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

Master's Thesis in Computer Science

submitted
by

Timo Bohnstedt

born 17.02.1992 in Gunzenhausen

Written at

Lehrstuhl für Mustererkennung (Informatik 5)
Department Informatik
Friedrich-Alexander-Universität Erlangen-Nürnberg.

Advisor: Mathias Seuret M. Sc.,Dr.-Ing. Vincent Christlein

Started: 15.12.2021

Finished: 15.06.2022

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Die Richtlinien des Lehrstuhls für Studien- und Diplomarbeiten habe ich gelesen und anerkannt, insbesondere die Regelung des Nutzungsrechts.

Erlangen, den 12. Januar 2022

Übersicht

Antike Papyri sind häufig in mehrere Fragmente zerrissen, und die Aufgabe der Papyrologen besteht darin, diese Fragmente zusammenzusetzen und zu entziffern. Einmal erfolgreich rekonstruiert, bietet antikes Papyrus die Möglichkeit, wichtige Informationen über vergangene Zeiten zu sammeln. Das Zusammensetzen von Hand ist jedoch zeitaufwändig, da sich die Fragmente in Farbe, Struktur und Form unterscheiden. Mit anderen Worten: Sie passen nicht perfekt zusammen - wie ein künstlich hergestelltes Spielzeugpuzzle. Ein Algorithmus, der eine Auswahl von passenden Fragmenten zu weiteren Fragmenten vorschlägt, spart Papyrologen viel Zeit. Hierfür wurde bereits in der Vergangenheit gezeigt, dass tiefes metrisches Lernen ein vielversprechender Ansatz ist [Pir] In der nachfolgenden Arbeit wird gezeigt, wie ein Algorithmus mit einer mean average percesion (mAP) von 70% geeignete Kandidaten findet. Es konnte außerdem gezeigt werden, dass der Algorithmus sich hauptsächlich an der Struktur der Papyrus Fasern orientiert und der Text vorerst irrelevant ist. Hierfür wurden Text und Fasern mittels Binarisierung separiert. Der mAP ändert sich nicht signifikant, wenn man die Textinformation herausfiltert und nur mit den Fasern arbeiten. Des Weiteren konnte gezeigt werden, wie Papyrus Fasern erweitert werden können, um eine geometrische Zuordnung möglicherweise passender Fragmente zu erreichen. Die Genauigkeit ist dabei mit der von Graph Algorithmen vergleichbar benötigt allerdings weniger Trainingsdaten.

Abstract

Ancient papyri are frequently torn into several fragments, and the task of papyrologists is to assemble and decipher these fragments. Once successfully reconstructed, ancient papyrus offers the opportunity to gather crucial information about past times. However, reassembling by hand is time-consuming because fragments differ in color, structure, and shape. In other words, they do not fit together perfectly - like an artificially designed toy puzzle. An algorithm that suggests a selection of matching fragments to on specific fragment saves papyrologists a time. For this, it has been shown in the past that deep metric learning is a promising approach [Pir]. In the following thesis it is shown how an algorithm with a mean average precision (mAP) of 70% finds suitable matching candidates. It was also shown that the algorithm is mainly guided by the structure of the papyrus fibers and the text is irrelevant for that approach. For this purpose, text and fibers were separated using binarization. The mAP does not change significantly if the textual information was filtered out. Furthermore, it is shown how Papyrus fibers can be extended to achieve geometric matching of possibly matching fragments. The accuracy is comparable to that of graph algorithms but requires less training data.

Contents

Chapter 1

Introduction

Ancient papyri are [Pir] frequently torn into several fragments, and the task of papyrologists is to assemble and decipher these fragments. Once successfully reconstructed, ancient papyrus offers the opportunity to gather crucial information about past times. However, reassembling by hand is time-consuming because fragments differ in color, structure, and shape. Since the phrase puzzling is used, it implies that those fragments are perfectly designed puzzle pieces. Usually, it is the opposite. That means that non-professionals can not tell if two images belong together at all. An example is shown in Figure 1. It can be observed from the Figure that the fibers, the color, and the structure of the two fragments do not fit perfectly into each other. Nevertheless, they belong to the same papyrus. It is hard to tell whether fragments belong together or not because they age differently. Environmental local factors such as exposure to sunlight determine the altering process differently. For example, the medium color of two fragments is inconsistent if one fragment was buried and the second fragment was not. That implies that color is not a good feature for matching fragments.

Finding meaningful features (semi) automatically on historical documents and reassembling them has become a popular challenge in the computer vision community. The researchers apply machine learning algorithms to the data and train a model. Those models can then find potential matching candidates for a specific fragment. Deep Learning algorithms are among the commonly discussed types of algorithms when it comes to supporting papyrologists. Research has shown that the use of Deep Learning can increase the efficiency of papyrologists. However, even though the results are promising, there are still many unanswered questions that we do not understand. Once a better understanding of the features is obtained, the algorithms can increase the papyrologist's efficiency by a greater chance.

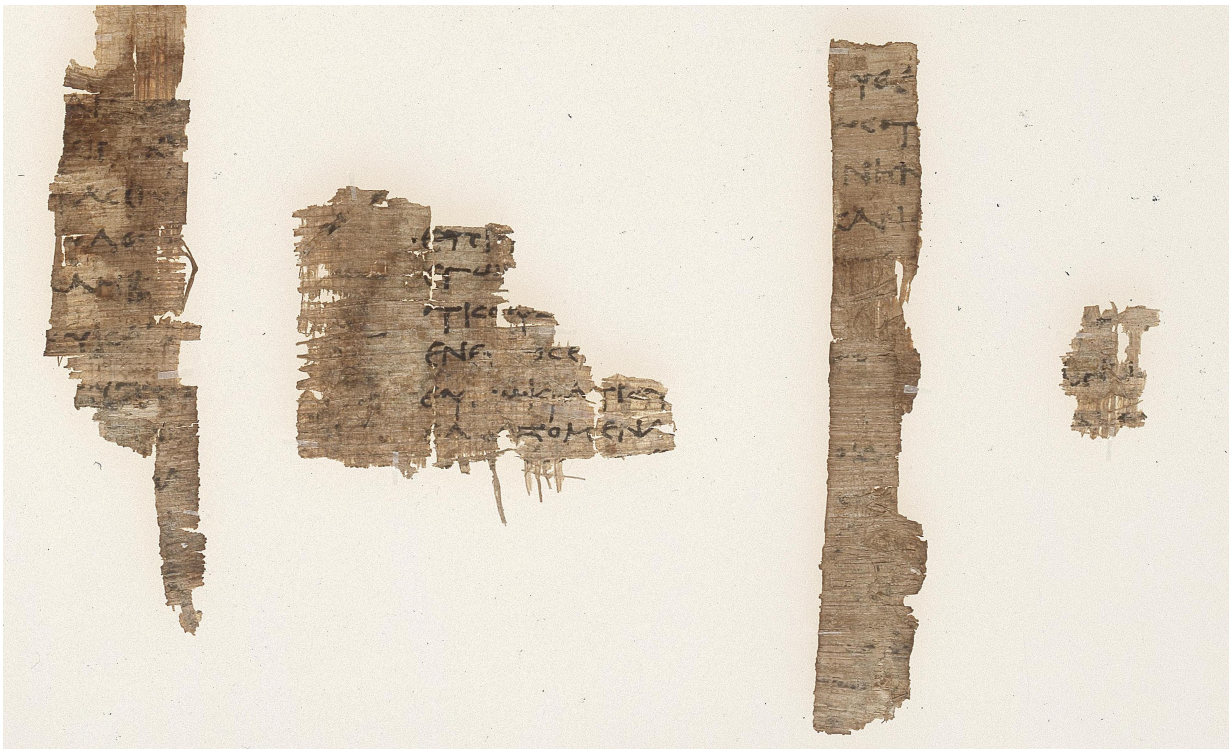

Figure 1.1: A papyri torn into several fragments

1.1 Contribution

The general objective of this thesis is to make the work of papyrologists easier and increase their efficiency by partially automating the reassembling process. To this end, an algorithm is designed to infer a smaller sub-selection of fragments with a high likelihood of being a potential fit. In the following, this algorithm is called puzzle-helper. Additionally, the thesis explores the use of papyrus fibers to determine an accurate spatial position of two potential matching fragments. Also, using deep learning implies that a vast amount of (labeled) data is required. The database from the University of Michigan offers plenty of it. Once the data is downloaded, it must be correctly preprocessed, like removing low contrast images or labeling the data. In particular, this thesis is centered around the following research questions:

RQI: Does the puzzle-helpers-accuracy differ significantly when only the text or only the fibers are used as input as opposed to the unprocessed data?

RQII: With the help of papyrus fibers, is it possible to determine the position of a fragment out of several matching candidates?

1.2 Outline

In the Chapter 1, the context of this master thesis was explained, and the research questions got defined. In the following, it is explained how the thesis is structured. Chapter 2 presents groundbreaking work in all areas that are relevant for this thesis. That includes binarization, inpainting, and deep metric learning. The chapter's goal is to show the reader a quick overview of actual results in the field of historical fragment retrieval. Chapter 3 aims to explain how a ground truth is computed with the help of Deep Metric Learning (DML) for comparison later results. In Chapter **??** the reader will learn how different datasets are obtained with the help of binarization and inpainting techniques. Furthermore, it is stated how results differ if the DML algorithm of the previous chapter is evaluated onto the different datasets. The results of the best-performing dataset from the previous chapter are used to determine spatial positions of potential matching candidates. That approach is explained in the Chapter **??**. A discussion about how the results of the different chapters determine the other results is stated in Chapter **??**. Finally, a conclusion is presented in Chapter **??**, where the reader will be informed about lessons learned and potential future work.

Chapter 2

State of the Art

This chapter summarizes state-of-the-art publications in the fields of fragment retrieval, historical document image binarization, and impainting. Fragment retrieval is the main objective of this thesis, whereas binarization and impainting are used to determine the influence of specific document characteristics on the retrieval approach.

Antoine Pirrone, Marie Beurton Aimar, Nicholas Journet are convinced that semi-automatic fragment retrieval is necessary to help papyrologists. Otherwise, they must review many fragments manually to find those that go together and then assemble them to analyze the text finally, as well. That is why they provide a solution where an expert uses a fragment as a request element and get fragments that belong to the same papyrus (puzzle helper). Their main contribution is the proposal of deep siamese network architecture, called Papy-S-Net for Papyrus-Siamese-Network, designed for papyri fragment matching. Their network was trained and validated on around 500 papyrus fragments. Since no one had approached historical fragment retrieval with a siamese network before they compared their results with the paper of Koch et al., He uses the approach within another domain. To train and validate the network, they created fragments semi-artificially. Precisely, they divided the papyri images where why could find natural fragments. Papy-S-Net outperforms Koch et al.s network. On their assembled ground truth, why could re-able 79% of the fragments correct.

In their follow-up work, the authors explored more ways such that papyrologists can obtain valuable matching suggestions on new data using Deep Convolutional Siamese-Networks. This time they focused on the low data regime, and they claimed that less labeled data is available to train sophisticated deep learning models. However, they proved that the from-scratch self-supervised approach is more effective than knowledge transfer from a large dataset. Furthermore, the paper is more precise in the evaluation section, and they planned to offer a publicly available

Task	Publication	Year	Method	Dataset	Public	Train Data	Val Data	Test Data	Metric	Result
Fragment Retrieval	Self-supervised deep metric learning for ancient papyrus fragments retrieval	2021	Self-supervised deep metric learning	Michigan	yes	800 papyri	100 papyri	100 papyri	Top-1 Accuracy	0.73
Fragment Retrieval	Self-supervised deep metric learning for ancient papyrus fragments retrieval	2021	Self-supervised deep metric learning	Hisfrag	yes	9000 papyri	1000 papyri	100 papyri	Top-1 Accuracy	0.87
Fragment Retrieval	Papy-S-Net: A Siamese Network to match papyrus fragments	2019	Siamese Network	B500	no	8500 patches	2000 patches	1000 patches (50 fragments)	True Pos. / Accuracy	0,79
Position Estimation	Using Graph Neural Networks to Reconstruct Ancient Documents	2021	Graph Neural Networks	B500 (subset)	no	3394 imgs	500 images	200 images	Accuracy	0.85
Data Argumentation	Data Augmentation Generative Adversarial Networks	2018	Generative Advesarial Networks	EMNIST	yes	-	-	-	Accuracy	+0.13

Table 2.1: The table summarizes the state-of-the-art publications for this thesis. In addition to the results and metrics used, other features are also presented.

dataset. Unfortunately, that never happened until now.

The work of Pirrone and his colleagues is used for comparison within this thesis. Building on top of their work, it determines how specific image characteristics determine the success of deep metric learning on historical fragment retrieval.

The review paper of Tensmeyer and Martinez provides a detailed view of the field of historical document image binarization. Therefore the authors focus on the contributions made in the last decade. They explain how the Document Image Binarization Contest and the corresponding standard benchmark dataset raised interest in that particular research field. The paper provides an overview of the standard image thresholding, preprocessing, and post-processing methods. Furthermore, the writers review the literature on statistical models, pixel classification with learning algorithms, and parameter tuning methods. In addition to reviewing binarization algorithms, they debate available public datasets and evaluation metrics. They suggest separating metrics onto whether they require pixel-level ground truth or not. Finally, they offer guidance for future work.

Chapter 3

Algorithms as Puzzle Helper

The following chapter explains the first steps towards a puzzle solver algorithm. In the beginning, the characteristics of the dataset and the performed preprocessing steps are explained. Afterward, the necessary theory behind a puzzle helper is explained. Finally, the results of various experiments and their evaluation are presented. The main objective is to show that the algorithm can provide a subset of potential matches for a given papyrus fragment.

3.1 Methods

3.1.1 Data

Beispielbilder - Wie sehen die Daten aus?

Histogram der gedownloadeten Bilder. 3.1.1 3.1.1 3.1.1 3.1.1

Boxplots fÃ¼r die Grosse.

Pipelinediagramm 3.1.1. Was haben wir mit den Daten gemacht?

Bild metrisches lernen erklÃ¤ren - wie funktioniert unser Algorithmus?

Schaubild Metric Welche Metriken werden verwendet?

Was bedeutet die Zufallsmetrik.

3.2 Results

Tabelle mit Hyperparametern und Ergebnissen - Welche Experimente wurden durchgefÃ¼hrt?
Graph mit mAP

U-MAP Accuracy Plot?

3.3 Evaluation

Was lief gut? Was lief schlecht?

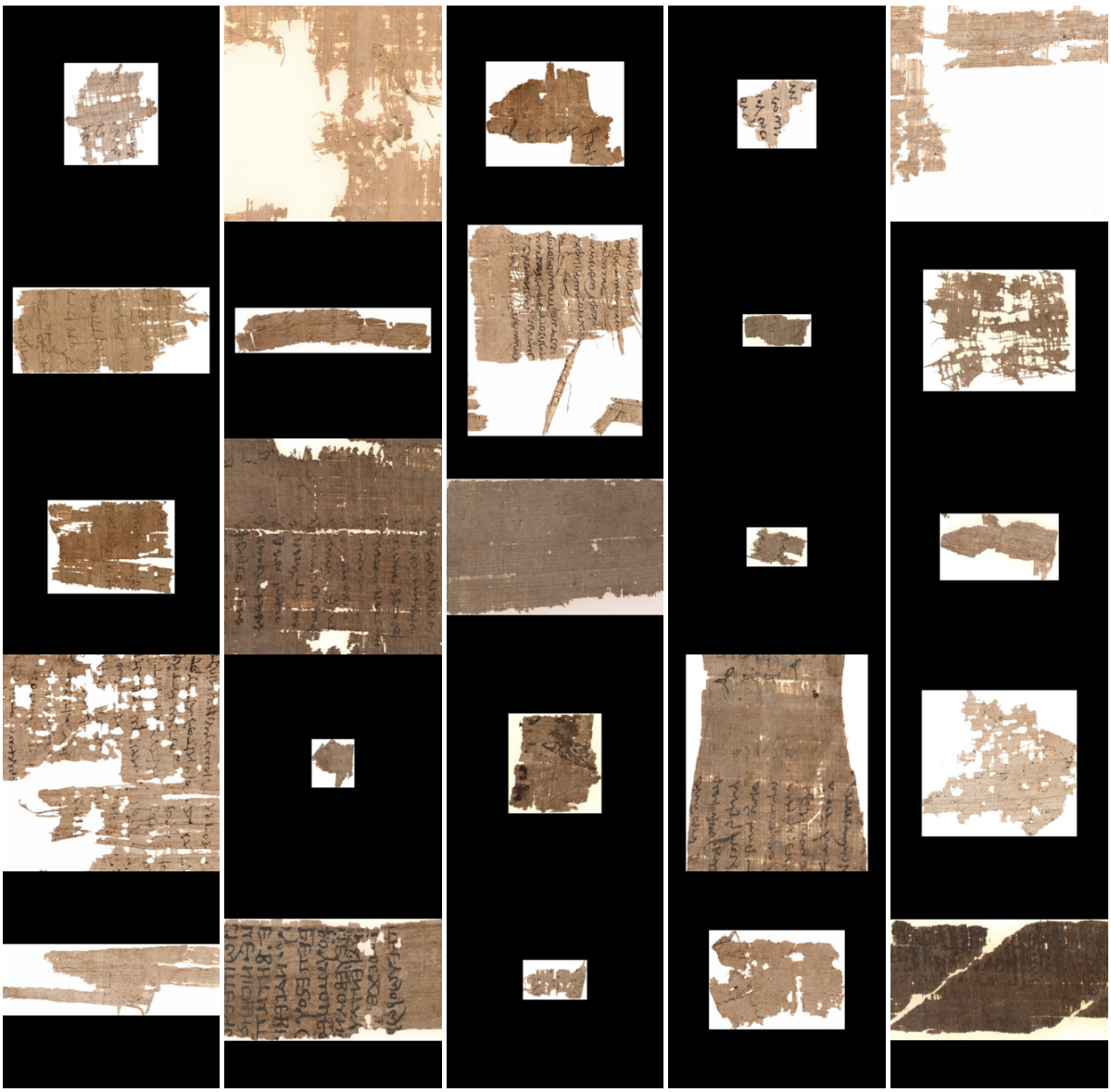

Figure 3.1: Each papyri was turned into grayscale, blured ...

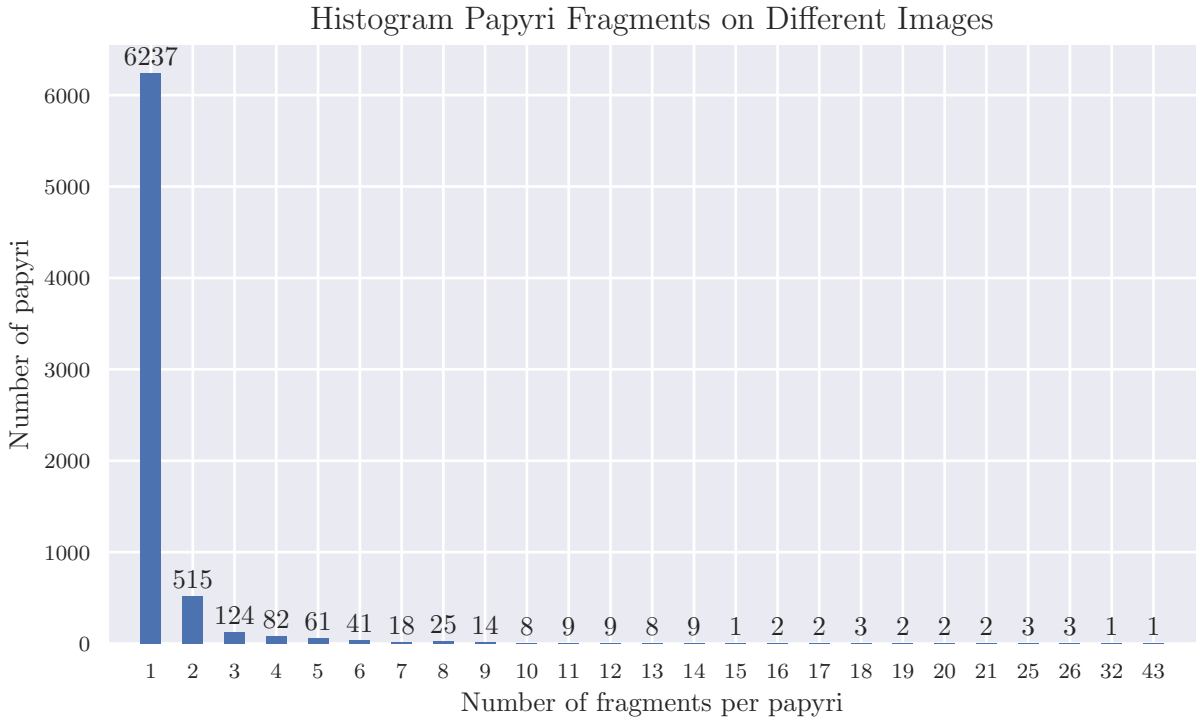

Figure 3.2: Each papyri was turned into grayscale, blured ...

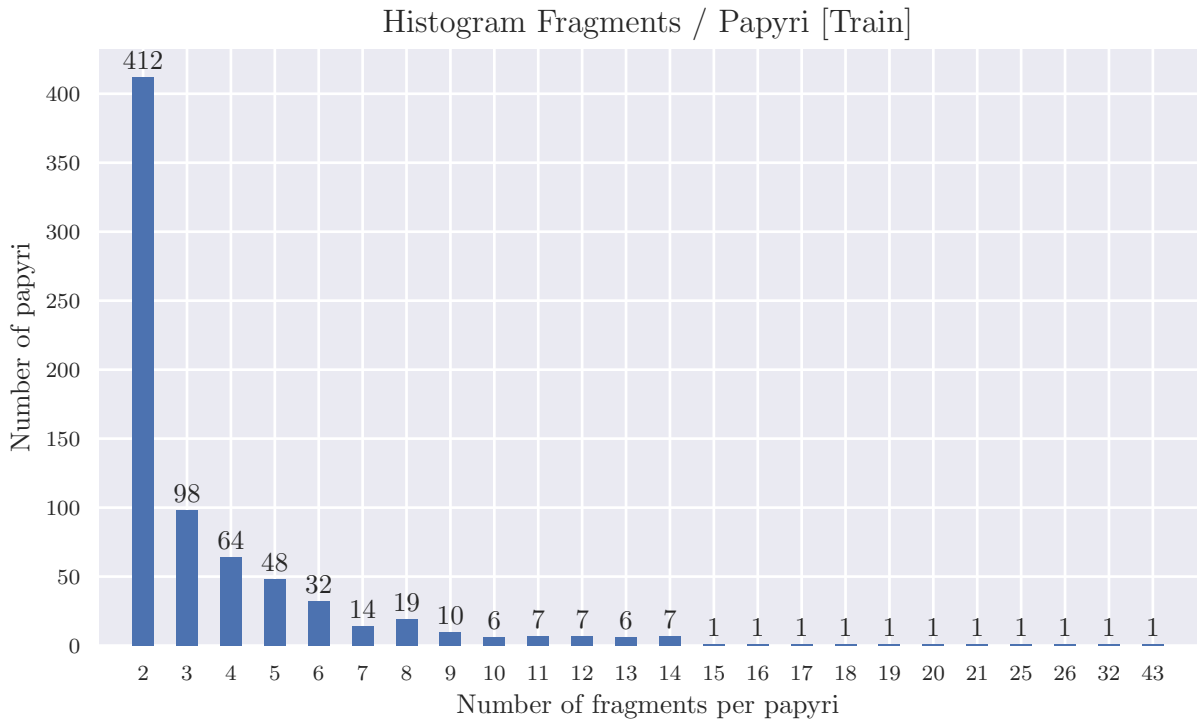

Figure 3.3: Train Histrogram

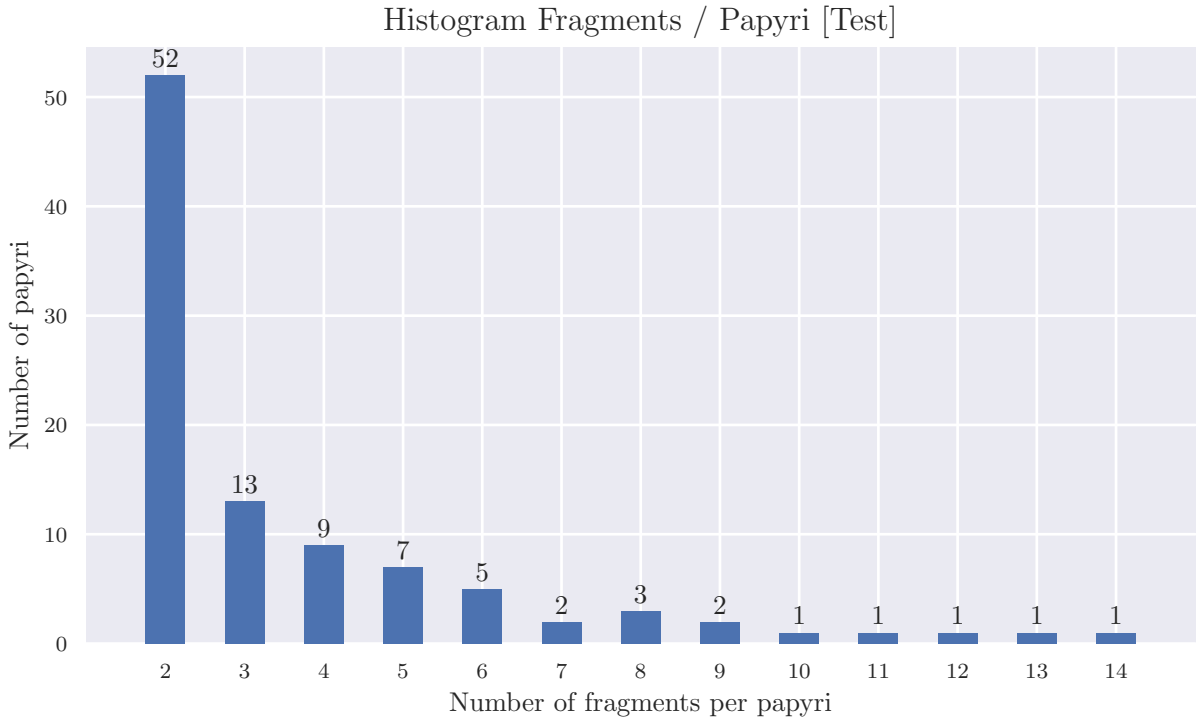

Figure 3.4: Test Histogram

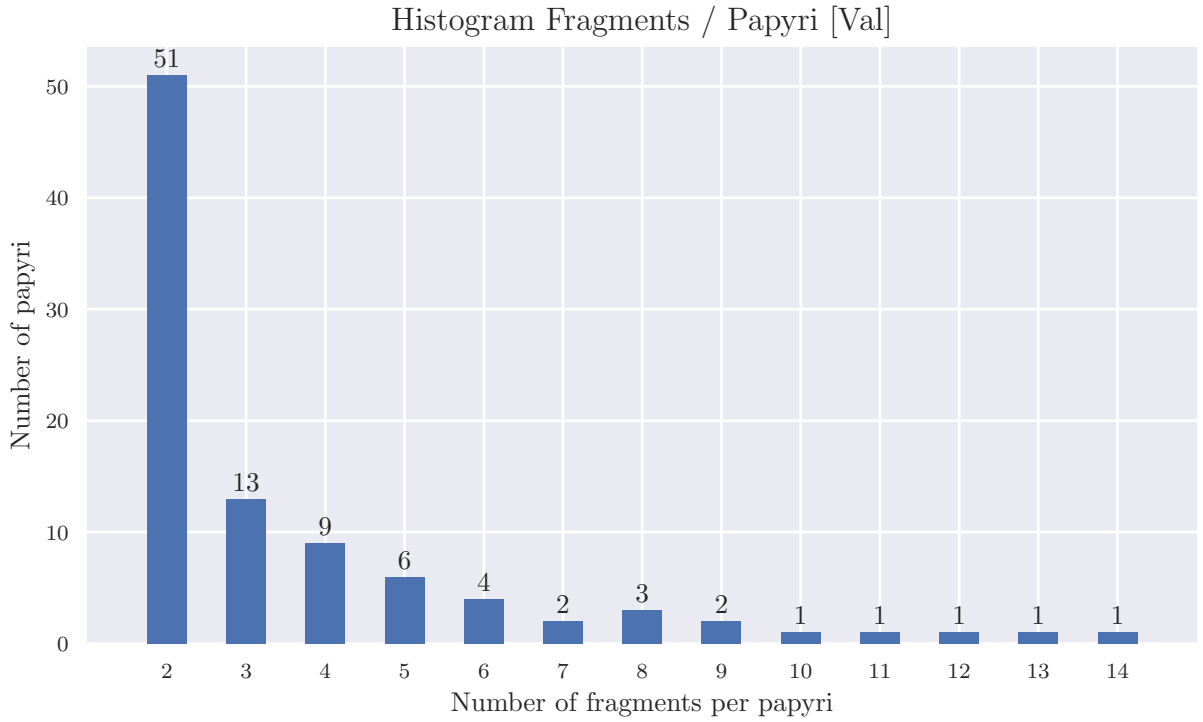

Figure 3.5: Valdidation Histrogram

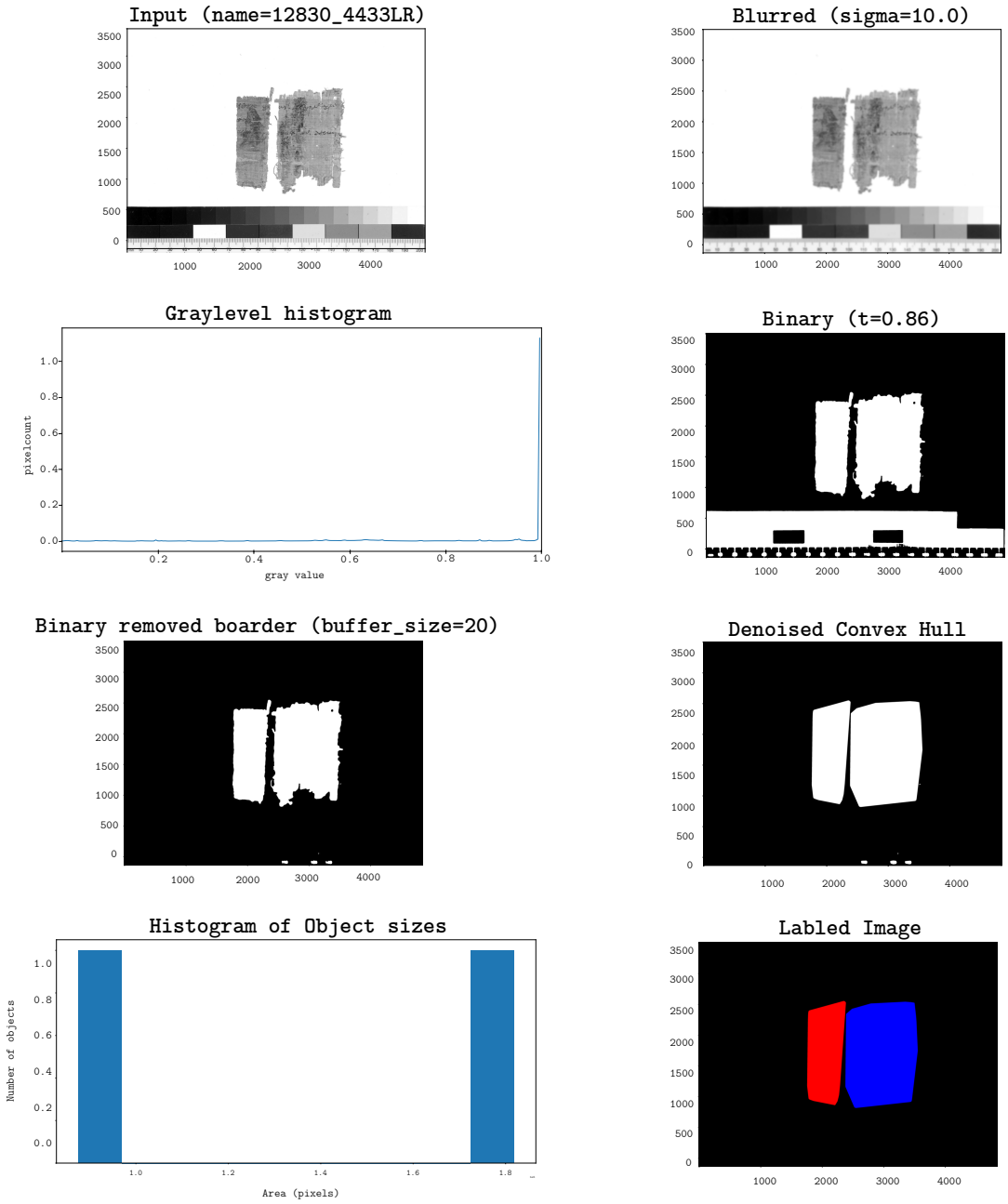

Figure 3.6: Each papyri was turned into grayscale, blured ...

List of Figures

List of Tables

Bibliography

[Pir] Antoine Pirrone, Marie Beurton-Aimar, and Nicholas Journet. Self-supervised deep metric learning for ancient papyrus fragments retrieval.