Automatic Clustering of Hebrew Manuscripts

DARIAH2024 Paper Abstract

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

February 11, 2024

Abstract

This paper presents the interdisciplinary research conducted at Tel Aviv University in the framework of the ERC Synergy project, MiDRASH. Our work combines scholarly domains of Hebrew paleography and deep machine learning. We aim to automatically cluster medieval Hebrew script types-modes beyond the limits of contemporary human paleography. Currently, we are working on Ashkenazi square script. Successful algorithms will be applied to other medieval Hebrew script type-modes, allowing improved clustering of the least-studied script types, such as Byzantine and Yemenite, and deepening our understanding of the sub-clustering of Oriental, Sephardic, and Italian scripts. This, in turn, will lead to the discovery of new paleographic patterns, improved layout segmentation based on script types, and more.

We begin with a dataset of single pages from 55 manuscripts in Ashkenazi square produced by the project's EPHE team. Beyond the general definition of the script type-mode, the dataset is not labeled with either the date or exact region of copying. Current paleographic research suggests that Ashkenazi square is divided into French and German groups, with stand-alone manuscripts produced in England (before 1290) [1]. In this innovative research, we aim to verify if the automatic clustering corresponds to the existing paleography theory and to try to find more subclusters based either on time or place of copying.

So far, we have used several methods for automatic clustering. Following the method outlined in [2, 3, 4], we employed SIFT and Bag of Visual Words (BOVW) to represent the text region images. Subsequently, we created a 2D plot of the text regions using PCA, allowing the paleographer to visually verify the presence of any meaningful clusters (Figure 1). Based on paleographers' intuition that the text regions may construct clusters by their global patterns, we used Local Binary Pattern (LBP) to represent the text region images and we created a 2D plot of the text regions using PCA, allowing the paleographer to visually verify the presence of any meaningful clusters (Figure 2).

Since the conventional algorithms were not successful at detecting the features that lead to paleographical clusters, we decided to experiment

with a more deterministic way. A paleographer defined a set of high-priority paleographical features and we designed these as a hierarchical multi-labeling classification problem that is a hierarchical tagging with multi-labels and mutually exclusive sub-labels.

We are developing an algorithm to measure the similarity of text region images by leveraging style features. Our approach employs ControlNet [5], a neural network architecture designed to incorporate spatial conditioning controls into diffusion models [6]. We suggest incorporating graph representation [7] for handwritten text as an additional control input. This representation encodes letters similarly to human pose labels and can be computed automatically. By training ControlNet with the control input of handwriting's graph representation, we anticipate that the model will learn style features and facilitate the measurement of paleographic similarity in text region images.

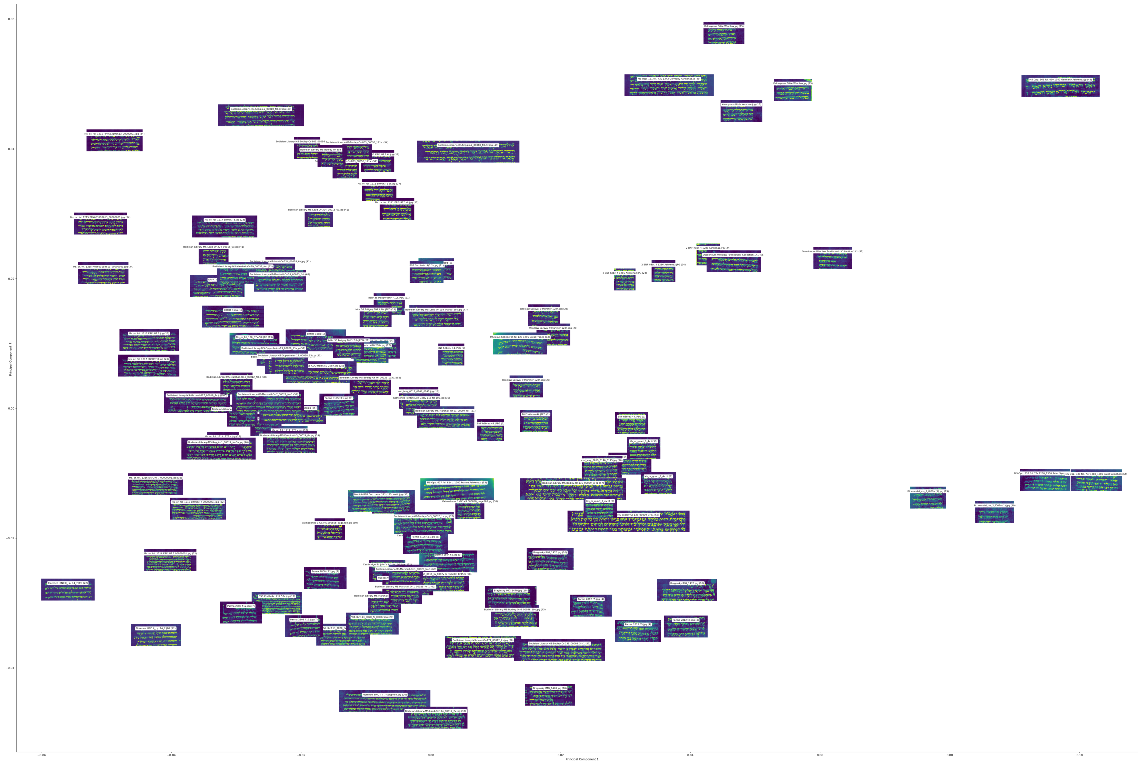

Figure 1: Visualization of main text regions in 2D plot by projecting dimensionality-reduced SIFT+BOVW features.

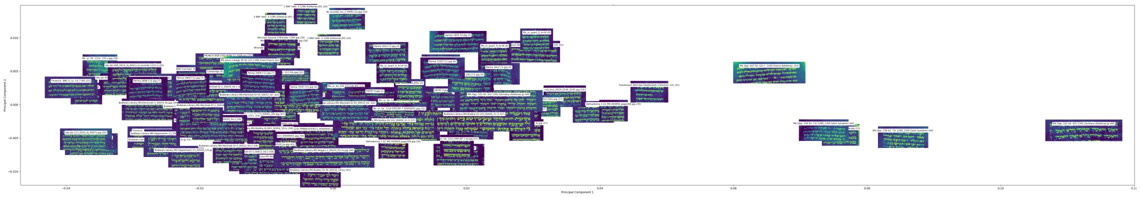

Figure 2: Visualization of main text regions in 2D plot by projecting dimensionality-reduced LBP features.

References

[1] J. Olszowy-Schlanger, "The early developments of Hebrew scripts in north-western Europe," *Gazette du livre médiéval*, vol. 63, no. 1, pp. 1–19, 2017.

[2] L. Wolf, N. Dershowitz, L. Potikha, T. German, R. Shweka, and Y. Choueka, "Automatic palaeographic exploration of Genizah manuscripts," in *Kodikologie und Paläographie im Digitalen Zeitalter 2 – Codicology and Palaeography in the Digital Age 2*, ser. Schriften des Instituts für Dokumentologie und Editorik, F. Fischer, C. Fritze, and G. Vogeler, Eds. Norderstedt, 2011, vol. 3, pp. 157–179.

[3] L. Wolf, L. Potikha, N. Dershowitz, R. Shweka, and Y. Choueka, "Computerized paleography: tools for historical manuscripts," in *2011 18th IEEE International Conference on Image Processing*. IEEE, 2011, pp. 3545–3548.

[4] L. Wolf, R. Littman, N. Mayer, T. German, N. Dershowitz, R. Shweka, and Y. Choueka, "Identifying join candidates in the Cairo Genizah," *International Journal of Computer Vision*, vol. 94, pp. 118–135, 2011.

[5] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[7] M. Stauffer, A. Fischer, and K. Riesen, "Graph-based keyword spotting in historical handwritten documents," in *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, S+ SSPR 2016, Mérida, Mexico, November 29-December 2, 2016, Proceedings*. Springer, 2016, pp. 564–573.