# Computational Qumranic Paleography

A Research Proposal for Postdoctoral Fellowship

Berat Kurar Barakat

11.07.2023

## 1 Background and potential significance

The Dead Sea Scrolls, dated to the turn of the Common Era, are of enormous historical significance. They are the oldest witnesses of the biblical books and contain a treasure-trove of texts that continue to shed light on ancient Judaism and on the origins of Christianity. Unfortunately, the scrolls—or, rather, scroll fragments—were discovered in the 20th century in very poor condition, having deteriorated over the millennia. During the past decades, all the texts have been painstakingly transcribed by scholars. The extant fragments have recently been digitized by the Israel Antiquities Authority using high resolution state-of-the-art multispectral imaging.[1] Older infrared images, photographed under the auspices of the Palestine Archaeological Museum in the 1950s, are likewise available in retrodigitized form. The overwhelming majority show only a small number of words or even just a few letters, many of which are only partially visible. The median number of certain or probable letters is only 13 per fragment!

Three main computational challenges face us:

1. *Provenance:* Determine when each manuscript fragment was written. The Qumran writings in Hebrew and Aramaic are traditionally assigned to one of the Persian, Hellenistic, Hasmonean, Herodian, and Roman historical periods. As there are uncertainties and disputes regarding many fragments and occasional inconsistencies between manual paleographic dating and $^{14}$C dates, new computational evidence can be of significant value.
2. *Clustering:* Group fragments based on their degree of similarity. This can be performed at various levels of similitude, ranging from high-level categories, like same alphabet and same script, through level of formality and historical features like scribal culture, down to individual idiosyncrasies. The connection between the formality of the writing and the intended purpose of the manuscript has been the subject of recent investigations [13].
3. *Joins:* Pair fragments that may have been part of the same original manuscript, written by the same scribe, based on the handwriting and codicological features. Every single new join is of great significance to Qumran studies.

In previous work, fragments of the Cairo Genizah [4] were successfully clustered by computational means, [18] thousands of new joins were discovered among the fragments, [3] and a tool was created to help scholars find similar-looking fragments [19]. This work was widely reported in the press.[2] But all that was in the pre-deep–learning era, using features like SIFT. With the enormous recent strides in machine learning and computer vision, it is time to revisit the methodologies.

[1] See The Leon Levy Dead Sea Scrolls Digital Library and the experimental Scripta Qumranica Electronica in which we participated.

[2] E.g. *The New York Times*, "Computer network piecing together a jigsaw of Jewish lore", May 26, 2013.

## 2 Objectives

The goal of this research project is to apply modern computer-vision tools to analyze paleographic features of the handwriting of ancient *fragmentary* texts that are now becoming available in the increasingly many large corpora of digitized manuscripts. Specifically, we will concentrate on the Dead Sea Scroll fragments.

Though most of the modern methods require very substantial quantities of labeled training data, it should be possible to create sufficient synthetic data for our purposes. Naive application of state-of-the-art methods results in networks that concentrate more on the texture of the fragments than on the shape of the handwriting. This problem can be overcome with substantive use of augmentation [11]. Our recent work on Hebrew computational paleography has demonstrated the feasibility of this approach [9,10,8].

In the proposed work, the methods will need to be adapted before they can be fruitfully applied to the highly-degraded Dead Sea Scrolls.

## 3 Research methods and expected results

A convolutional neural network (CNN) reduces human intervention as it is a feature extractor plus a classifier that is end-to-end trainable. In contrast to models that use hand-crafted features, a CNN consist of filters that can extract features from input images. Their weights are learned through backpropogating the error between the prediction and the label. Hence the model figures out which features are useful according to the labels. This eliminates the effort of designing features that may be inappropriate due to limited human intuition or can be affected by scale and rotation [15].

We plan to use a CNN classifier for provenance prediction. Hand crafted features have an advantage [5,16] over deep-learning features due to the scarce amount of labeled data for the Dead Sea Scrolls. We intend to overcome this challenge by experimenting with several data regularization methods such as synthetic data generation [11], image transformations [9], and pretraining [17].

Pairing fragments is a classification problem with a large number of unknown classes, each class having only a few samples. Therefore we believe that distance metric learning models such as a siamese network with contrastive loss or a triplet network with triplet loss might provide an ideal solution for fragment pairing. These models not only aim to maximize the inter-class distances but also aim to minimize intra-class distances. Additionally, considering the raw input images includes information from both, the writing style and the background texture, which help for pairing fragments. We used a siamese network for retrieving historical handwritten Hebrew words in a past work [2]. In preliminary experiments for fragment pairing with a triplet network, we've worked with both binarized and full color images. For augmentation, we used letter-level elastic transformations and image-level color transformations (gradient and white gaussian noise).

One way to achieve deep-learning–based clustering of fragments is first to learn feature representations and then run a conventional clustering algorithm using the learned feature representations. The very first challenge is to direct the learning through features of the desired similitude level despite the absent labels. Previously we designed surrogate tasks that require the machine to learn to recognize handwritten text lines [12,1,7]. For this work, we hypothesize that spatial context prediction [6] or context encoding [14] of binarized images—where only the handwriting strokes are visible—will learn the features of handwriting patterns. These feature vectors can be used for further clustering of alphabet types or writing styles. Preliminarily, we have experimented with a CNN that achieves over 90% accuracy in spatial context prediction of historical handwritten word images, which is considerably above human performance.[3]

[3] Berat Kurar Barakat, "Human performance evaluation in spatial context prediction of handwritten words".

## 4 Research plan and resources required

We plan this research to be performed over the course of one year (including writing up results), and using available servers on the school's network.

## References


1. Barakat, B.K., Droby, A., Alaasam, R., Madi, B., Rabaev, I., Shammes, R., El-Sana, J.: Unsupervised deep learning for text line segmentation. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 2304–2311. IEEE (2021)
2. Barakat, B.K., El-Sana, J., Rabaev, I.: The pinkas dataset. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 732–737. IEEE (2019)
3. Ben-Shalom, A., Choueka, Y., Dershowitz, N., Shweka, R., Wolf, L.: Where is my other half? In: DH (2014)
4. Choueka, Y.: Computerizing the cairo genizah: Aims, methodologies and achievements. Ginzei Qedem **8**, 9–30 (2012)
5. Dhali, M.A., Jansen, C.N., De Wit, J.W., Schomaker, L.: Feature-extraction methods for historical manuscript dating based on writing style development. Pattern Recognition Letters **131**, 413–420 (2020)
6. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE international conference on computer vision. pp. 1422–1430 (2015)
7. Droby, A., Barakat, B.K., Madi, B., Alaasam, R., El-Sana, J.: Unsupervised deep learning for handwritten page segmentation. In: 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 240–245. IEEE (2020)
8. Droby, A., Kurar Barakat, B., Vasyutinsky Shapira, D., Rabaev, I., El-Sana, J.: Vml-hp: Hebrew paleography dataset. In: International Conference on Document Analysis and Recognition. pp. 205–220. Springer (2021)
9. Droby, A., Rabaev, I., Shapira, D.V., Kurar Barakat, B., El-Sana, J.: Digital hebrew paleography: Script types and modes. Journal of Imaging **8**(5), 143 (2022)
10. Droby, A., Shapira, D.V., Rabaev, I., Barakat, B.K., El-Sana, J.: Hard and soft labeling for hebrew paleography: A case study. In: International Workshop on Document Analysis Systems. pp. 492–506. Springer (2022)
11. Keret, S., Wolf, L., Dershowitz, N., Werner, E., Almogi, O., Wangchuk, D.: Transductive learning for reading handwritten tibetan manuscripts. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 214–221. IEEE (2019)
12. Kurar Barakat, B., Droby, A., Saabni, R., El-Sana, J.: Unsupervised learning of text line segmentation by differentiating coarse patterns. In: International Conference on Document Analysis and Recognition. pp. 523–537. Springer (2021)
13. Longacre, D.: Paleographic style and the forms and functions of the dead sea psalm scrolls: A hand fitting for the occasion? Vetus Testamentum **72**(1), 67–92 (2021)
14. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2536–2544 (2016)
15. Popović, M., Dhali, M.A., Schomaker, L.: Artificial intelligence based writer identification generates new evidence for the unknown scribes of the dead sea scrolls exemplified by the great isaiah scroll (1qisaa). PloS one **16**(4), e0249769 (2021)
16. Wahlberg, F., Mårtensson, L., Brun, A.: Large scale continuous dating of medieval scribes using a combined image and language model. In: 2016 12th IAPR Workshop on Document Analysis Systems (DAS). pp. 48–53. IEEE (2016)
17. Wahlberg, F., Wilkinson, T., Brun, A.: Historical manuscript production date estimation using deep convolutional neural networks. In: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 205–210. IEEE (2016)
18. Wolf, L., Dershowitz, N., Potikha, L., German, T., Shweka, R., Choueka, Y.: Automatic palaeographic exploration of Genizah manuscripts. Books on Demand (BoD) (2011)
19. 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**(1), 118–135 (2011)