# Midrash Ashkenazi Square script clustering progress report

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

Tel Aviv University

12.12.2023

## 1 Problem definition

We are tasked with clustering Ashkenazi Square scripts. Essentially, we are aware of two groups, German and French scripts ([4]). However, no labels are provided. The problem is entirely unsupervised, and the output needs to be evidence-based, providing the user with the reasons that lead to a specific clustering.

## 2 Dataset

The dataset comprises 55 document images provided by Judith Olszowy-Schlanger. The images exhibit heterogeneity in terms of layout, resolution, brightness, and contrast (Figure 1). Assuming that the script of interest is not present in the side texts, we detected the main text regions and cropped them. A document image may contain multiple main text regions; thus, we had a total of 118 main text regions, where several text regions may contain the same script.

Subsequently, we calculated the average component height and utilized it as a reference value to standardize resolutions. Finally, we converted the images to grayscale and adjusted the brightness and contrast of the text regions (Figure 2).

## 3 First method

### 3.1 Method

We conducted experiments employing the Unsupervised Feature Learning using K-means (UFLK) method ([5]). Patches were randomly sampled from the normalized main text regions, and K-means was employed to extract centroids for the codebook (Figure 3).

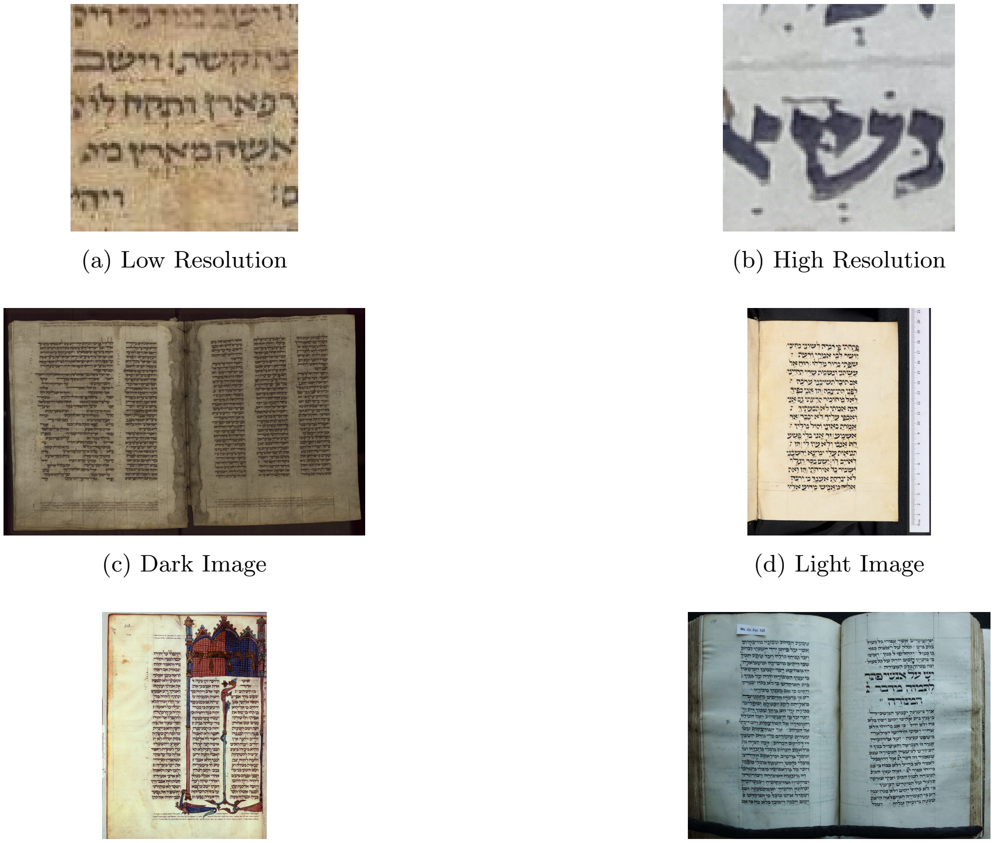
### 3.2 Visualization

Subsequently, we computed the normalized occurrence frequency of these centroids for each main text region, forming the feature vector of the text region. We then reduced the dimensionality of the feature vectors and plotted the text regions in 2D to observe potential groupings (Figure 4).

To provide evidence for the user, we also visualized the regions that highly activate a given centroid (Figure 5).

### 3.3 Limitations

However, we find the results obtained using UFLK to be less useful due to the dataset comprising multi-domain samples, where each document is gathered from a different manuscript and, consequently, from a different domain (Figure 1). As expected, UFLK would typically extract centroids representing the average of the most prevalent strokes. However, we observed that the extracted centroids did not capture the necessary complexity to represent the fundamental building blocks of script strokes across all documents from various domains.



(a) Low Resolution

(b) High Resolution

(c) Dark Image

(d) Light Image

(e) Three columns layout

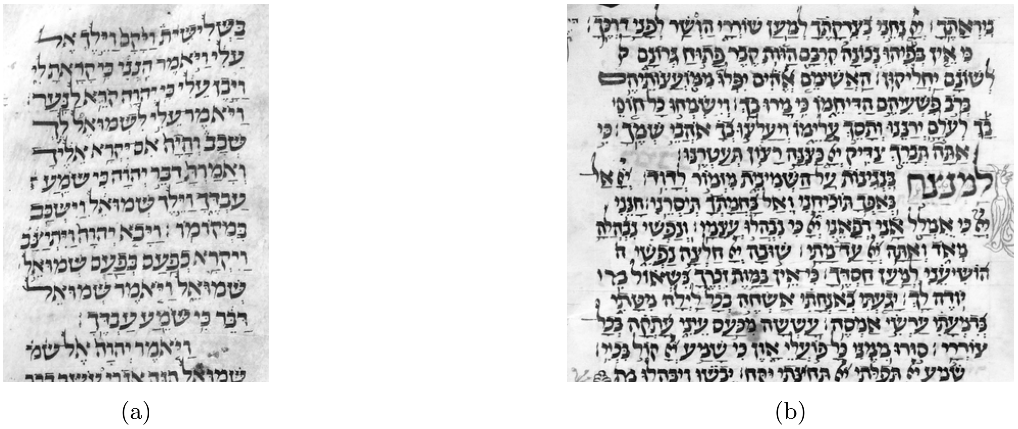
(f) Single column layout

Fig. 1: Heterogeneity of the document images in terms of resolution (a, b), brightness, and contrast (c, d), and layout (e, f)."

### 3.4 Future directions

Our emphasis lies in the idea that various spatial regions within a document, while comprising different letters, might exhibit the same handwriting style. Hence we suppose that the handwriting style is characterized by consistent patterns and arrangements of pen strokes, eliminating the need for formal definitions. Let's refer to these consistent patterns and arrangements as style elements. Our intention is to utilize these style elements to measure the pairwise stylistic similarity of handwriting in document images without relying on predefined labels. Therefore, we plan to experiment with four methods to represent these style elements:

- Conventional filters (SIFT, Gabor, etc)
- CNN with gram matrix ([2])
- Bilinear CNN features ([3])
- Texture filters with CNN ([1])



(a)

(b)

Fig. 2: Cropped main text regions (a, b) from two different documents, resized to an average character height of 15 pixels, with adjusted brightness and contrast.

## 4 Second method

In general, the direction is based on trying to use the latest state-of-the-art models for image generation and manipulation, like the backbone model used in the trendy system of dal-e, where you input a text prompt, and the model synthesizes an image based on it.

We started by thinking simple, where the images in the dataset was encoded into compact representations, by the so called “diffusion” model mentioned before, so it can be clustered in the next stage by a classical clustering algorithm in unsupervised Machine Learning, which is the setting where the dataset isn’t annotated by a human expert, like in our case for this part of experiments.

The results weren’t good enough to impress Daria, so we decided to move on to the second phase of our experiments. We are considering the following ideas right now:

– Based on a recent article published in a top tier computer vision conference, we want to adapt their way of representing images, which can learn to focus on the important part of the image leading it to be categorized under some cluster over the another. For example, if an image contains a letter written in a specific style, the representation of it, using this methodology, could account for this decisive piece of information, so it can be put under the correct cluster by the clustering algorithm in the next stage.
– We also plan to use those representations as an input for the vanilla clustering algorithm used in the first trial as a base line maybe, but we plan to incorporate the usage of another algorithms, like the ones which might consider the hierarchical nature of the clusters of our dataset, as we come to know it’s the case by a note from Judith.

## 5 Third method

This approach 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. We experiment with networks like VGGs, ResNets etc help us concentrate on minute details within images, emphasizing smaller receptive fields. Subsequently, the obtained feature embeddings are subjected to dimensionality reduction through principal component analysis (PCA). This step condenses the high-dimensional

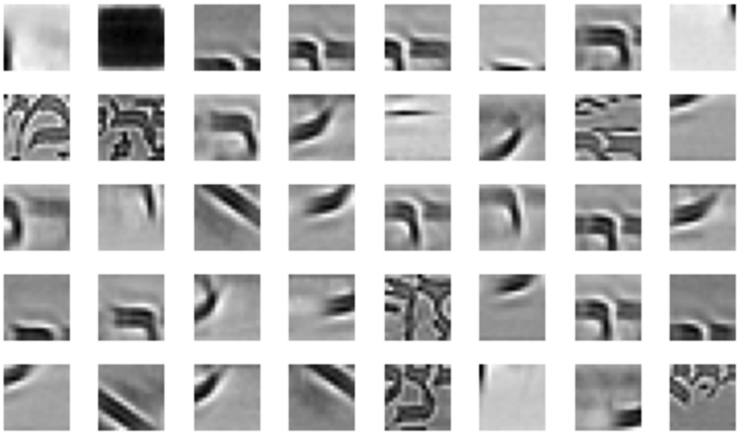


Fig. 3: A set of centroids extracted using the UFLK method.

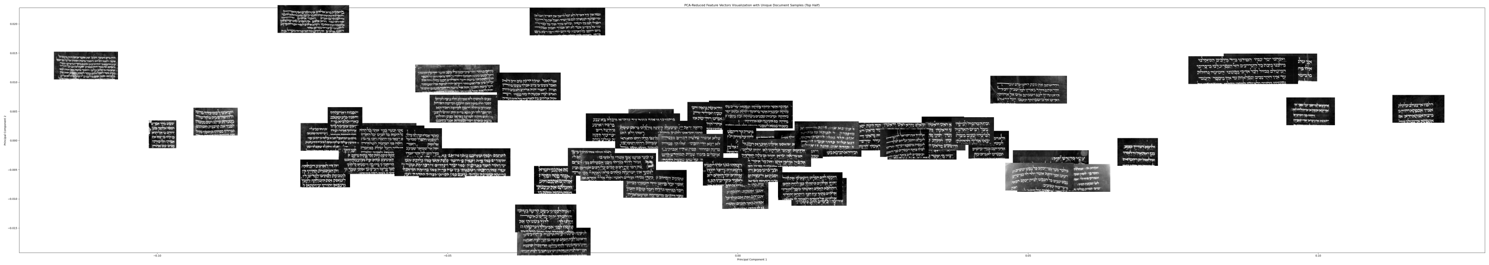


Fig. 4: Visualization of main text regions in 2D plot by projecting dimensionality-reduced UFLK features.

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.

## References

1. Cimpoi, M., Maji, S., Vedaldi, A.: Deep filter banks for texture recognition and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3828–3836 (2015)
2. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2414–2423 (2016)
3. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear cnn models for fine-grained visual recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 1449–1457 (2015)
4. Olszowy-Schlanger, J.: The early developments of hebrew scripts in north-western europe. Gazette du livre médiéval **63**(1), 1–19 (2017)
5. Wang, T., Wu, D.J., Coates, A., Ng, A.Y.: End-to-end text recognition with convolutional neural networks. In: Proceedings of the 21st international conference on pattern recognition (ICPR2012). pp. 3304–3308. IEEE (2012)

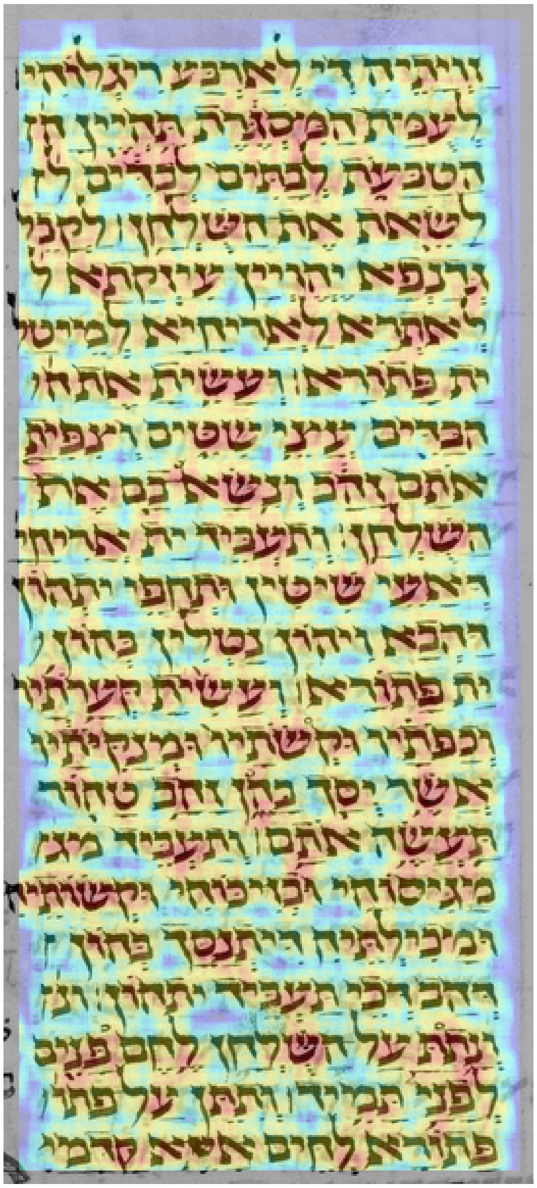


Fig. 5: Visualization of regions highly activating a specific centroid, serving as evidence for the user.