Human vs. Computational Paleography: a case of Ms. Hunt. 200*

Berat Kurar-Barakat Daria Vasyutinsky Shapira
Mohammad Suliman Sharva Gogawale
Omer Ventura Gal Grudka Nachum Dershowitz

This research is conducted at Tel Aviv University as part of the "MiDRASH" ERC Synergy project [1]. One of our current goals is to develop algorithms and methods to improve and automatize traditional Hebrew paleography. One of the problems we are trying to solve is clustering within the known script type-modes [2, 3] and differentiating between scribes within the same type-mode. Here we will present the results of our work on a paleographically very interesting manuscript, Huntington 200, from the library of Corpus Cristi College in Oxford. The manuscript, which was produced in Egypt in 1279, is scribed in Oriental non-square (semi-square/semi-cursive) script by a number of scribes. SfarData[1] indicates eleven hands (scribes) in this manuscript. Judith Olszowy-Schlanger, in her paper dedicated to this manuscript [4], says that the manuscript was copied by six different scribes who worked in a kind of workshop. We are trying to find a computational solution to confirm or challenge this data.

Our approach to analyzing paleographical similarity is based on letter-level rather than page-level analysis. Traditional methods, such as the bag of words model, face challenges in paleography because they assume a consistent number of features per object, which does not hold for handwriting at page level. Since we aim to make letter-level comparisons, we face the challenge of letter detection. Training a machine learning model is not feasible due to the additional challenge of data annotation. Therefore, we adopt an approach inspired by how a nonreader of a text would compare handwriting: by looking at repeating shapes regardless of their identity.

We begin by detecting all the connected components on a manuscript page. We then perform pairwise Scale-Invariant Feature Transform (SIFT) matching among all component combinations. A SIFT match is defined as having at least m feature matches between two components, each satisfying the Lowe ratio test with a threshold r. Using a union-find algorithm, we group the matched com-

*This research was funded in part 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.

[1] http://sfardata.nli.org.il/

ponents into g groups, where each group contains c similar-shaped components.

We then measure the number of similar components (s) between two pages of the manuscript as a measure of paleographical similarity. While a higher s value suggests stronger similarity between pages, it does not necessarily imply the same scribe, as certain common shapes may appear frequently by chance. Our goal is to refine this metric to account for scribal attributions.

An example of our method's results is illustrated in Figure 1, showing clusters of similar components extracted from a single page. Additionally, in Figure 2, our algorithm provides evidence supporting the findings of Judith Olszowy-Schlanger, demonstrating that specific letters are written in the same way between two pages, contrary to the attribution in SfarData.

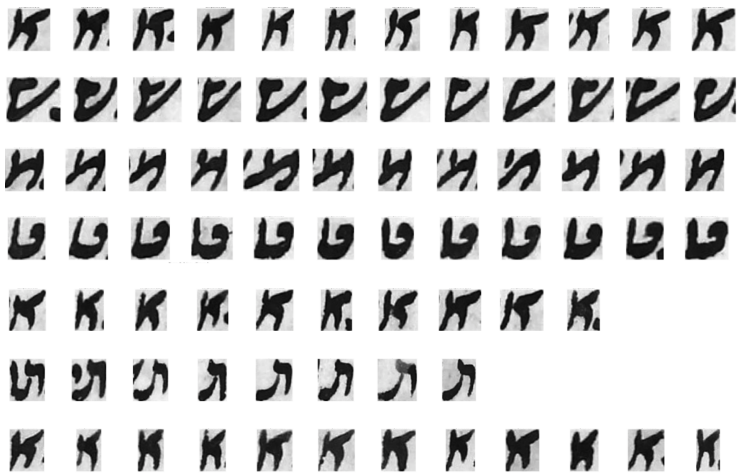

Figure 1: Groups of similar components detected within a single manuscript page.

Figure 2: Similar letters found between Bodleian-Library-MS-Huntington-200_00021_9r (scribe 1) and Bodleian-Library-MS-Huntington-200_00015_6r (scribe 2), providing evidence that challenges the attribution in SfarData.

We are continuing this experiment now; the final goal is to identify the number of scribes in the whole manuscript. This method can be further applied to other manuscripts in the Oriental non-square script (first and foremost, Genizah fragments) to identify individual scribes and distinguish between hands.

We assume this method will also produce great results with other non-square medieval Hebrew scripts.

References

[1] 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*, M. Corsini, D. Ferdani, A. Kuijper, and H. Kutlu, Eds., 2024. [Online]. Available: https://diglib.eg.org/server/api/core/bitstreams/e3fbf6c8-a64d-4bdd-a10e-70eb89b0c4bd/content

[2] B. Kurar-Barakat, D. Vasyutinsky-Shapira, S. Gogawale, M. Suliman, and N. Dershowitz, "Computational paleography of medieval hebrew scripts," 2024.

[3] D. V. Shapira, B. Kurar-Barakat, M. Suliman, S. Gogawale, and N. Dershowitz, "Automatic clustering of hebrew manuscripts," 2024.

[4] J. Olszowy-Schlanger, "User-production of hebrew manuscripts revisited: the case of manuscript oxford, bodleian library, huntington 200," *Personal Manuscripts: Copying, Drafting, Taking Notes*, vol. 30, p. 335, 2023.