

# GRK-Papyri: A Dataset of Greek Handwriting on Papyri for the Task of Writer Identification

Hussein Mohammed\*, Isabelle Marthot-Santaniello<sup>‡</sup>, Volker Märgner\*<sup>†</sup>

\* Cluster of Excellence: Understanding Written Artefacts, Universität Hamburg, Hamburg, Germany

Email: hussein.adnan.mohammed@uni-hamburg.de

<sup>‡</sup> Universität Basel, Basel, Switzerland

Email: i.marthot-santaniello@unibas.ch

<sup>†</sup> Technische Universität Braunschweig, Braunschweig, Germany

Email: maergner@ifn.ing.tu-bs.de

**Abstract**—Presenting actual research questions from academia through publishing datasets is a practice of great importance in order to generate relevant solutions. Therefore, we propose a dataset of handwriting on papyri for the task of writer identification. This dataset is derived directly from research questions in the field of Papyrology, and the samples are selected by experts from the respective field of research. This dataset consists of 50 handwriting samples in Greek on papyri approximately from the 6th century A.D., which belong to 10 different scribes. It is prepared and made freely available for non-commercial research along with their confirmed ground-truth information related to the task of writer identification. This paper presents not only the details of the dataset but also its relation to research questions and how the results of computational analysis can support scholars from manuscript research. Some preprocessing and experimentation results are provided as well in order to highlight the difficulties posed by the image degradation of this dataset.

## 1. Introduction

Several datasets have been published for the task of writer identification in the recent years. Most of them contain only contemporary handwriting samples [1]–[6] where the text is easily separable from the background. An exception was the recently published Historical-WI dataset [7], which contains manuscript samples from the digital archive of the Universitätsbibliothek Basel from 13th to 20th century <https://www.e-manuscripta.ch/>. The Historical-WI dataset mostly contains samples of English language, and some of other languages (e.g. Greek and Latin), the samples of the aforementioned dataset have been selected randomly. Publishing such dataset is a step forward for the community of computational document analysis toward providing practical solutions for the scholars from the field of manuscript research. Nevertheless, these samples have been selected automatically by an algorithm with a selection criteria set by computer scientists rather than by experts of respective manuscript research. This could render such datasets irrelevant to research questions of scholars from the Humanities,

as scholars typically need support of computational methods in cases when handwriting samples have been written during the same period of time, using the same writing materials, and the same script type.

Therefore, we propose a dataset driven directly from a research question in the field of Papyrology as will be described in Section 2. All samples in the proposed dataset have been selected by experts from the respective field of research in order to apply computational-based handwriting style identification methods.

Papyri are among the most ancient writing material that has survived to our times, their number is huge and unparalleled for documents from Antiquity. These artefacts are, however, often without explicit date nor information on the identity of their writer (as opposed to colophons in medieval manuscripts). Typically, no more than a couple of samples can be assigned with confidence to the same scribe. Our dataset is small; nevertheless, its labelling is extremely solid: the texts come from the largest archive of the 6th century A.D. and are all explicitly signed.

This dataset is very interesting for the community of computational document analysis, because it requires the development of novel methods and ideas. This dataset has a very limited number of samples with a very heavy degradation, which reflects the type of challenges in many other manuscripts research fields such as old palm leave manuscripts and Chinese bamboo strips.

Identifying the handwriting style of different scribes is done until now by experts from the field of Papyrology. The question here is whether it is possible for computational methods to provide some quantitative measures of similarities that can be used as a supporting information for the task of handwriting style identification.

This paper is organised as follows: In Section 2 we present the proposed GRK-Papyri Dataset, in Section 3 we provide some experimental results for the proposed dataset, and the conclusions will be given in the final section.

## 2. The GRK-Papyri Dataset

The proposed dataset consists of 50 handwriting samples in Greek on papyrus, approximately from the 6th century A.D. All images are taken from The Bank of Papyrus Images of Byzantine Aphrodite (BIPAb: [http://www.misha.fr/papyrus\\_bipab/](http://www.misha.fr/papyrus_bipab/)); see Fig. 1 for a sample image. BIPAb is created by Prof. Jean-Luc Fournet and funded by the Association Internationale de Papyrologues and Strasbourg University. The right of reproduction and online display have been acquired and each set of images is accompanied with the photographic credit and copyright of the owner.

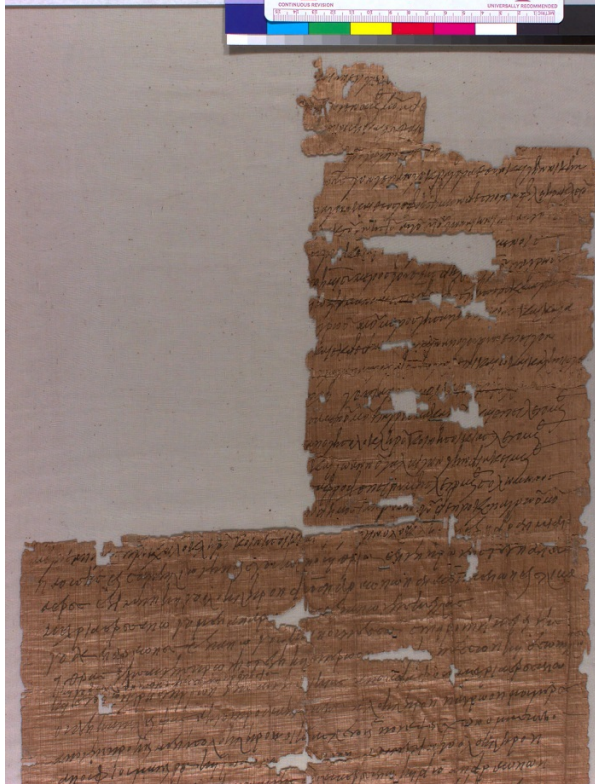


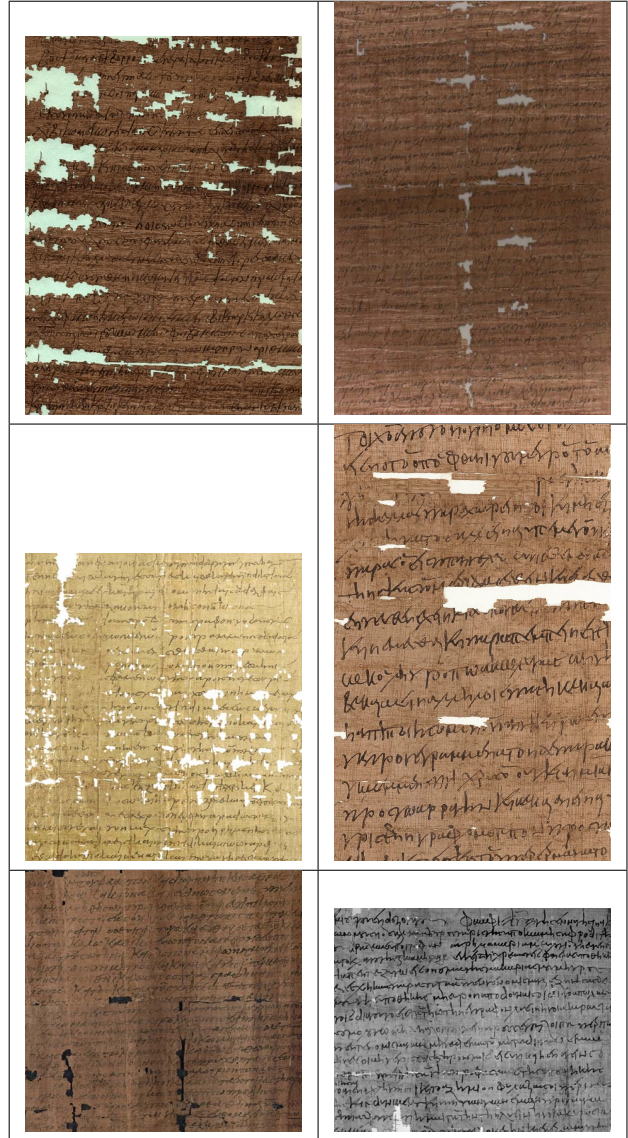
Figure 1: Sample image from BIPAb: ([http://www.misha.fr/papyrus\\_bipab/](http://www.misha.fr/papyrus_bipab/)).

The selected samples are contracts written by ten different notaries. The minimum and maximum number of samples per scribe are 4 and 7 respectively. The ground-truth of these samples is confirmed by experts using both palaeographic evidences and signatures of scribes on the contracts. This dataset is prepared and made freely available for non-commercial research along with its confirmed ground-truth related to the task of writer identification.

The selection of images is based on the relevance to scribe identification research of scholars from the field of Papyrology. In addition, the amount of handwritten text is taken into consideration so that only images with sufficient amount of text are considered. Finally, the images have been cropped so that they contain only the parts with handwritten

text written by the corresponding individuals; see the figures in Table 1 for some samples.

TABLE 1: Samples from GRK-Papyri. The original images are taken from BIPAb: [http://www.misha.fr/papyrus\\_bipab/](http://www.misha.fr/papyrus_bipab/).



The proposed dataset can be accessed and downloaded from this link: [www.d-scribes.org/en/gkr-papyri](http://www.d-scribes.org/en/gkr-papyri). Two versions of the dataset are provided, one for the leave-one-out scenario and one for the training-test scenario in order to permit the application of wider range of computational methods. The leave-one-out version contains all the 50 images together, while the training-test version contains two folders. The training folder contains 20 images, two images for each scribe in order to offer a balanced training set. The test folder contains 30 images with different number of samples per scribe. We are planning to increase the size of this dataset gradually and to add samples from different regions and periods of time.

## 2.1. Relation to Research Questions from Papyrology

Papyrology provides unparalleled data for Ancient Historians, Classicists, and scholars studying the history of law, economics, philosophy, medicine, etc. Identifying similar handwriting styles across papyrus fragments is a fundamental task for papyrologists, not only for piecing up fragments of the same document, but also to provide better understanding of its content. Gathering texts written by the same individual reveals chronological as well as rich socio-economical information. An example of the research questions from Papyrology can be found in d-scribes.org ([www.d-scribes.org](http://www.d-scribes.org)). This project aims to develop new digital palaeography approaches to identify similar scribes or handwriting styles. One of its case studies is the Dioscorus archive from the village of Aphrodito. This group of around 700 Greek papyri from the 6th A.D. addresses the specific issue of cursive handwriting, where letters can not be segmented reliably by a computational algorithm. The chosen dataset samples (from chronologically limited village archive) is well adapted to test computational approaches that can later be applied to other groups of Greek papyri. The results of such computational methods will also be relevant for several other projects working on texts written on papyri (e.g. Dead Sea scrolls in Groningen and Tel Aviv, Demotic in Heidelberg, Hieratic in Basel).

Digitisation has only started in Papyrology, but it is the natural evolution of this field, and the concern now of more and more institutions and collections around the world is to move toward virtual libraries. There are now 70,000 papyri already published, and many more are still unpublished.

Many joins have already been established in the Dioscorus archive, but a computational survey could find more. Furthermore, computational methods could help identifying identical notaries, secretaries or private individuals over different documents, and thus helping to improve our understanding of this key source for late Antiquity. Additionally, this presents a unique opportunity to study literacy levels in the Egyptian countryside and the question of the transmission of knowledge within a rural population. Can schools of scribes be spotted? How do the habits of such writers compare to inhabitants of the provincial capital of that time, Antinoopolis, or other well-documented cities like Hermopolis a couple of hundreds kilometres north?

Computational methods can give supporting information or even make it possible to give answers for the questions mentioned above and many other questions from this field of research.

## 2.2. Description of GRK-Papyri Images

The proposed GRK-Papyri (Greek on Papyri) dataset consists of heavily degraded images due to ageing and preservation conditions resulting in all kinds of distortions and ink fading. Furthermore, it consists of grey and coloured samples of different image resolution, quality, and illumination conditions. The reason is that these samples are

provided by various institutions and collections which own the papyri and are in charge of their digitisation. In addition, almost all samples contain many holes and bending marks.

All samples in the proposed dataset are in JPG format, with image resolution ranging from height of 796 to 6818 pixels and from width of 177 to 7938 pixels. The dpi is ranging from 96 to 2000. Some images are grey-scale images and others are RGB-colour space images. All samples suffer from heavy degradation including low contrast, several holes and bending marks, and even reflection of glass which covers some samples for preservation purposes. The samples are not equally distributed over the 10 scribes. The minimum and maximum number of samples per scribe are 4 and 7 respectively.

The background of the handwriting samples in this dataset typically contains papyrus fibres of different sizes and spatial frequencies. These fibres not only add complexity to the already degraded images, but also add irregularity to the overlapping contour of ink-trace. This might have negative effect on the repeatability of the keypoint detection algorithms (the possibility of detecting the same feature in different samples on the same location). Furthermore, it is worth noting that manuscripts from the same scribe and with the same exact writing material have different visual appearance due to the different preservation conditions of each sample.

The heavy degradation of samples in this dataset is not unique to handwriting on papyrus. Similar degradation can be found in old palm leave manuscripts and handwritings on Chinese bamboo strips. Therefore, developing computational solutions for this dataset could have even a wider application field for different manuscript types with similar types of degradation.

## 3. Experimental Results

Due to the challenging degradation in the proposed dataset and the low image quality of the samples, pre/post processing operations are needed to enhance the performance of computational methods. We provide here only preliminary experimental results on the GRK-Papyri dataset in order to demonstrate the complexity of the images in the proposed dataset. We hope that this actual research problem will attract the attention of researchers in the community of computational document analysis and invites computational solutions with novel ideas.

### 3.1. Image Enhancement

Given the heavy degradation of images in this dataset, applying image enhancement techniques is needed in order to enhance the performance of any computational method related to handwriting analysis. In this section, we demonstrate the impact of two image processing and enhancement techniques in a qualitative manner.

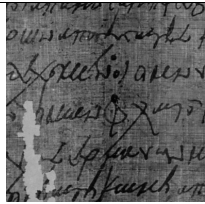
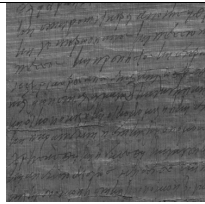
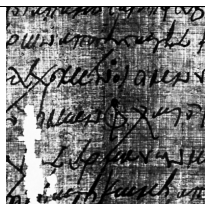
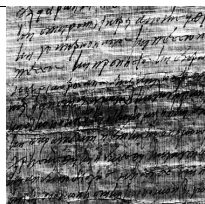
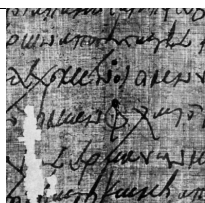
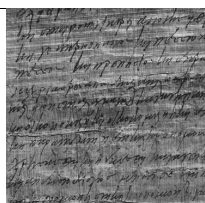
**3.1.1. Contrast Limited Adaptive Histogram Equalisation (CLAHE).** The Contrast Limited Adaptive Histogram



Equalisation (CLAHE) [8] is a modified version of the Adaptive Histogram Equalisation (AHE). AHE has a drawback of over-amplifying noise. CLAHE limits this negative effect by clipping the histogram at a specified threshold, called clip limit, before computing the Cumulative Distribution Function (CDF).

The effect of the CLAHE on papyri samples can easily be observed visually in the figures of Table 2. Images in this figure are produced using clip limit of 2.

TABLE 2: Comparison between the visual effect of the standard histogram equalisation and the CLAHE on the grey scale of two selected samples.

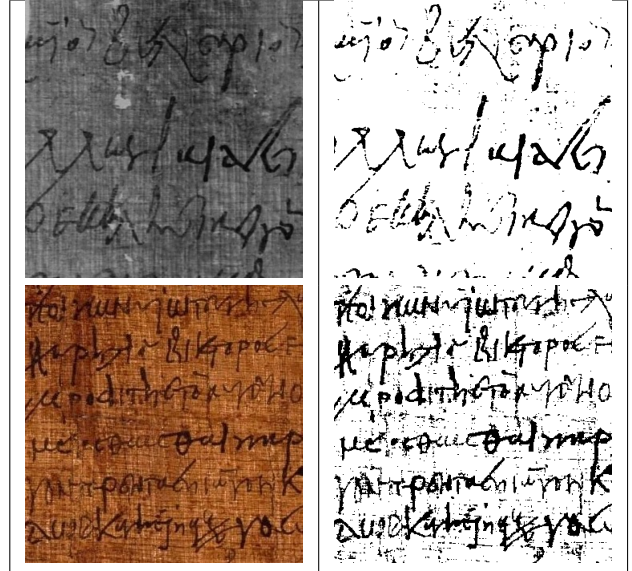
Grey Scale		
Histogram Equalisation		
CLAHE		

**3.1.2. Binarisation.** Most of state-of-the-art binarisation methods are learning-based and require several training samples. Furthermore, they typically require pixel-wise annotation and labelling which is a very demanding task that is beyond the scope of this dataset which is proposed for the task of writer identification. On the other hand, the Wolf Jolien binarisation [9] is a learning-free method that uses an adaptive threshold technique. This algorithm is developed to binarise detected text in images and videos and it can achieve very satisfactory results on challenging images and degradation conditions. Nevertheless, less impressive outcome can be obtained on this dataset as can be seen in the figures of Table 3. Better results are expected with more advanced techniques and pre/post processing operations such as bilateral filtering [10] and connected component analysis.

### 3.2. Writer Identification

Since the number of samples per writer is extremely small (between 4 and 7), learning-free methods seem to be a

TABLE 3: Parts of the samples from the proposed dataset and their binarisation results using the Wolf Jolien binarisation method [9].

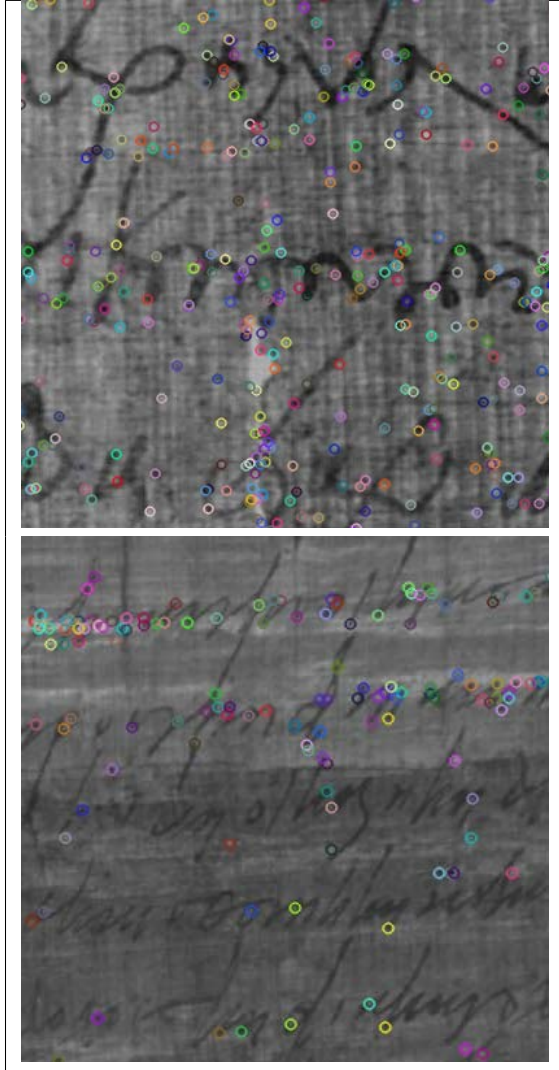


practical choice for this dataset; nevertheless, novel training techniques could offer some solutions. Furthermore, character/word segmentation would be extremely difficult given the heavy degradation and that the handwriting is cursive in these samples. Also, we suspect that extracting contours of letters would not be reliable as well. Therefore, we carried out basic tests using the Normalised Local NBNN [11] as a learning- and segmentation-free method which does not require any contour extraction. Tests have been done without any pre- or post-processing.

Although Normalised Local NBNN achieved state-of-the-art results in both contemporary [11] and historical [12] datasets, this method was not able to cope with the degradation types posed by the proposed dataset. Our experimentation results showed that the Normalised Local NBNN with FAST keypoints using the same parameters used in [12] resulted in only 30.0% identification rate on the GRK-Papyri dataset with leave-one-out criteria, and only 26.6% identification rate with training-test criteria.

We suspect that several parts of the Normalised Local NBNN method contribute to the very low identification rate on this dataset. For example, the FAST keypoints in the figures of Table 4 clearly not concentrated on the contour of the ink trace compared to the case shown on samples from ICFHR-2016 dataset [6] and St. Gall dataset [13] in [12], see reproduced figures in Table 5 from the aforementioned publication. Furthermore, the repeatability of the detected FAST keypoints is expected to be very low even if they were located on the contour of the ink-trace due to the rich texture of the background, partially due to papyrus fibres.

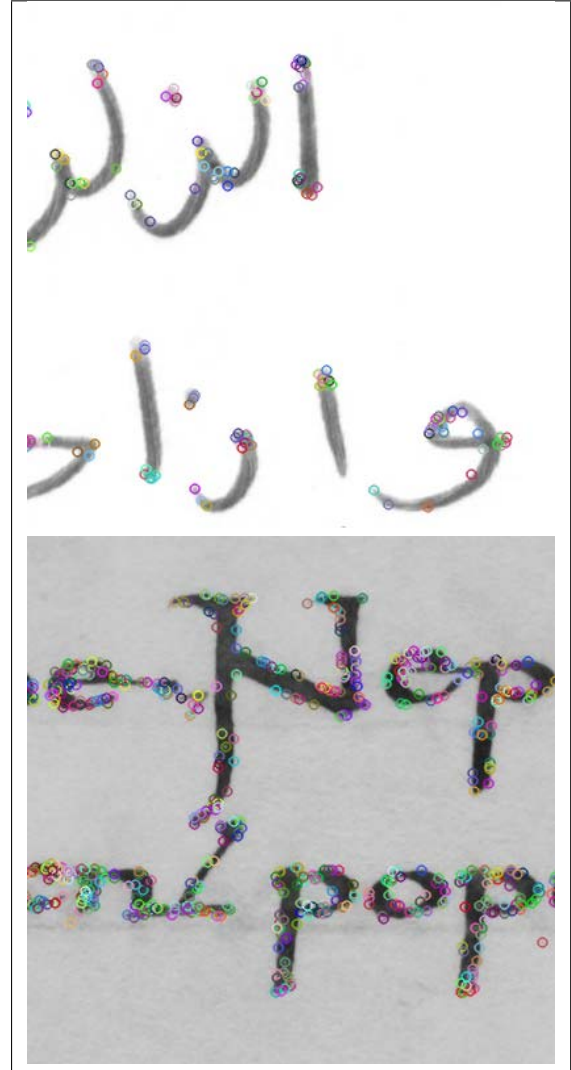
TABLE 4: FAST keypoints visualised as coloured circles on parts of samples from the proposed dataset. Only 5% of the detected keypoints using the same criteria as proposed in [12].



## 4. Conclusion

We present in this paper a dataset of Greek handwriting on papyri for the task of writer identification. The proposed dataset is derived from research questions from the field of Papyrology. Furthermore, the samples of this dataset are selected by scholars from the respective research field in order to insure that the dataset is representative of actual research problems. The ground-truth information for the task of writer identification has been confirmed both by palaeographic evidence and by the signatures of scribes on the contracts themselves. The dataset is prepared and published to be used by researchers for further experimentation. Two versions of the dataset are provided one for leave-one-out and one for train-test scenario.

TABLE 5: FAST keypoints visualised as coloured circles. Figures are reproduced from [12].



Writer identification is cited as a use case in order to guide the selection of data; nevertheless, this dataset can be used for different research topics such as image enhancement, binarisation, and line/word segmentation.

Preliminary experiments are presented in order to demonstrate the complexity of degradation in the samples of this dataset, and the need for new computational solutions and/or pre- and post-processing operations. We are planning to enrich this dataset with more samples from different regions and periods of time, and with different research questions as well.

## Acknowledgement

This work has been funded by the German Research Foundation (DFG) of the Sonderforschungsbereich (SFB 950) within the scope of the Centre for the Study of

Manuscript Cultures (CSMC) at Universität Hamburg. In addition, it has been funded by the SNSF as part of Ambizione project n° PZ00P1\_174149 “Reuniting fragments, identifying scribes and characterizing scripts: the Digital paleography of Greek and Coptic papyri”.

## References

- [1] G. Louloudis, N. Stamatopoulos, and B. Gatos, “Icdar 2011 writer identification contest,” in *Document Analysis and Recognition (ICDAR), International Conference on*. IEEE, 2011, pp. 1475–1479.
- [2] A. Fornes, A. Dutta, A. Gordo, and J. Lladós, “The icdar 2011 music scores competition: Staff removal and writer identification,” in *Document Analysis and Recognition (ICDAR), International Conference on*. IEEE, 2011, pp. 1511–1515.
- [3] G. Louloudis, B. Gatos, and N. Stamatopoulos, “Icfhr 2012 competition on writer identification challenge 1: Latin/greek documents,” in *Frontiers in Handwriting Recognition (ICFHR), International Conference on*. IEEE, 2012, pp. 829–834.
- [4] G. Louloudis, B. Gatos, N. Stamatopoulos, and A. Papandreou, “Icdar 2013 competition on writer identification,” in *Document Analysis and Recognition (ICDAR), 12th International Conference on*. IEEE, 2013, pp. 1397–1401.
- [5] F. Kleber, S. Fiel, M. Diem, and R. Sablatnig, “Cvl-database: An off-line database for writer retrieval, writer identification and word spotting,” in *Document Analysis and Recognition (ICDAR), 12th International Conference on*. IEEE, 2013, pp. 560–564.
- [6] C. Djeddi, S. Al-Maadeed, A. Gattal, I. Siddiqi, A. Ennaji, and H. El Abed, “ICFHR2016 competition on multi-script writer demographics classification using” quwi” database.”
- [7] S. Fiel, F. Kleber, M. Diem, V. Christlein, G. Louloudis, S. Nikos, and B. Gatos, “Icdar2017 competition on historical document writer identification (historical-wi),” in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2017, pp. 1377–1382.
- [8] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, “Adaptive histogram equalization and its variations,” *Computer vision, graphics, and image processing*, vol. 39, no. 3, pp. 355–368, 1987.
- [9] C. Wolf, J.-M. Jolion, and F. Chassaing, “Text localization, enhancement and binarization in multimedia documents,” in *ICPR*, 2002.
- [10] F. Banterle, M. Corsini, P. Cignoni, and R. Scopigno, “A low-memory, straightforward and fast bilateral filter through subsampling in spatial domain,” *Computer Graphics Forum*, vol. 31, no. 1, pp. 19–32, 2012. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8659.2011.02078.x>
- [11] H. Mohammed, V. Märgner, T. Konidaris, and H. S. Stiehl, “Normalised local naïve bayes nearest-neighbour classifier for offline writer identification,” in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2017, pp. 1013–1018.
- [12] H. Mohammed, V. Märgner, and H. S. Stiehl, “Writer identification for historical manuscripts: Analysis and optimisation of a classifier as an easy-to-use tool for scholars from the humanities,” in *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Aug 2018, pp. 534–539.
- [13] e-codices Virtual Manuscript Library of Switzerland. St. gallen, stiftsbibliothek. [Online]. Available: <http://www.e-codices.ch>