# Clustering Ashkenazi Manuscripts

DH2024 Short Paper Abstract

Berat Barakat Mohammad Suliman Sharva Gogawale
Daria Vasyutinsky Shapira Nachum Dershowitz

December 2023

We present the initial results of ongoing research on computational paleography being conducted at Tel Aviv University as part of the international ERC Synergy project, MiDRASH.

Improving the accessibility of primary sources to researchers and the general public is essential today when, on the one hand, extensive collections of manuscripts are accessible online. On the other hand, the recent pandemic placed us in a situation where working from home is not a choice but a necessity. A deeper understanding of ancient scripts makes primary sources more available to the research community. Machine-learning algorithms allow for very exact and nuanced categorization of scripts, improving both explainability and variability of the results.

A dataset of sixty manuscripts written in the Ashkenazi square type-mode of medieval Hebrew script has been prepared by the paleography team of MiDRASH, led by Judith Olszowy-Schlanger of the EPHE.

We have experimented with several different deep-learning approaches for clustering those images. Our goal was to identify and categorize subtypes of scripts within the Ashkenazi square type-mode, one of the primary type-modes of the Hebrew script that emerged in the 12th century and is still widely used today. Subcategorization will allow for more precise identification of the likely region in where the manuscript was written and for more precise dating of undated manuscripts. There exist many thousands of Ashkenazi square manuscripts, and that we will run the algorithms on all of them eventually.

The concept of handwriting style in Hebrew paleography lacks precise definitions and is primarily limited to the recognition of a few expert paleographers. The challenge is to group an unlabeled set of documents based on script style without predefined definitions. Traditional supervised methods become impractical due to the absence of labeled examples guiding the learning process towards useful features. We suppose that handwriting style is characterized by consistent patterns and arrangements of pen strokes, transcending the necessity for formal definitions. Our method focuses on the notion that different spatial regions within a document, despite containing distinct letters, may share the same handwriting style.

We propose two approaches for document analysis and understanding, each leveraging advanced techniques to extract and interpret information from document images:

**Method I:** One of our approaches utilizes deep neural networks to extract intricate patterns from document images and transform it into high level dimensions. We focus on isolating the primary text regions within the document images and then employ deep neural networks to extract essential features. Subsequently, the obtained feature embeddings are subjected to dimensionality reduction through principal component analysis (PCA). This step condenses the high-dimensional embeddings into a more manageable and informative representation, preserving the crucial aspects of the original data. Then PCA-transformed embeddings are then fed into a clustering algorithm. To further derive meaningful insights, a clustering technique, specifically $k$-means, is applied. This allows us to group similar text regions into distinct clusters, offering a robust solution for document understanding and interpretation.

**Method II:** In this research, we also propose utilizing the latest state-of-the-art image synthesis and manipulation models based on the diffusion architecture to generate image encodings for document manuscripts. These models will facilitate the extraction of image encodings for the manuscripts, incorporating a range of features such as stylistic, textual attributes, and general-purpose attributes. These features may be localized, extracting information from specific image segments that could prove pivotal or global, encompassing the entire image. A combination of both local and global feature extraction approaches may also be explored. These encodings will serve as input for the subsequent clustering stage, where various clustering techniques will be evaluated, ranging from simple methods like $k$-means to more advanced algorithms like hierarchical clustering.

In this context, we use statistical methods to extract representatives of the most prevalent strokes in a text region, denoted as style elements. We employ these style elements to measure the pairwise stylistic similarity of handwriting in document images without relying on predefined labels. To enhance interpretability, we visualize the identified style elements on the document images, providing evidence of their presence and locations. This visualization aids in the qualitative assessment of detected stylistic similarities, empowering researchers to gain a deeper understanding of the stylistic patterns uncovered by our approach.

**Preliminary results** We are currently experimenting with clustering into two to six clusters. At each stage, the result is evaluated by a Hebrew paleographer. We expect the deep learning approach to allow a more refined and exact subcategorization of the Ashkenazi square script than the existing paleographical solutions.

# References


[1] Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "Image style transfer using convolutional neural networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[2] Caron, Mathilde, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. "Deep clustering for unsupervised learning of visual features." Proceedings of the European Conference on Computer Vision (ECCV), pp. 132-149. 2018.

[3] Chen, Ting, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A simple framework for contrastive learning of visual representations." In International Conference on Machine Learning, pp. 1597-1607. PMLR, 2020.

[4] Olszowy-Schlanger, Judith. "The early developments of Hebrew scripts in north-western Europe." Gazette du livre médiéval 63.1 (2017): 1-19.

[5] Coates, Adam, Blake Carpenter, Carl Case, Sanjeev Satheesh, Bipin Suresh, Tao Wang, David J. Wu, and Andrew Y. Ng. "Text detection and character recognition in scene images with unsupervised feature learning." In 2011 International Conference on Document Analysis and Recognition, pp. 440-445. IEEE, 2011.

[6] Malachi Beit-Arie, Edna Engel. Specimens of Mediaeval Hebrew Scripts, Vol. III: Ashkenazic Script. The Israeli Academy of Sciences and Humanities, 2017.