

Determining the Influence of Papyrus Characteristics and Data Argumentation on Fragments Retrieval with Deep Metric Learning

In practice, ancient papyri are torn into several fragments. The task of papyrologists is to assemble and decipher these fragments. Together with historians, this can yield important results about antiquity. However, the process of manual piecing together is very time-consuming and difficult.

In the following work, the aim is to facilitate the work of papyrologists by partially automating the process of assembling. For this purpose, an algorithm is designed which suggests a smaller selection of possibly matching fragments for a papyrus fragment. Summarized in the following this algorithm is called a puzzle help function. With such a puzzle help function the effectiveness and efficiency can be increased when assembling fragments by papyrologists. However, a puzzle help function must fulfill several conditions.

First, the creation of such a function may not require a large dataset of assembled Papyri fragments, as few such data exist in the world. Deep Metric Learning (DML) has been used so far to create a puzzle help function and yields good accuracy. If one wants to further refine DML to achieve very good accuracy, an immense amount of data will be needed. Otherwise, the model would very likely not converge. If one were to manually assemble the fragments to thus obtain a large data set, the purpose of the puzzle help function would be obsolete. The time required of the papyrologist would remain the same at best or be increased at worst. Generative-Adversarial-Networks (GANs) are a well researched way to generate more artificial data across domains. It has been shown that it can significantly increase accuracy across different application domains. Our goal is to use this method to overcome the hurdle of insufficient data.

Another condition is that the false-negative hits are as small as possible. Thus, the pre-selection of the puzzle help function should contain as many potentially matching fragments as possible. If the matching fragment is not contained in the subset, the papyrologist has no chance to identify the correct fragment. Unless he extends his search again to the entire data set, which makes the puzzle help function obsolete again. A further drawback is that he may spend an extremely long time searching for the matching fragment, which cannot be found. Target metrics to measure this property and to get an overall picture of the model are the mAP, top-1, pr@10, pr@100. In the following we talk simplified only about accuracy. This accuracy depends strongly on which image section is used as input. More precisely, what is suggested to the model as equal and unequal during training determines the accuracy of the model. Therefore, our approach is to statistically investigate how certain papyrus characteristics affect the accuracy of such a model. For this purpose, we compare the accuracy of different models by using only the foreground of a papyrus (text) or only the background (fibers) as input for training. Separating text and phasors is a challenge of its own with many obstacles. We use different methods and evaluate them on overall accuracy. Separated evaluation would not be possible without pixel accurate ground truth. Also, we are interested in the influence on the accuracy and not the separation itself. For text recognition, we first use a threshold, which separates out the text and returns a mask. The extracted areas can now be filled using inpainting and the mask. The inpainting procedure is also evaluated only on the basis of the overall accuracy.

Finally, we investigate whether certain characteristics (e.g. fibers) can be used in other ways. For

example, to help papyrologists not only identify potential candidates, but also determine their exact location. This would be the logical step towards a fully automated function. So in our terminology, a transformation from a puzzle helper to a puzzle solver.

In summary, the work is divided into the following milestones:

- Incorporating thresholding [3, 4] and inpainting [5] into the training pipeline of a DML model [6].
- Use the preprocessed data to create a large dataset suitable for DML network throughout a GAN-based method [2].
- Evaluate the performance of the thresholding and inpainting techniques on the Michigan dataset [1] and the large dataset created by the GAN.
- Further experiments regarding fragment matching via papyri-fibers.

The implementation will be done in Python.

Supervisors: Dr.-Ing. V. Christlein, Prof. Dr.-Ing. habil. A. Maier, Mathias Seuret M. Sc.

Student: Timo Bohnstedt

Start: November 8th, 2021

End: April 8th, 2022

References

- [1] A. Pirrone, M. Beurton-Aimar, and N. Journet, “Self-supervised deep metric learning for ancient papyrus fragments retrieval,” vol. 24, no. 3, pp. 219–234. [Online]. Available: <https://doi.org/10.1007/s10032-021-00369-1>
- [2] A. Antoniou, A. Storkey, and H. Edwards, “Data augmentation generative adversarial networks,” eprint: 1711.04340.
- [3] C. Tensmeyer and T. Martinez, “Historical document image binarization: A review,” vol. 1, no. 3, p. 173. [Online]. Available: <https://doi.org/10.1007/s42979-020-00176-1>
- [4] I. Pratikakis, K. Zagoris, X. Karagiannis, L. Tsochatzidis, T. Mondal, and I. Marthot-Santaniello, “ICDAR 2019 competition on document image binarization (DIBCO 2019),” in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1547–1556.
- [5] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, “Image inpainting for irregular holes using partial convolutions,” vol. abs/1804.07723. [Online]. Available: <http://arxiv.org/abs/1804.07723>
- [6] K. Musgrave, S. Belongie, and S.-N. Lim, “PyTorch metric learning,” eprint: 2008.09164.