Artificial Intelligence in Historical Document Analysis
Dhali, Maruf A.

DOI:
10.33612/diss.869247881

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2024

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 01-02-2024
Part V

DISCUSSION AND CONCLUSION

What did we achieve? What more can be done?
DISCUSSION AND CONCLUSION

The first four parts of this thesis present a digital palaeographic approach toward writer identification and dating of the DSS by employing artificial intelligence and pattern recognition techniques that can be extended to many ancient historical documents.

The thesis started by presenting the results of writer identification experiments on a limited number of scribes using various shape-based and grapheme-based features on the Brill images. The QuadHinge method performed the best in the top-1 hit lists for most cases, except for a specific scenario where CoHinge outperformed it slightly. However, the Hinge feature consistently gave better results in all the top-10 hit lists, which can be attributed to its design criteria and the incorporation of FCM following the JFD principle. On the other hand, due to the uniformity of ink traces in medieval scripts, the Quill feature proved to be a weak candidate for the writer identification task. The Junclets feature also had lower performance compared to cross-script writer identification.

This pilot work served as a baseline measurement for future experiments in the digital palaeography of DSS. The study highlighted the need for statistical modeling, transfer learning, and data augmentation to account for the diverse collection of manuscripts and varying performance results for different sets of writers. The idea to combine specifically designed shape features with deep learning methods to generate new empirical data for the study of the DSS also arose. This initial work emphasized the importance of a sophisticated character extraction technique for robust and accurate feature calculations, mainly when dealing with a larger dataset of IAA images. Traditional intensity-based methods were found to be inadequate for the diverse image types in the IAA collection, necessitating the introduction of BiNet, a new binarization technique.

The thesis then presented BiNet for efficiently binarizing degraded historical manuscript collections, explicitly focusing on the DSS. BiNet demonstrated high-performance binarization results by effectively isolating the original written content from the background, which is crucial for task identification and document dating tasks. The network’s ability to segment non-ink information from the manuscript and its robustness across diverse document textures and layouts highlight its multi-purpose usability. The work also introduced a simple and effective ground truth labeling technique and an image fusion technique that improved binarization results, allowing for better extraction of original contents. The fusion technique showed promise for application in other collections with multi-spectral images. BiNet's excellent performance on (H-)DIBCO datasets further validates its potential as a...
versatile tool for manuscript binarization. The success of BiNet paved the way for in-depth experiments on the challenging and diverse images of the DSS, including writer identification, dating, character reconstruction, and image-based material analysis.

Once a better character extraction technique is achieved, the thesis took the Great Isaiah Scroll (1QIsa\(^a\)) as a test case for writer identification; the evidence collected through various feature extraction methods supports the presence of two main scribes responsible for copying 1QIsa\(^a\), with a clear separation between their work at columns 27 and 28. The statistical tests and stress tests confirmed the robustness of the findings. The study also introduced innovative techniques for generating palaeographic charts and demonstrated using heatmaps to bridge quantitative analyses and traditional palaeography.

While acknowledging the variables of material degradation, writing implements and conditions, and limitations on character extraction, the study’s methodology remained solid and provides statistically proven probabilities for writer identification. This research represented a significant advancement in pattern recognition and artificial intelligence applied to palaeography, particularly in the context of the DSS. The findings for 1QIsa\(^a\) revealed two clusters with a statistically significant break in the middle of the manuscript, indicating the presence of two distinct writing styles by different scribes, contrary to the previous hypothesis. The similarity in handwriting between the scribes suggested a formal training or close social setting, such as a school or family context. The writer identification study highlighted the need for collaboration between palaeographers and researchers utilizing pattern recognition methods to understand ancient manuscripts’ complexities better.

Finally, the thesis focused on the dating issue. It first started by demonstrating that feature extraction techniques can be used in estimating time periods of the DSS. The reasons behind this have been examined, and potential solutions to the identified challenges have been proposed. Next, this study provides a comprehensive overview of the current state of dating the DSS and highlights the remaining problems and obstacles. Once the groundwork is done, the thesis focuses on the final date prediction model.

Enoch, the final date prediction model for the DSS, was designed by combining AI with radiocarbon dating. The study focused on collecting and preparing data, utilizing neural network architectures for preprocessing and feature extraction, and employing a Bayesian regression approach for date prediction. The study again highlighted the importance of high-quality binarized images in achieving accurate predictions and emphasized the significance of data balancing techniques to address training data bias. The model’s reliability was assessed through various training options and validation tests using unseen data. The successful development of this date prediction model demonstrated the integration of knowledge from different domains and offered promising prospects for dating undated historical documents. Furthermore, the statistical approach employed contributes to a better understanding
of historical chronology, builds trust in the models, and sets the stage for future advancements in the field.

To the best of our knowledge, Enoch is the first complete AI-based methodology for estimating the dates of manuscripts from their raw image inputs. In contrast to earlier models, Enoch uses a probability-based strategy, utilizing palaeographic input and the full probability distribution from $^{14}$C output to ensure transparency and interpretability. Due to this, Enoch can be used by palaeographers and historians to examine date predictions, facilitate decision-making, and explain qualitative palaeographic reasoning. To resolve uncertainties and lack of agreement related to exact dating, the study suggests avoiding palaeographic estimates as target values for machine learning and choosing physical and geometric evidence instead. The outcomes show that the $^{14}$C-based training does not fundamentally constrain the style-based predictions, enabling a wider time axis for predictions.

With a date prediction success rate of $67\%$, Enoch shows promise when working with completely undated documents. Better image treatment of the test data suggests the possibility of increased success rates. More text samples can improve Enoch’s accuracy even more and make it possible to calculate the effects and dangers of fresh additions to the $^{14}$C reference collection. In contrast to conventional palaeographic methods, Enoch’s approach to text analysis emphasizes shared traits and similarity matching rather than concentrating on minute stylistic distinctions. By discovering hidden patterns and improving contextual information, this interdisciplinary fusion may enhance historical interpretations and comprehension. New $^{14}$C evidence and other date-bearing manuscripts can be added to Enoch’s training data to improve refinement and precision further.

### 7.1 Addressing the Research Questions

Several research motivations were presented in Chapter 1 along with four main research themes. The research themes were focused on integrating the palaeographic knowledge and $^{14}$C data to develop AI-based methods for analyzing the DSS. This thesis tackles several challenges in addressing those research themes. In this section, we address those themes and provide some answers.

**Theme: Identifying the scribes**

Chapter 2 introduced the digital palaeography of the DSS in this thesis and presents a pilot experiment on writer-identification tests using specific features and provisional labels. We gave the results of writer identification on a limited number of scribes. We confirmed the feasibility of the writer identification test on the DSS images using several feature extraction techniques. The hinge feature gave the best result in all the top-10 hit lists for writer identification. The results highlighted the need for
a better binarization technique and statistical modeling for the diverse collection of manuscripts. The chapter served as a baseline for future experiments, which combines shape features and deep-learning methods to generate new empirical data. The initial investigation enabled the articulation of challenges in this ancient collection and possible resolutions to achieve better results.

*Theme: Enhancing handwriting*

Following the initial writer identification tests in Chapter 2, Chapter 3 explored a new binarization technique for the grayscale, color, and multi-spectral images of DSS. The degradation and damage of the DSS posed challenges for existing intensity-based binarization techniques. The new method, BiNet, addressed these issues using a novel neural network architecture based on U-Net models specifically for the DSS collection, which had proven effective for other historical manuscripts. Image fusion techniques were also explored to enhance the results using multi-spectral images. The chapter concluded with the presentation of binarization results for the entire DSS collection, providing a foundation for subsequent chapters on writer identification, dating, image-based material analysis, and character-based augmentation of the DSS.

With enhanced character extraction from BiNet, Chapter 4 presented new findings regarding a breaking point in the series of columns within the Great Isaiah Scroll without making any prior assumptions about the writer’s identity. These findings were based on point clouds obtained from the reduced-dimensionality feature space. This study’s successful implementation of the feature extraction and pattern recognition pipeline implementation provided confidence. Furthermore, it served as a motivating factor for developing the date prediction model in subsequent sections of this thesis.

In connection with the original research theme of binarization in Chapter 3, this thesis also explored tangential research of image-based material analysis presented in the appendix as Chapter 8. A good binarization technique provides the foreground (ink) and means to extract the background (parchment and papyrus in the DSS collection). This enabled various pattern recognition techniques in the background materials without any handwriting. So, as additional research, this thesis presented three different techniques for classifying the DSS images based on their material and opened up new doors for future research.

Another tangential research is presented in Chapter 9 as the final appendix to this thesis. This chapter presented the effects of data augmentation by elastic morphing on the dating of historical manuscripts. As additional research, this chapter experimented with character-level augmentation on three different datasets and presented new research opportunities.
Theme: Dual-perspective time axis

In Chapter 5, the experimental setup and results of feature extraction techniques for estimating the periods of the DSS manuscripts from the writing styles were presented. This chapter established the groundwork for developing a comprehensive date-prediction model in the following chapter. Finally, in Chapter 6, the $^{14}$C dating results and palaeographic knowledge were integrated to propose Enoch, the final date prediction model. This chapter employed the Bayesian regression technique to generate probabilities along the time axis for any sample manuscript from the DSS. As a result, a dual-perspective time axis was introduced.

Theme: Ingenuity, adaptation, interpretability, and explainability

Creativeness (handcrafting) played a vital role in the design of AI models within the scope of this thesis, requiring innovative thinking to tackle complex problems and understanding the strengths and limitations of different algorithms aided in the development of models tailored to specific tasks and data constraints. The thesis also explored statistical methods, including Bayesian statistics, to uncover underlying data patterns, estimate uncertainties, and make reliable predictions with limited data. Incorporating domain knowledge and learning from experts in other disciplines was essential for informed decision-making in model design. Additionally, the thesis considered the importance of adaptation in facilitating effective integration, collaboration, and contribution across various disciplines. The focus was always on interpretability and explainability, ensuring the systems’ transparency, reproducibility, and scalability. Such clarity allowed scholars from different domains to verify the accuracy and reliability of the results, a crucial aspect in the context of historical document analysis, particularly for the DSS. Furthermore, explainability instilled trust and confidence in the AI-based models among domain experts. Therefore, the thesis maintained close collaboration with scholars and prioritized achieving the highest level of explainability in all research endeavors.

7.2 Future Research

7.2.1 Character reconstruction

BiNet focused on the binarization aspect and maintained a conservative approach to preserving the accuracy of characters; future research can explore reconstruction-based binarization methods to address extreme degradation at the character level. Additionally, further improvements can be made to the image fusion technique by considering more than three channels and addressing border areas in larger images. The article also acknowledges the potential of more advanced network architectures
if more training data becomes available. The second sub-research question of the second theme (Tangent 2 in 1.2.2.2) can be further addressed in future works on character reconstruction.

7.2.2 Scaling up writer identification test

Extensive writer identification tests, like the one of the Great Isaiah Scroll, can be utilized for many other manuscripts. A complete FRR-FAR curve for the collection can also be developed. Most of this research focused on paleography. However, the content and context can be incorporated for more refined results. A comprehensive character recognition model can also be developed for the DSS.

7.2.3 Neural network-based dating models

While artificial neural networks were used in preprocessing and feature extraction stages for the date prediction model, high-parametric models or neural networks were not fully utilized for date prediction due to limited training data and the absence of dated manuscripts. A pretrained PNASnet has been employed to check the feasibility of a deep network in dating the DSS. Even though the results are inconclusive, the potential for exploring such models in the future was acknowledged, given the availability of more labeled data.

7.2.4 Full page text recognition and transcription

With modern large language models (LLMs) [27], future works can be done on full-page text recognition and automatic transcription using transformer-based architectures [301]. The context and content of the manuscript can be used to improve the performance.

Models like GPT-4 (Generative Pre-trained Transformer) [199] and T5 (Text-to-Text Transfer Transformer) [233] have proven their capacity to comprehend and produce human-like texts. LLMs and Transformers can also be used to process entire manuscripts to recognize and transcribe the text. The process may involve encoding the input manuscript into a sequence of tokens, which the LLM or Transformer can then process. These models efficiently capture the contextual dependencies and semantic linkages between words and sentences (essential for proper transcription) by utilizing attention mechanisms. The models can also benefit from their extensive pre-training on diverse textual data, which enables them to comprehend a wide range of linguistic idioms and adapt to multiple writing styles.
7.3 Conclusion

7.2.5 Material analysis and localization

Even though this thesis introduces image-based material analysis (in the appendix in Chapter 8) for the DSS, further experimentation is needed to incorporate the material’s information in writer identification and dating experiments. Information from non-destructive chemical analysis (e.g., Raman Spectroscopy [169]) can be incorporated into the existing image-based material analysis pipeline to improve the classification results further. Thus, the physical materials can also be localized using enhanced material analysis, providing invaluable insights into the overall writer identification and dating.

7.3 Conclusion

In conclusion, this thesis’s novel contribution to digital palaeography and historical manuscript analysis can be summarized in four themes: writer identification, enhanced preprocessing, dating, and interpretability & explainability. This thesis explores the intersection of AI and historical manuscript analysis through the study centered around the DSS. With the digital scans in hand, this thesis sheds light on the invaluable insights AI can provide in understanding ancient texts, identifying authors, and dating historical documents by exploring several machine learning and pattern recognition techniques on the DSS. Integrating radiocarbon data into a machine learning-based date prediction model will also remain a benchmark for dating using multiple information sources. Throughout the journey, this thesis continued incorporating knowledge from different domains while ensuring transparency and interpretability through its explainable design of AI and PR models. While the research of this book focused on writer identification and dating of the DSS, several future research directions are suggested, including complete page transcription, character reconstruction, material knowledge incorporation, etc.

Finally, this thesis’s interdisciplinary knowledge fusion with AI may enrich the understanding of ancient manuscripts’ writer identity, dating, material properties, and historical context, leading to enhanced interpretations of the past and opening new doors for future research.