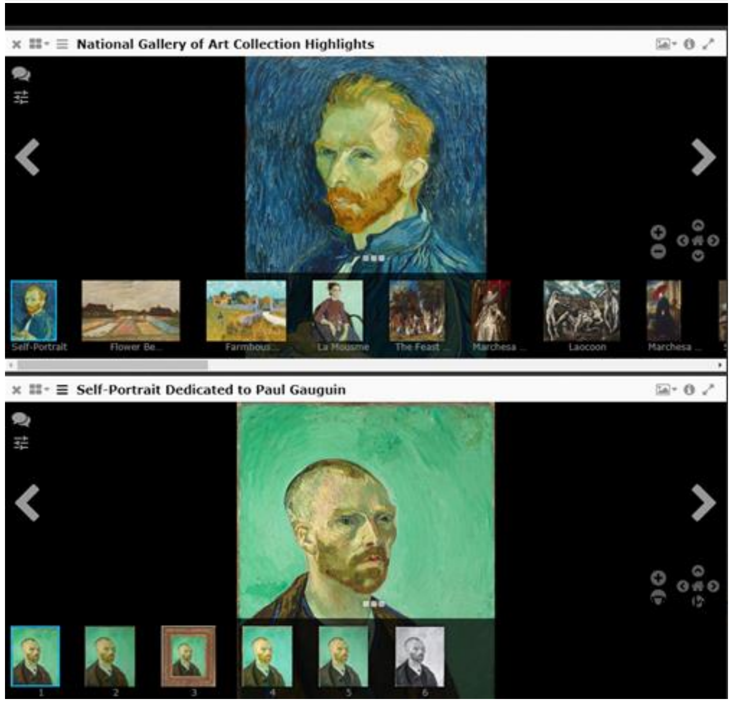
2. Background and Related Work

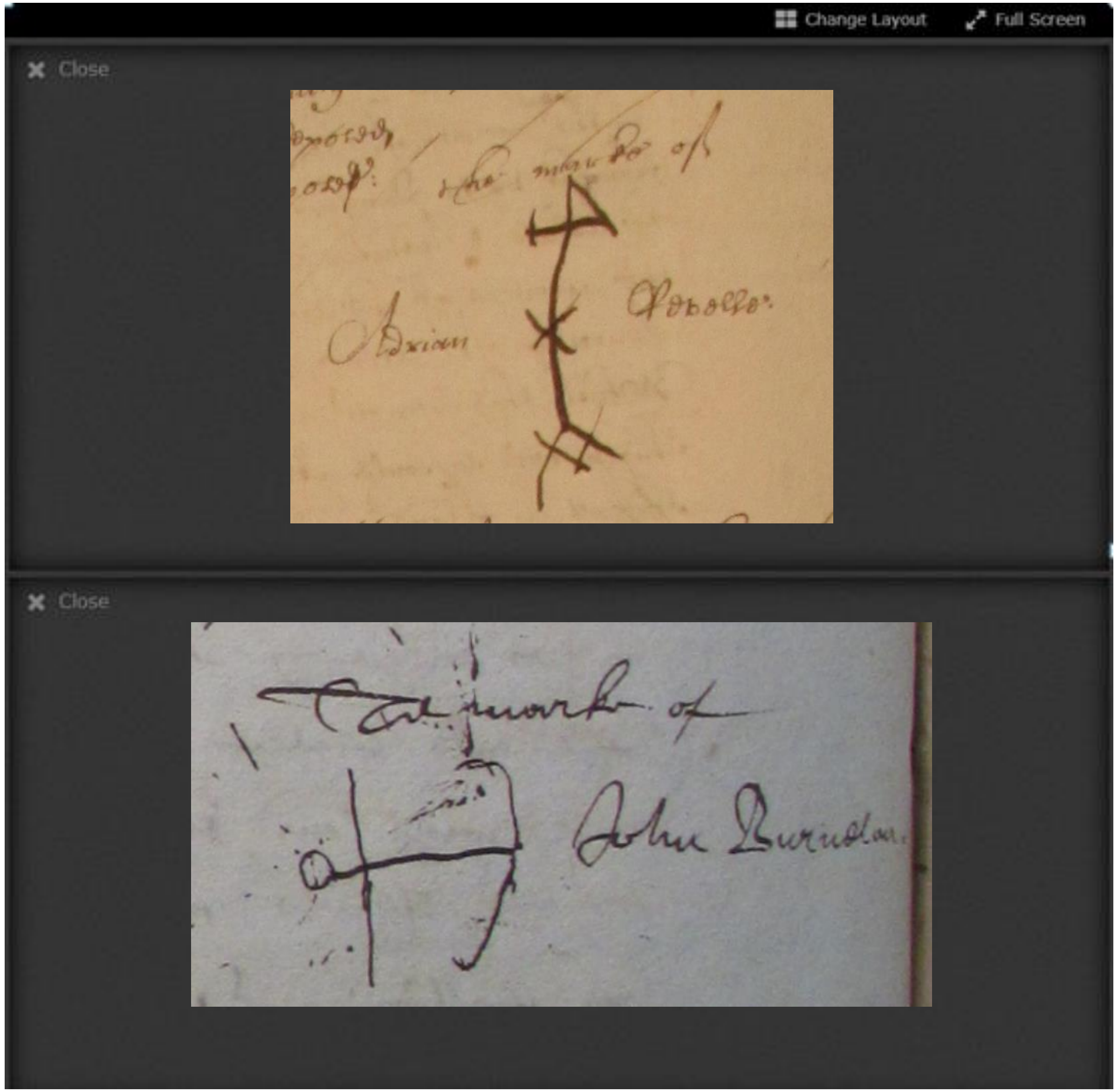
Our objective fits well with the Siamese CNN neural network architecture, which was first developed in 1993 to tackle the signature verification problem. [3] This type of architecture takes in two inputs and outputs a distance metric for the inputs. Bromley et al. was able to detect 95% of genuine signatures using this architecture. However, note that the signature verification problem expects a pair of inputs to be very similar to each other to be considered a match. This setup would not be effective for the problem we are trying to tackle, because our system should be agnostic to the actual text in a writing sample.

Other researchers have focused more closely on the authorship identification problem. A study in 2015 by Xing et al. reported an accuracy of 97% in classifying English writing samples for 657 authors. [11] They used the same dataset we will be using in this paper, the IAM Handwriting Database, and a 4-layer CNN. This study gave us confidence that we can achieve high accuracies on authorship problems using the IAM dataset. In a very recent research study from 2016, Yang et al. was able to achieve a 95% accuracy in classifying the authors for Chinese text samples,

Potential tool: conjoint analysis IIIF viewer plugin







Adrian Revelle, twenty-three year old mariner, of Dunquirke in fflanders, November 12th, 1653; “hee onely speaketh the flemmish speech” ([HCA 13/68 f.183v])

John Burnelau, twenty-eight year old sailor, of Mornar, France,March 30th, 1661 ([HCA 13/73 f.486v])

Jack Reed, David Rumsey MapTab Chrome extension; http://projectmirador.org/demo/; Conjoint analysis IIIF viewer plugin vapourware