Midrash TAU group progress report 29.01.2024

Our abstract about page layout was accepted to Magazen (Venice), we will finalize the paper by mid-March.
Submitted abstract to DH2024 in Potsdam.
Submitted abstract to DH2024 in Washington (online). Daria serves as reviewer for this conference.

1 Ashkenazi square clustering

1.1 Conventional feature descriptors

Previously, we conducted experiments using UFLK (Unsupervised Feature Learning using K-means) and found it to be less effective. This was attributed to the dataset containing multi-domain samples, where each document is sourced from a different manuscript and, consequently, from a distinct domain. As anticipated, UFLK typically extracts centroids representing the average of the most prevalent strokes. However, we observed that these centroids failed to capture the necessary complexity required to represent the fundamental building blocks of script strokes across all documents from various domains.

In light of this, we hypothesized that different spatial regions within a document, even comprising different letters, might exhibit the same handwriting style. Thus, we posit that handwriting style is characterized by consistent patterns and arrangements of pen strokes, eliminating the need for formal definitions. We refer to these consistent patterns and arrangements as style elements. We employed conventional feature descriptors to represent the handwriting style in a single document image without relying on predefined labels.

SIFT and BOVW Following the method outlined in [1,3,2], 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). Although the samples from the same manuscript appeared at closer proximity,there were no clusters in terms of paleography.

We visualized the centroids (Figure 2) and their respective spatial locations on the text regions to present evidence to the paleographer (Figure 3). However, since no meaningful clusters emerged, the visualization was ineffective in providing evidence. Through this visualization, we discovered a critical fact: theoretically, the BOVW method cannot adequately represent paleographical features as required. This limitation stems from the fact that BOVW was developed in Computer Vision to represent object features. In natural images, objects often display a consistent number of features; for instance, a cat consistently has two ears. In contrast, in paleography, a text region may exhibit a particular characteristic with varying occurrences. For example, a text region may contain a variable number of bited alephs.

SIFT and FVE We employed SIFT and Fisher Vector Encoding (FVE) 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 4). Although the samples from the same manuscript appeared at closer proximity,there were no clusters in terms of paleography.

Distances in between the letters and the words The paleographers had the intuition that letters and words in some text regions are closer to each other than those in other text regions. They suggested that text regions could be clustered based on the distances between letters and the distances between words. Consequently, we plotted the samples according to the distances between letters versus the distances between words.

For each text region, we used the projection profile to detect text lines. We considered the first n profiles with the maximum sum, leading to lines passing through the top part of letters. Additionally, we employed run length to determine

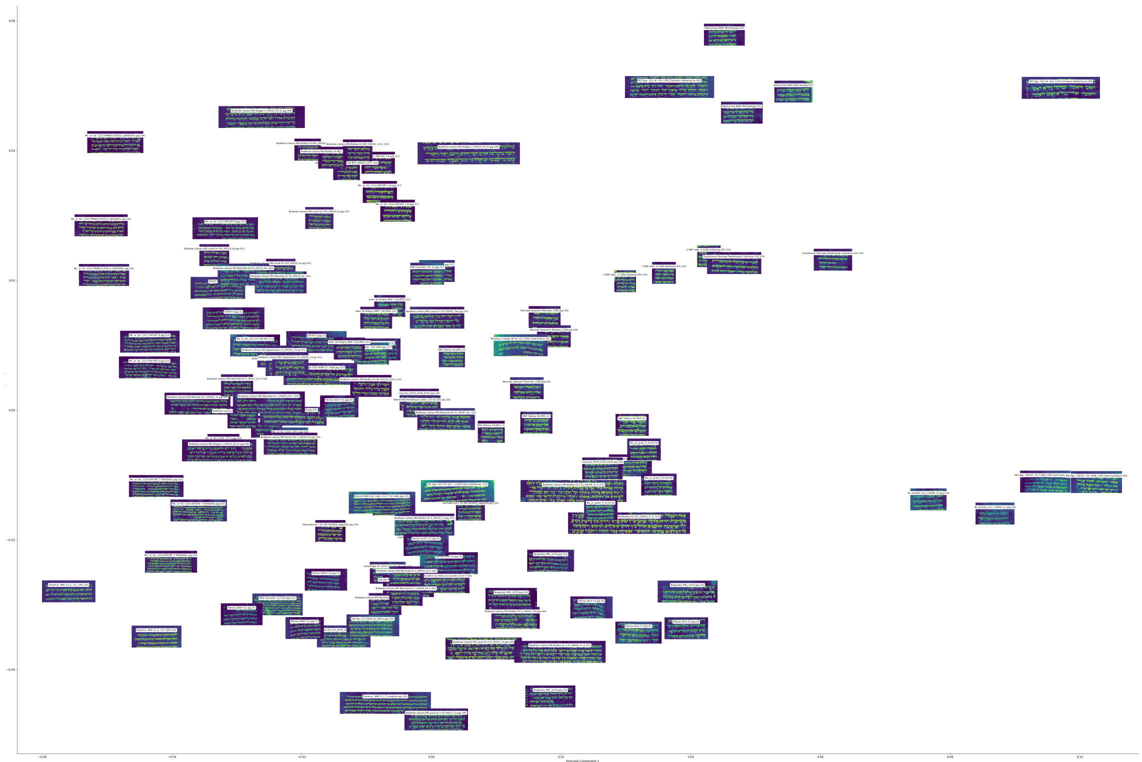

Fig. 1: Visualization of main text regions in 2D plot by projecting dimensionality-reduced SIFT+BOVW features.

the distribution of consecutive background pixels, focusing on run length values less than 20, as the average letter size is 15. We assumed that the average of the most frequent two values represents the distance between letters, and the average of the most frequent third, fourth, and fifth values represents the distance between words.

We observed that all the text regions have nearly identical distances between their letters and words (Figure 5).

Local Binary Pattern (LBP) The paleographers had the intuition that the text regions may construct clusters by their global patterns. Hence we used Local Binarry Pattern (LBP) to 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 6).

1.2 Hierarchical multi-label classification

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. Daria is working to add pages to the existing dataset to make it suitable for training a classifier. She 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 shown in the following list:

– **Label: nikud**

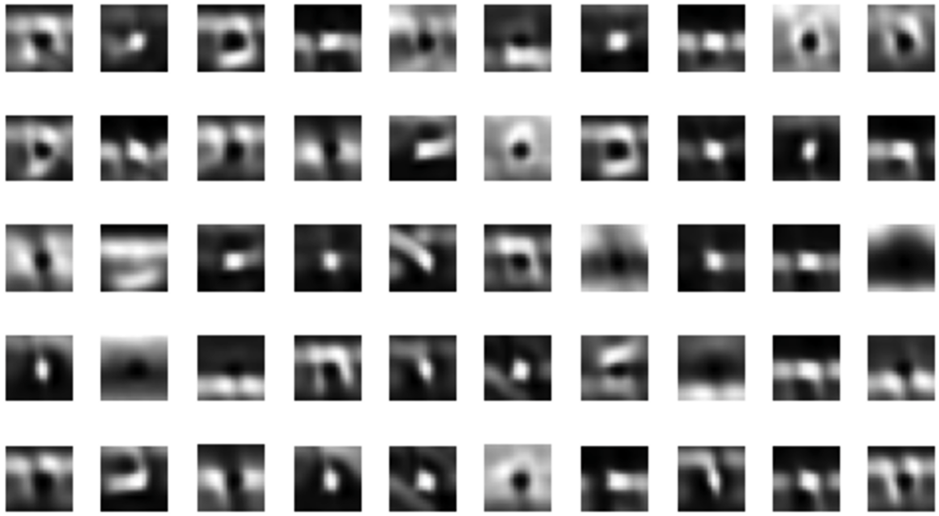

Fig. 2: Visualization of visual words extracted using the SIFT+BOVW method.

  - nikud_0
  - nikud_1
- **Label: bited_aleph**
  - bited_aleph_0
  - bited_aleph_1
- **Label: fish_tail**
  - fish_tail_1
  - fish_tail_2
- **Label: short_descender**
  - short_descender_0
  - short_descender_1
- **Label: shading**
  - shading_0
  - shading_1
- **Label: string**
  - string_1
  - string_2
- **Label: left_justify**
  - left_justify_0
  - left_justify_1
  - left_justify_2
- **Label: vertical_stretch**
  - vertical_stretched_0
  - Vertical_stretched_1
- **Label: left_slanted**
  - left_slanted_0

Fig. 3: Visualization of corresponding spatial locations for the 25th centroid of SIFT+BOVW method on three different text regions.

  - left_slanted_1
- **Label: nesting**
  - nesting_0
  - nesting_1
  - nesting_2

Once we train a machine that can predict n paleographical labels for a given text region, we plan to use all possible combinations of $1, 2, 3, ..., n$ features and find the optimal combinations that lead to the most cohesive clusters.

1.3 Text line detection

Nachum, Lior, and Mohammad proposed that the handwriting style could be discoverable at the text line level. Consequently, we enriched the dataset by incorporating the spatial locations of text lines in a coco-json format file. This file contains bounding box coordinates of the text lines within each cropped main text region. These bounding boxes were predicted using a Faster-RCNN model trained on manually labeled samples (Figure 8).

To access the code for this algorithm, along manually labeled training data, prediction results in COCO JSON format, and visualized bounding boxes on text regions, please visit the GitHub repository: TAU-CH midrash_textline_detection Repository.

2 Layout classification

We are ready to start working on connected pages reading order.

References

1. Wolf, L., Dershowitz, N., Potikha, L., German, T., Shweka, R., Choueka, Y.: Automatic palaeographic exploration of Genizah manuscripts, vol. 3. Books on Demand (BoD) (2011)
2. Wolf, L., Littman, R., Mayer, N., German, T., Dershowitz, N., Shweka, R., Choueka, Y.: Identifying join candidates in the cairo genizah. International Journal of Computer Vision **94**, 118–135 (2011)
3. Wolf, L., Potikha, L., Dershowitz, N., Shweka, R., Choueka, Y.: Computerized paleography: tools for historical manuscripts. In: 2011 18th IEEE International Conference on Image Processing. pp. 3545–3548. IEEE (2011)

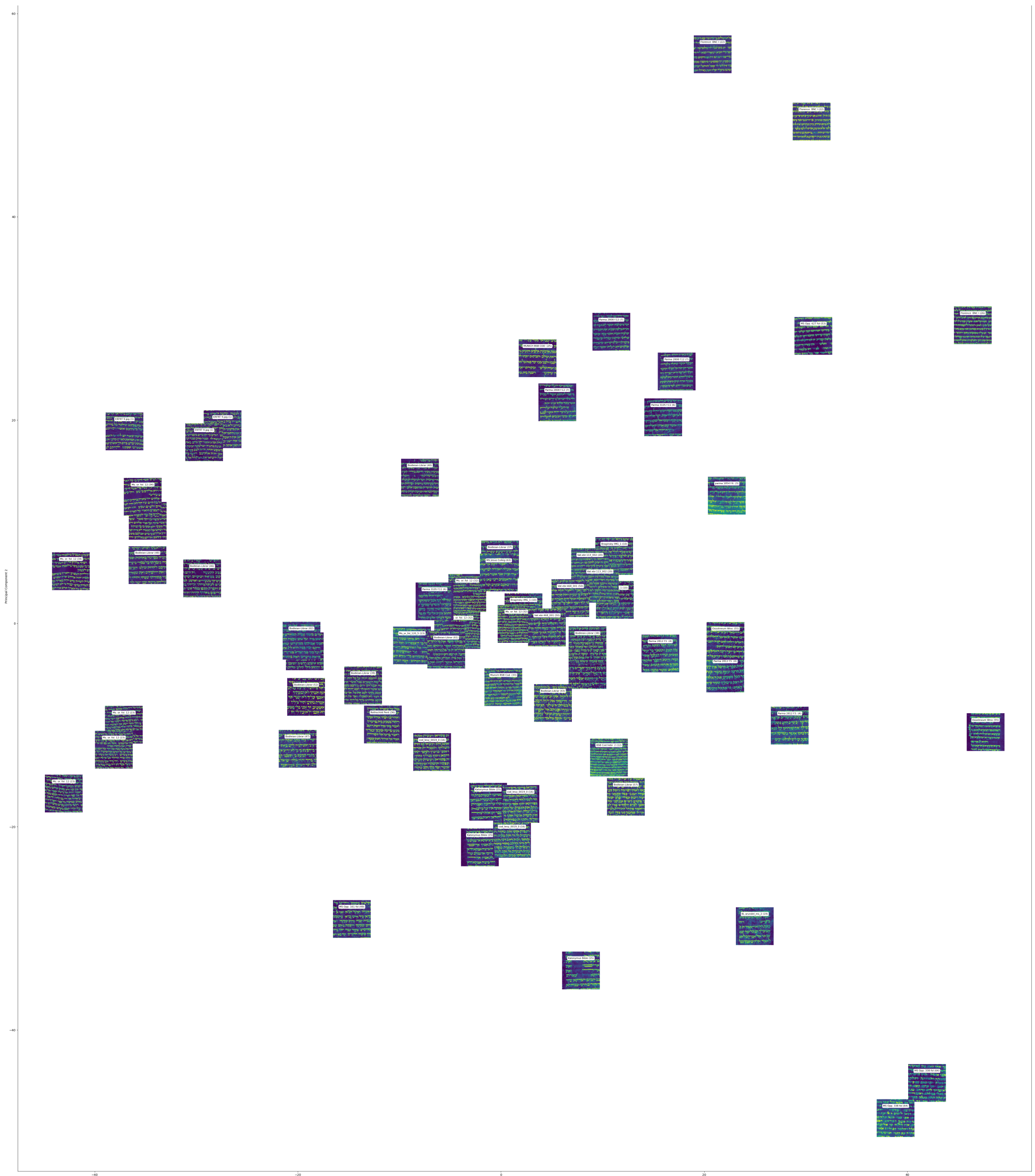

Fig. 4: Visualization of main text regions in 2D plot by projecting dimensionality-reduced SIFT+FVE features.

Fig. 5: Visualization of main text regions in 2D plot by projecting distances between the letters and the distances between the words.

Fig. 6: Visualization of main text regions in 2D plot by projecting dimensionality-reduced LBP features.

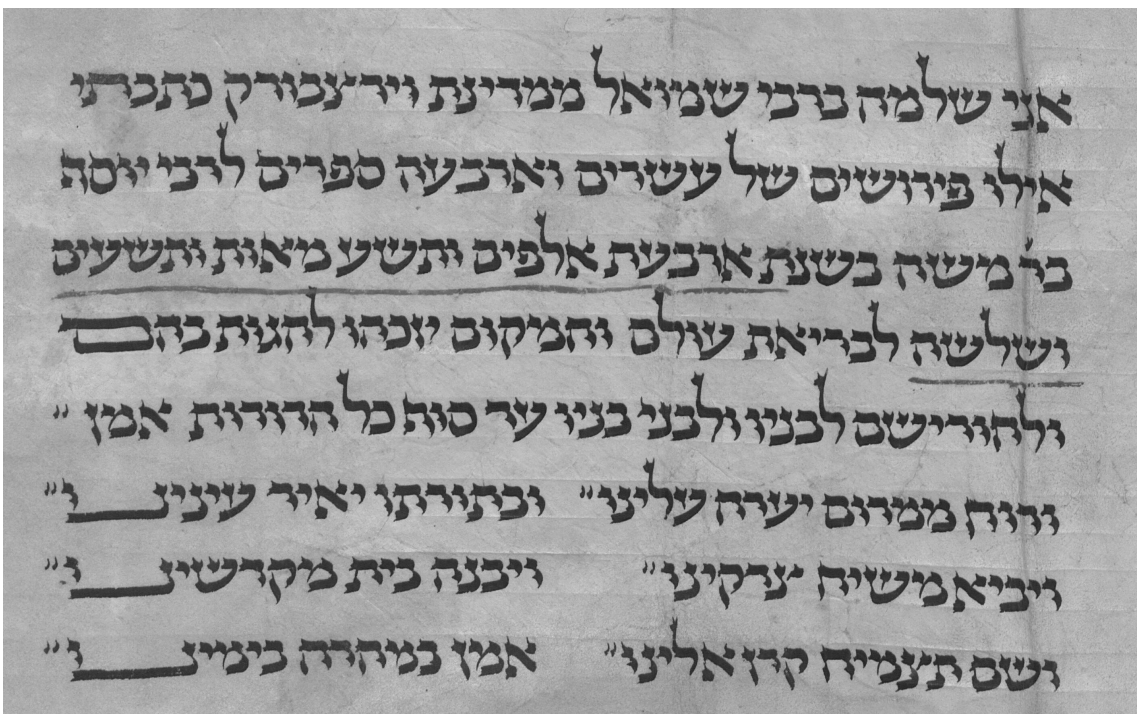

Fig. 7: An example text region labeled with the attributes: bited-aleph-0, fish-tail-1, left-justify-2, left-slanted-1, nesting-1, nikud-0, shading-0, short-descender-1, string-1, and vertical-stretch-0.

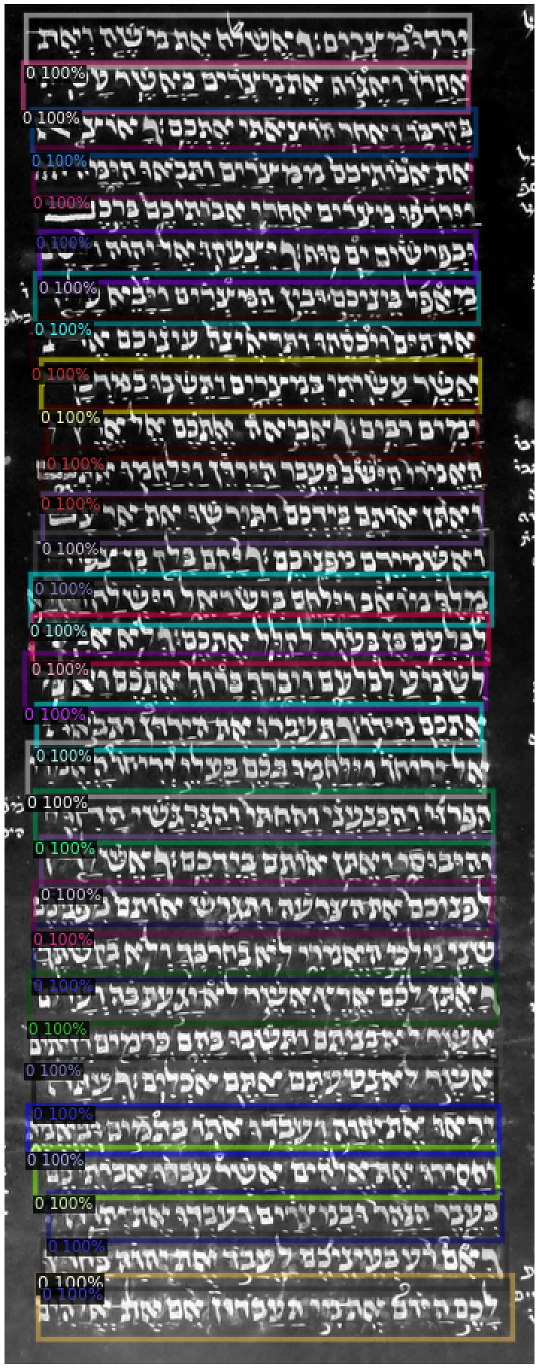

Fig. 8: Visualization detected bounding boxes for the text lines in a main text region.