Computational Paleography of Medieval Hebrew Scripts*

Berat Kurar-Barakat*, Daria Vasyutinsky-Shapira, Sharva Gogawale, Mohammad Suliman and Nachum Dershowitz

Tel Aviv University

Abstract

We present ongoing work as part of an international multidisciplinary project, called MiDRASH, on the computational analysis of medieval manuscripts. We focus here on clustering manuscripts written in Ashkenazi square script using a dataset of 206 pages from 59 manuscripts. Collaborating with expert paleographers, we identified ten critical features and trained a multi-label CNN, achieving high accuracy in feature prediction. This should make it possible to computationally predict the subclusters already known to paleographers and those yet to be discovered. We identified visible clusters using PCA and χ^2 feature selection. In future work, we aim to enhance feature extraction using deep learning algorithms and provide computational tools to ease paleographers' work. We plan to develop new methodologies for analyzing Hebrew scripts and refining our understanding of medieval Hebrew manuscripts.

Keywords

Medieval Hebrew manuscripts, computational paleography, convolutional neural networks, image clustering, recurrent neural networks

1. Introduction

"MIDRASH: Migrations of Textual and Scribal Traditions via Large-Scale Computational Analysis of Medieval Manuscripts in Hebrew Script," supported by an ERC Synergy grant, is an international effort to develop a revolutionary, computational approach to manuscript studies. Among other aspects, it combines traditional, digital, and computational paleographic methods to refine and potentially rewrite our understanding of Hebrew scripts, particularly their geographical variation in scribal practices [8]. The project is led by Daniel Stökl Ben Ezra (École pratique des hautes études [EPHE], Paris Sciences-Lettres University), Judith Olszowy-Schlanger (École pratique des hautes études, Paris Sciences-Lettres University and Oxford University), Nachum Dershowitz (Tel Aviv University [TAU]), and Avi Shmidman (Bar-Ilan University [BIU]), with the participation of the National Library of Israel (NLI) and Haifa University.

The main goal of the project is to develop new methodologies for studying medieval Hebrew manuscripts. In addition to employing handwriting text recognition to extract text from images, We will analyze these manuscripts using computational tools from paleographic, codicological,

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark

*Corresponding author.

berat@tauex.tau.ac.il (B. Kurar-Barakat); dariashap@tauex.tau.ac.il (D. Vasyutinsky-Shapira); sharvag@mail.tau.ac.il (S. Gogawale); suliman@mail.tau.ac.il (M. Suliman); nachum@tau.ac.il (N. Dershowitz)

https:cs.tau.ac.il/~berat (B. Kurar-Barakat)

0000-0002-7240-7286 (B. Kurar-Barakat)

linguistic, and literary perspectives. This analysis will contribute to the understanding of the manuscripts' materiality, textuality, transmission, and the historical and intellectual context of their creation and readership. By combining traditional philology with machine learning, computer vision, and computational linguistics, we will process large amounts of textual and paleographical data that traditional philology cannot handle. See Figure 1.

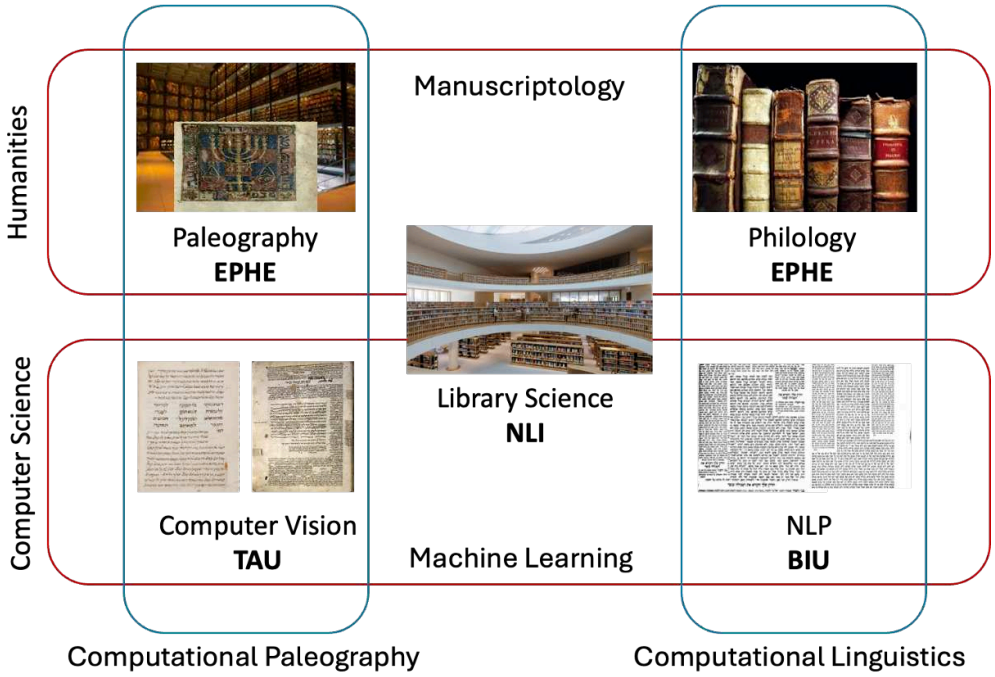

Figure 1: Synergy in computational manuscriptology.

The principal aims include:

1. Develop optical character recognition (OCR) algorithms to convert manuscript images into searchable text.
2. Implement text mining algorithms to compare a large corpus of texts and identify quotations, paraphrases, borrowings, allusions, and other intertextual relationships.
3. Train machine learning models to perform handwriting analysis and predict each manuscript's geographical and temporal origins.
4. Design natural language processing (NLP) algorithms to extract and analyze linguistic features for improved textual searches and historical context placement.
5. Integrate traditional and computational methodologies for paleographic, philological, and textual analysis.

2. Computational Paleography Tasks

Accessing manuscripts' textual and non-textual information is valuable only if we can understand the texts in their specific context of place and time. Out of all the medieval Hebrew

manuscripts, only about 3,500 are dated and have colophons (scribes' notes) or other identifying marks. Paleography (the study of handwriting) and codicology (the study of the physical aspects of books) are the primary methods used to determine the provenance of manuscripts. SfarData (sfardata.nli.org.il) is the only existing database that focuses primarily on codicology for dated medieval Hebrew manuscripts.

However, reliance on paleography is essential for a project, like ours, studying document images. The MiDRASH traditional paleography team aims to make precise regional and chronological classifications. They scan well-defined manuscript samples to find correlations between their textual features and scripts. We use HebrewPal (hebrewpalaeography.com), an ongoing effort to build a comprehensive database of Hebrew paleography. Processing this data involves synergetic collaboration between the traditional and computational paleography teams.

As the computational paleography team, we are currently working on solving the problem of finding subgroups among the Ashkenazi square script documents. Paleographers describe medieval Hebrew manuscripts according to their script mode (square, cursive, and semicursive) and geographical type. The six geographical types are Oriental (Egypt, Palestine, Syria, Lebanon, Iraq, Iran, Uzbekistan, and Bukhara, Eastern Turkey), Sephardic (the Iberian Peninsula, Provence and Languedoc, North Africa, and Sicily), Italian, Ashkenazi (France and England, the Holy Roman Empire, Central and Eastern Europe), Byzantine (Greece, the Balkans, Western Asia Minor, and regions surrounding the Black Sea), and Yemenite (Figure 2). This level of codicological classification for Hebrew manuscripts and initial automatic dating has already been successfully performed using computational means [9, 10, 5].

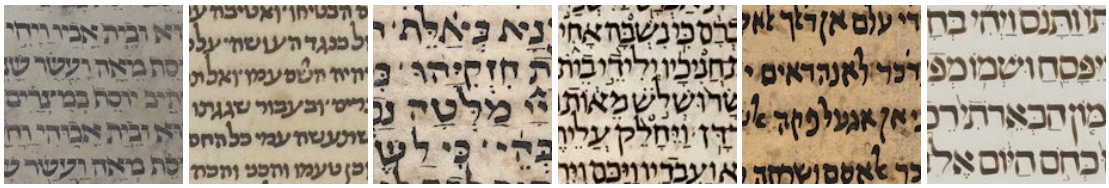

Figure 2: Medieval Hebrew script types in square mode.

Within certain script type-modes, there are distinct subclusters. Only in rare cases have these subclusters been relatively well-studied. This is, for example, the case of the Ashkenazi square script, which has been well-studied and clustered [1, 4, 3, 6]. However, even the most experienced paleographers are most familiar with the manuscripts they work with frequently, and no human memory can retain thousands of script examples. Moreover, the variations within some script type-modes are very subtle. Therefore, we are working to develop computational methods to identify clusters, and subclusters within different script types that have yet to be discovered (or those that are not discoverable) by paleographers. This work will contribute to identifying the place of copying for manuscripts of unknown provenance, more exactly than the current results [2].

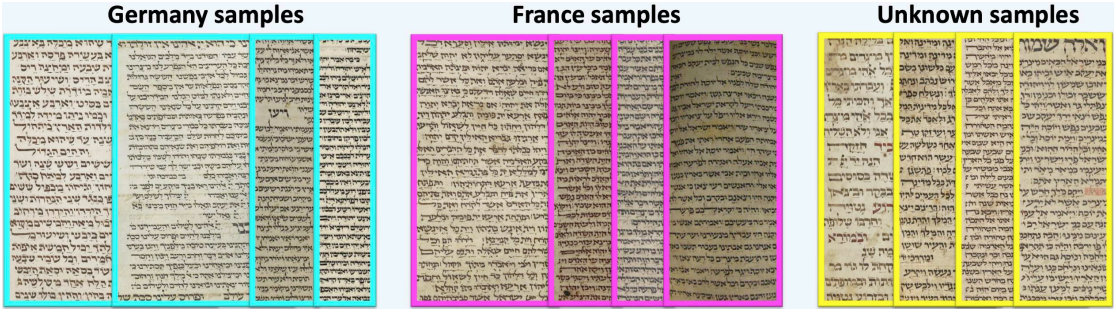

Figure 3: Sample page images from the ASC dataset.

3. Data

The Ktiv project at the National Library of Israel has led a significant digitization campaign of Hebrew-character manuscripts from collections worldwide. It has accumulated more than 80% of extant manuscripts, making tens of thousands of manuscripts accessible via a unified catalog. The Friedberg Genizah Project has contributed images and metadata of approximately 350,000 fragments from medieval book and document depositories, known as *genizot* (*geniza* or *genizah* in the singular. This digital corpus serves as the source material for our project. It includes relatively well-preserved scrolls and codices, as well as hundreds of thousands of fragments. For the clustering task, we used high-resolution pages from well-preserved manuscripts.

We are using a dataset of images built for us by Judith Olszowy-Schlanger specifically for the Ashkenazi-square clustering problem, a style for which she is the leading expert. She challenged us (as part of the MiDRASH project) with the task of automatically clustering within this specific type-mode, and potentially revealing additional subclusters as yet uncategorized by traditional Hebrew paleography. This "ASC" dataset, publicly available online at github.com/TAU-CH/midrash_ASC_dataset, contains 206 images, each depicting part of a page from 59 manuscripts, with approximately four pages from each manuscript (Table 1). It also includes an annotation file for the bounding boxes of the main text regions and text lines. Samples are unlabeled, but it is known that 17 manuscripts are from Germany and 11 are from France, while the origins of the remaining 31 manuscripts are currently unknown (Figure 3). All the manuscripts are written in Ashkenazi square script, and we aim to discover potential subclusters within these manuscripts based on their script types, which have slight variations.

Table 1
Statistics of the ASC Dataset

	Germany	France	Unknown	Total
Manuscripts	17	11	31	59
Pages	62	35	109	206
Text regions	136	61	260	457
Text lines	4413	1799	8080	14292

4. Methods and Results

Our preliminary work focused on clustering medieval manuscripts written in Ashkenazi square script using the ASC dataset. Conventional computational methods, such as the bag-of-words approach, struggle to identify the intricate features necessary for effective paleographic clustering, as the frequency of occurrence of paleographical features varies even within the same script type. To address this, we had expert paleographers identify ten critical features that they use in their analyses of this script type. The ten features identified in this way are vocalization marks, left (end of line) justification, vertical stretch, strings, short descenders, fishtails, left slant, biting, nesting, and shading (Figure 4).

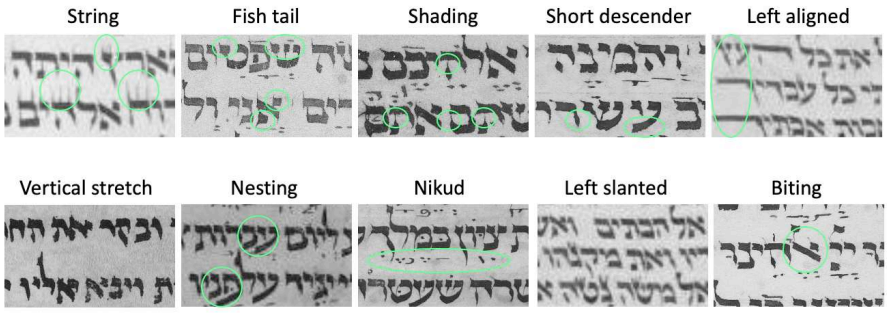

Figure 4: Figure showing all ten features identified in the dataset. Each image patch contains a specific feature, highlighted by green circles. Data annotation was facilitated using the Hasty AI assisted annotation tool (hasty.cloudfactory.com).

4.1. Predicting Paleographical Features

We trained a multi-label VGG-19 network [7] model to predict the presence of these features on a given page image. Treating this as a multi-label problem allowed us to account for the coexistence of multiple labels, as their spatial location and frequency of occurrence are not crucial for paleographical definition.

In order to prevent overfitting, we utilized the regularization approach known as early stopping. This technique stops the training process when the model's performance on the validation set stops improving, thus preventing the model from simply memorizing the training data. As a result, we ended the training when the validation loss reached 0.12, resulting in the model achieving its best validation performance (see Figure 5).

To evaluate model performance on unseen data, we split the dataset at the manuscript level (Figure 6). Splitting at this level ensures that entire manuscripts, rather than individual pages, were held out for testing. Unseen testing involves evaluating the model on data that contains patterns not seen during training. This is important for ensuring that the model can generalize well to unseen data, like in the real-world scenarios where new manuscripts are encountered.

The prediction performance on the unseen test set, as shown in the bar graph, demonstrates that our model can effectively automate the tasks performed by a paleographer, achieving accuracy levels of 98%. The performance graph (Figure 7) shows three types of F1 scores: macro

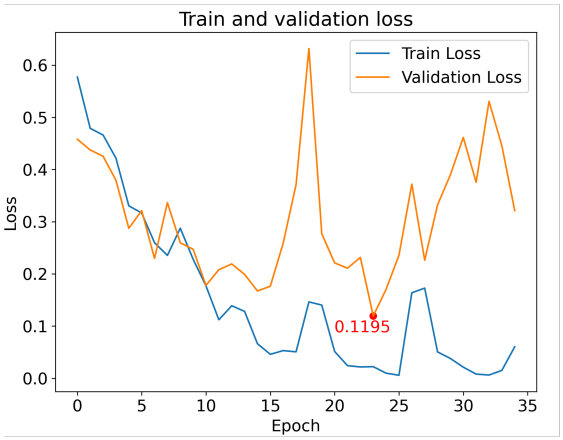

Figure 5: Training and validation loss across epochs, demonstrating the application of early stopping to achieve the model with the best validation performance when the validation loss reached 0.12.

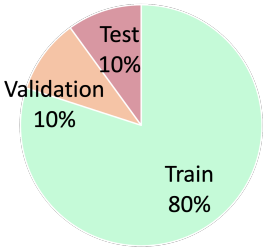

Figure 6: Pie chart showing the split percentages of the dataset at the manuscript level, used for unseen testing to ensure the model's generalization capabilities on unseen data.

average, micro average, and weighted average. The macro average F1 score calculates the F1 score for each class individually and then takes their average. It gives equal importance to all classes, regardless of their size. Therefore, every class contributes equally to the final score. The micro average F1 score combines the contributions of all classes to calculate the F1 score. The classes with larger samples influence the micro average F1 more. The weighted average F1 score calculates the F1 score for each class and then calculates the average, weighted by the number of samples in each class. This gives a balanced view by considering the size of each class, ensuring that larger classes have more influence on the final score.

During training, we monitored the prediction accuracy for each of the ten labels to identify the ease or difficulty of learning specific features (Figure 8). For instance, we observed that the "left slanted" feature took longer to learn due to its non-binary nature. It eventually achieved high accuracy because of its frequent occurrence in a single-page image. On the other hand, features such as "nesting," "shading," and "string" also took longer to learn due to their gradual values and finally resulted in lower accuracies due to their less frequent appearances.

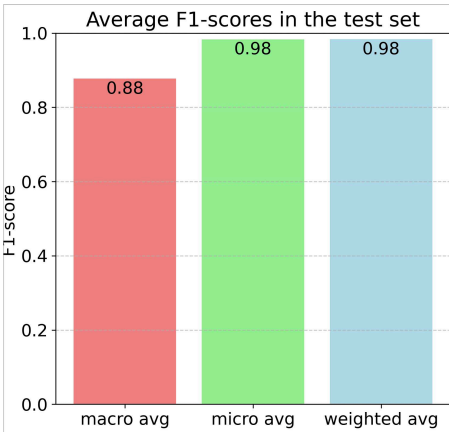

Figure 7: Bar chart showing the average F1 scores for the prediction performance on the unseen test set, demonstrating the model's effectiveness in automating a paleographer task with an accuracy level of 98%

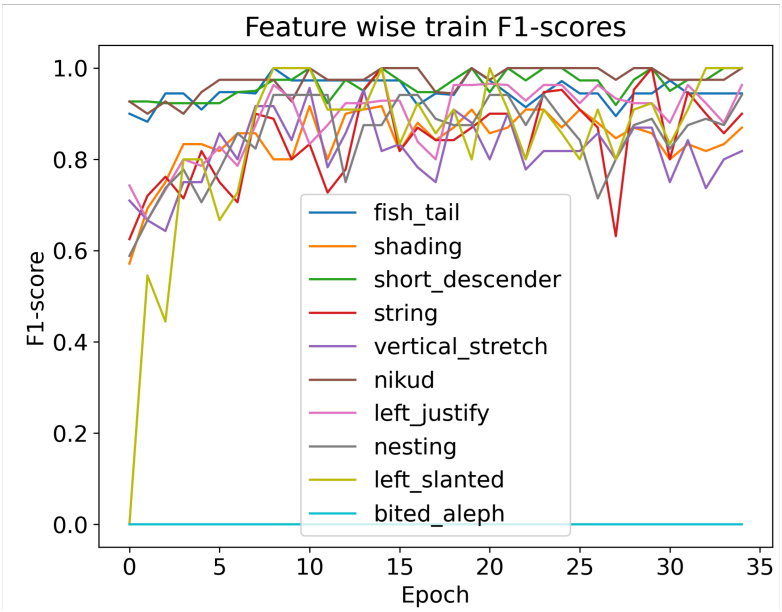

Figure 8: Feature-wise training F1 scores through epochs, showing the learning progress for each of the ten labels. The "left slanted" feature, despite taking longer to learn, eventually achieved high accuracy, while features such as "nesting," "shading," and "string" features exhibited lower accuracies due to their gradual values and less frequent occurrences.

4.2. Exploring Subclusters

To identify the subclusters, we performed a brute-force search to find the feature combinations that lead to the most cohesive subclusters. Principal component analysis (PCA) was used to visualize the samples in 2D and identify potential clusters (Figure 9). In Figure 10, you can see a sample page labeled for its regional origin by a colored frame in each cluster. We systematically

tested all features or selected features and found that χ^2 feature selection led to visible clusters. This feature selection process highlighted visible clusters based on the selected features (strings, left slanted, vertical stretch, and nesting), addressing the challenge faced by paleographers who can quickly identify individual features on a single page but struggle to simultaneously remember and analyze these features across multiple pages to discern grouping patterns. The clustering algorithm mainly successfully grouped manuscripts of known provenance and suggested some meaningful grouping of other manuscripts. One of the main challenges in computational paleography is the time and effort needed to build an initial dataset. We plan to enlarge the existing dataset and experiment on other known clusters (Oriental square and non-square; Sephardic square, non-square, and cursive, etc.) and cluster the lesser studied script types such as Yemenite. We expect this to improve the results and to significantly advance our knowledge of both human and computational paleography.

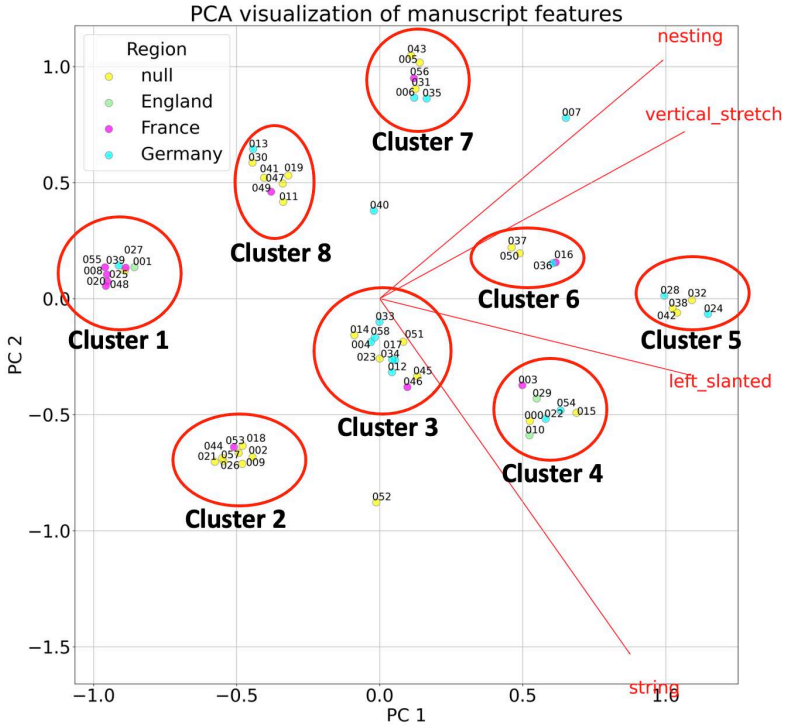

Figure 9: 2D PCA visualization of manuscripts based on χ^2 selected features, highlighting the formation of visible clusters using the identified features (strings, left slanted, vertical stretch, and nesting). Each dot is labeled with the identifier of the corresponding manuscript.

5. Conclusion and Future Work

Our approach tackles the challenge encountered by paleographers. They can quickly identify individual features on a single page but find it difficult to remember and analyze them across multiple pages to identify grouping patterns. We identified some clusters using this method

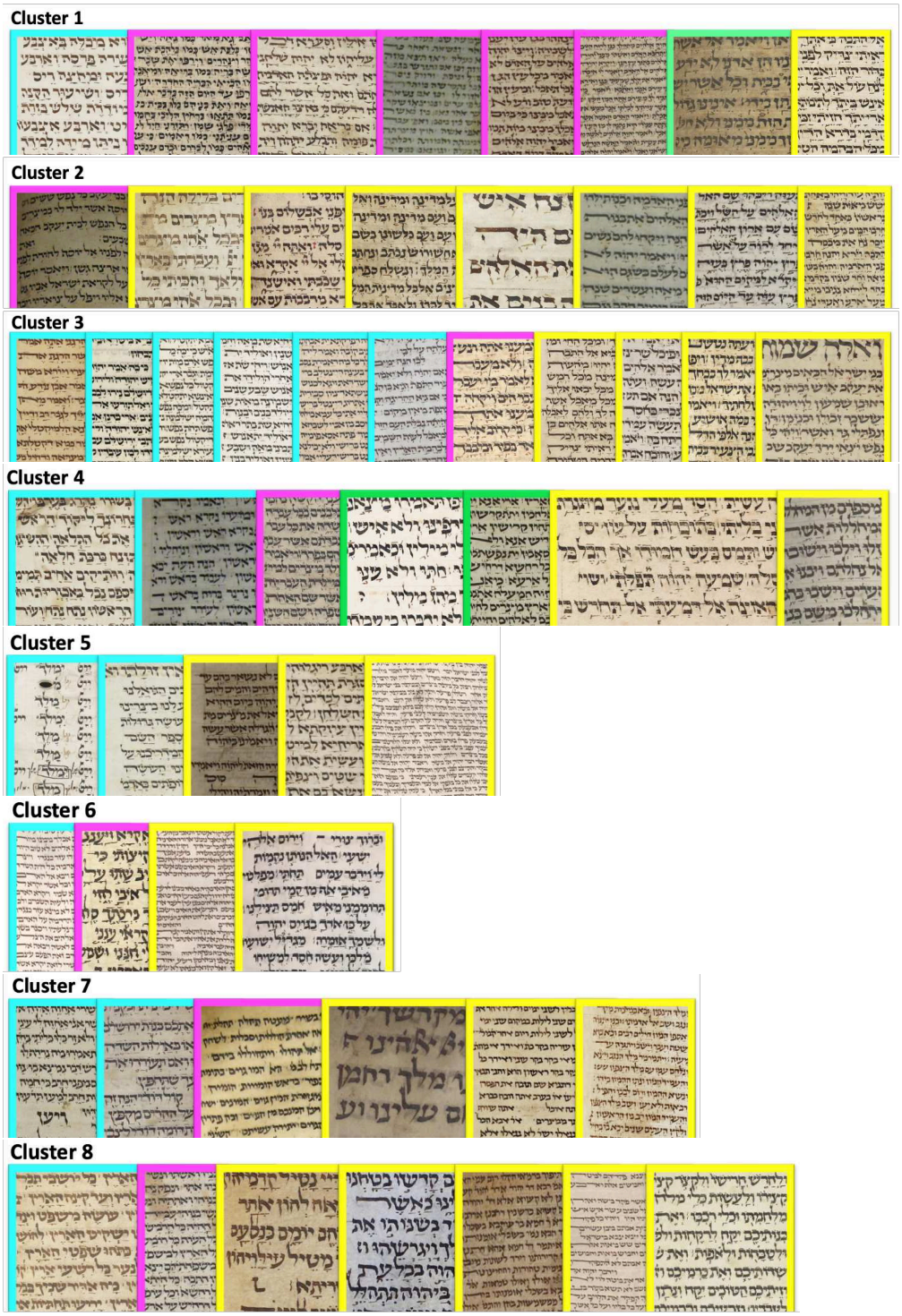

Figure 10: Sample patches from the manuscripts in each of the subclusters. Frames are color-coded: cyan for Germany, magenta for France, green for England, and yellow for unknown.

and provided insights into the paleographic features that drive these formations. Hence, we automated some of the constraints of traditional paleographic analysis.

In future work, we aim to explore methods to discover discriminative features besides those defined by paleographers. Assuming that a script type S' possesses n distinct paleographical features that are absent in a baseline script type S (Ashkenazi square script, in our case), we will train a multi-label CNN to predict the presence of all n features in images of script S', while predicting the absence of these features in images of the baseline S. We can visualize the spatial locations of these n features within the images of S' using gradient-weighted class activation mapping (Grad-CAM). This approach enables us to identify characteristics that may not be immediately apparent to human experts, furthering our understanding of these script types.

We plan to incorporate another deep learning architecture to further enhance the representation of handwriting style features. For instance, we will train a sequence-generating recurrent neural network (RNN) on the ordered sequence of contour tip points from letter strokes. The hidden state vectors from the RNN will then be used as embedding vectors, which are expected to capture stylistic features of the handwriting.

Acknowledgments

Funded by the European Union (ERC, MiDRASH, Project No. 101071829). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] M. Beit-Arie and E. Engel. *Specimens of mediaeval Hebrew scripts*. Israel Academy of Sciences and Humanities, 2017.

[2] A. Droby, I. Rabaev, D. V. Shapira, B. Kurar-Barakat, and J. El-Sana. "Digital Hebrew Paleography: Script Types and Modes". In: *Journal of Imaging* 8.5 (2022), p. 143.

[3] E. Engel. "Between France and Germany: Gothic Characteristics in Ashkenazi Script". In: *Manuscrits hébreux et arabes: Mélanges en l'honneur de Colette Sirat*. Publications de l'École Pratique des Hautes Études, 2014, pp. 197–219.

[4] E. Engel. "Calamus or Chisel: On The History of the Ashkenazic Script". In: *"Genizat Germania" – Hebrew and Aramaic Binding Fragments from Germany in Context*. Leiden, The Netherlands: Brill, 2010, pp. 183–197.

[5] B. Madi, N. Atamni, V. Tsitrinovich, D. Vasyutinsky-Shapira, J. El-Sana, and I. Rabaev. "Automated Dating of Medieval Manuscripts with a New Dataset". In: *Workshop on Computational Paleography (WCP)*. 2024, pp. 45–48.

[6] J. Olszowy-Schlanger. "The early developments of Hebrew scripts in north-western Europe". In: *Gazette du livre medieval* 63.1 (2017), pp. 1–19.

[7] K. Simonyan and A. Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *3rd International Conference on Learning Representations*. 2015, pp. 1–14.

[8] D. Vasyutinsky-Shapira, B. Kurar-Barakat, S. Gogawale, M. Suliman, and N. Dershowitz. “MiDRASH – A Project for Computational Analysis of Medieval Hebrew Manuscripts”. In: *Eurographics Workshop on Graphics and Cultural Heritage*. 2024, p. 0.

[9] L. Wolf, N. Dershowitz, L. Potikha, T. German, R. Shweka, and Y. Choueka. “Automatic paleographic exploration of Genizah manuscripts”. In: *Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age*. Norderstedt, Germany: Books on Demand, 2011, pp. 157–17.

[10] L. Wolf, L. Potikha, N. Dershowitz, R. Shweka, and Y. Choueka. “Computerized paleography: Tools for historical manuscripts”. In: *18th IEEE International Conference on Image Processing*. 2011, pp. 3545–3548.