Master's Thesis at the Pattern Recognition Lab, FAU Erlangen-Nuremberg

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 and the task of papyrologists is to assemble and decipher these fragments. Once successfully reconstructed, ancient papyrus is an important historical document. It offers the opportunity to gather important information about ancient cultures. However, reassembling by hand is very time-consuming and challenging because fragments differ in color, structure, and shape. In other words, they do not fit together perfectly - like an artificially designed toy puzzle.

The objective of the following thesis is to make the work of papyrologists easier by partially automating the process of reassembling. To this end, an algorithm is designed that infers a smaller sub-selection of fragments with a high likelihood of being a good fit given one particular fragment as input. In the following, this algorithm is called a puzzle-helper. The purpose of the proposed puzzle-helper is to increase the effectiveness and efficiency of the reassembling task. Nevertheless, in a real-world scenario, a puzzle-helper must fulfill several constraints.

First, training such a helper shall require just a tiny amount of data of assembled Payri fragments because only a tiny bit of such data exists. So far, Deep Metric Learning (DML) has been used to create puzzle helpers, achieving a top-1 accuracy of 0.73. Top-1 accuracy means the accuracy of the reassembling task, where only the candidate with the highest probability is considered at each fragment, but further improvement of DML models requires an immense amount of data. Otherwise, the model will probably not converge. If fragments are reassembled manually to get an extensive data set, the purpose of the puzzle-helper becomes obsolete because the time required by the papyrologist remains the same, so it is not helpful. Generative-Adversarial-Networks (GANs) are a well-known possibility to generate more artificial data. They could improve the overall performance in different domains. In this thesis, a GAN will also be used to overcome the problem of too little data.

Another constraint is to keep the number of false-negative candidates as low as possible. That means in the preselection of the puzzle-helper, all potentially matching fragments shall be included. If the perfect match is not included in the subset, the papyrologist has no chance to identify the correct fragment unless he extends his search again to the entire data set. Thus the puzzle-helper is obsolete again. Another disadvantage is that the approach could fail silently. The papyrologist might search long for a match because the expert is convinced that the algorithm works fine. However, if the perfect match is not among the candidates in the subset, the search would be useless. Metrics that measure the accuracy in a helpful manner are the mAP, top-1, pr@10, pr@100. On the one hand, these metrics show how accurate the model is in general and how precise the model is within a particular range of potential candidates. In the following, we will only talk about accuracy in a simplified way. This accuracy depends strongly on which part of the image is used for training such a puzzle helper. More precisely, what is suggested to the model during training as equal and unequal determines the model's accuracy. In this work, we will statistically measure how specific papyrus characteristics determine the accuracy of a DML model. For this purpose, the accuracy of different models is compared by using only the foreground of a papyrus (text) or the background (fibers) as input for training. Separating text and fibers is challenging because the background often has similar colors as the foreground, making thresholding way more complicated. Several methods will be used and get evaluated on the overall accuracy. A separate evaluation is not straightforward and is misleading because we do not have pixel-wise ground truth. Also, the focus of the thesis lies on the influence of specific characteristics and the influence of more data available instead of the separating process itself. For separating the text, a method based on a color threshold will be used(binarization), which separates the text and generates a text mask. The extracted areas can now be filled using an inpainting approach. The inpainting procedure is also just evaluated based on overall accuracy.

Finally, further experiments will examine whether specific characteristics (fibers) can be exploited in some way. Specifically, it will be tested if the fibers can be used to determine the exact position of the fragments to each other. Determining the position is the next logical step towards a fully automated helper. So, the final goal is to design a puzzle solver function by combining the DML model and a suitable method for position determination.

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

- Separating text and papyrus fibers by binarization and inpainting.
- Generating larger amounts of data through a GAN.
- Evaluation by means of a DML model using the original data in the original state (text and fibers), in the processed state (text only or fibers only). Also, an evaluation of the artificial data together with the original data in the original state and processed state.
- Review of the state of the art on the possibilities of using certain characteristics for position determination.

The following research questions emerge from the milestones:

RQI: Do the chosen metrics (mAP, top-1, pr@10, pr@100) differ significantly based on the binarization and impainting methods used?

RQII: Do the selected metrics (mAP, top-1, pr@10, pr@100) differ significantly when only the text or only the fibers are used as input as opposed to the unprocessed data?

RQIII: Do the selected metrics (mAP, top-1, pr@10, pr@100) differ significantly when additional artificial data generated by a GAN are used for training?

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 oth, 2022

References