PapyRow: a dataset of row images from ancient Greek Papyri for writers identification*

Nicole Dalia Cilia[1], Claudio De Stefano[1], Francesco Fontanella[1], Isabelle Marthot-Santaniello[2], and Alessandra Scotto di Freca[1]

[1] University of Cassino and Southern Lazio, Via di Biasio 43, Italy
nicoledalia.cilia|destefano|f.fontanella|a.scotto@unicas.it
[2] Universität Basel, Basel, Switzerland
i.marthot-santaniello@unibas.ch

Abstract. Papyrology is the discipline that studies texts written on ancient papyri. An important problem faced by papyrologists and, in general by paleographers, is to identify the writers, also known as scribes, who contributed to the drawing up of a manuscript. Traditionally, paleographers perform qualitative evaluations to distinguish the writers, and in recent years, these techniques have been combined with computer-based tools to automatically measure quantities such as height and width of letters, distances between characters, inclination angles, number and types of abbreviations, etc. Recently-emerged approaches in digital paleography combine powerful machine learning algorithms with high-quality digital images. Some of these approaches have been used for feature extraction, other to classify writers with machine learning algorithms or deep learning systems. However, traditional techniques require a preliminary feature engineering step that involves an expert in the field. For this reason, publishing a well-labeled dataset is always a challenge and a stimulus for the academic world as researchers can test their methods and then compare their results from the same starting point. In this paper, we propose a new dataset of handwriting on papyri for the task of writer identification. This dataset is derived directly from GRK-Papyri dataset and the samples are obtained with some enhancement image operation. This paper presents not only the details of the dataset but also the operation of resizing, rotation, background smoothing, and rows segmentation in order to overcome the difficulties posed by the image degradation of this dataset. 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.

Keywords: Greek Papyri · Row extraction · Dataset · Writer Identification.

* This work was supported in part by the Italian Ministry of University and Research, grant "Dipartimenti di Eccellenza 2018–2022', and the Swiss National Science Foundation as part of the project n° PZ00P1-174149 "Reuniting fragments, identifying scribes and characterizing scripts: the Digital paleography of Greek and Coptic papyri"

1 Introduction

The large majority of databases publicly available for writer identification tasks, are composed of contemporary handwritings [21, 14, 19, 20, 17, 12]. Historical documents, however, present specific difficulties (limited amount of data, complex material, degradation) that digital technologies must be able to cope with, in order to be successfully applied by Historians to actual research problems.

An exception in this context is represented by the Historical-WI dataset [13], containing about 3600 handwritten pages selected from manuscripts ranging from the thirteenth to the sixteenth century, which were produced by 720 different scribes. The main problem in the use of these data is that most of the documents were automatically selected by an algorithm made by computer scientists and not by manuscript experts: therefore, it is not tailored to answer actual research questions (see [22], p. 726).

Moreover, as discussed in [27], when computerized writer identification approaches are tried on historical texts, it is required that the handwriting samples are previously associated with the corresponding writers, in order to verify the correctness of the automatic classification results: such labeling often relies on the judgement of ancient writings specialists (epigraphers, paleographers) or forensic experts, which are all subjective and often not in accordance among them.

Recently-emerged approaches in digital paleography combine powerful machine learning algorithms with high-quality digital images. Machine-learning-based approaches have received increasing attention from researchers, thanks to their ability to handle complex and difficult image classification tasks for writer identification [3, 30, 11, 7]. Some of these methodologies have been used for feature selection [4] and extraction [18, 16], others for image segmentation in lines in order to improve or enlarge the dataset [23, 25, 24], or for studying the most appropriate dataset size [8, 5]. Finally, in recent years deep learning systems are not only tested for identifying sundry elements of interest inside document pages, but also with a specific focus on writer recognition problems [2, 6, 9, 10].

Ancient texts from antiquity survived only in rare cases, the wide majority as carved inscriptions on stone, in more specific contexts as impressions on clay tablets and writings on potsherds (ostraca[1]).

Papyri from Egypt stand as an exception: thanks to the dryness of the climate, an unparalleled quantity of texts reached out to us, literary pieces as well as documents from daily life. Papyri written in Ancient Greek cover a millennium of Egypt History, from Alexander the Great at the end of the fourth century BCE to the Arab conquest in the middle of the seventh century CE. There are already more than 70,000 published texts and hundreds of thousands still waiting for publication in museums and libraries, not to mention newly excavated material each year [28]. Therefore, papyri offer a unique opportunity to analyse ancient handwritings but with specific constraints: by definition a

[1] An ostracon (Greek term, whose plural is ostraca) is a piece of pottery on which an inscription is engraved. Usually these pieces are fragments of broken pottery vessels, on which inscriptions were subsequently made.

papyrus leaf, made of strips cut in the plant stem, is a brownish, complex background. The state of preservation is usually bad (breaks, holes, mud, erasure of ink) and the amount of text is limited, not to be compared with hundred-page long medieval manuscripts. Furthermore, the majority of papyri was acquired by collections in the antiquities market, so without archaeological context and often voluntarily broken into pieces to increase their sale value. Scholars have to reconstruct original documents as well as piece up ancient archives. Being able to identify the same writer over several fragments allow substantial progress in our understanding of Egypt ancient History.

For all these reasons, publishing a well-labeled dataset is always a challenge and a stimulus for the academic world, as researchers can test their methodologies and then compare their results from the same starting point. In this paper, we propose a new dataset of handwriting on Ancient Greek papyri for the task of writer identification. This dataset is derived directly from GRK-Papyri dataset [22] and the samples are obtained through complex image enhancement operations. Moreover, it is made of objectively attributed documents thanks to inner evidence, the subscription - signature - of their writers: reference data are therefore available to verify the actual classification results.

The samples have been selected to reflect real investigations from scholars, thus representing a severe test bed for writer identification algorithms: they come from the same kind of documents, the same period of time and the same geographic area but the images are from various collections (museums, libraries) and therefore vary greatly (see below, in Section 2). What is at stake is to be able to differentiate the handwriting of an individual from the ones of his contemporaries in a writing context as homogeneous as possible. If this extreme case gives successful results, then heterogeneous cases (different handwriting styles, periods, locations) and homogeneous data (same digitization process) should yield even better results.

The remainder of the paper is organized as follows: after a description of the original dataset GRK-Papyri in Section 2 and its extension, we present the operation of background smoothing in Section 3.1, resizing in Section 3.2, rotation in Section 3.3, and rows segmentation in Section 3.4, in order to overcome the difficulties posed by the image degradation of this dataset. Finally, in Section 4 we present the obtained row dataset, which is freely available for non-commercial research along with the confirmed ground-truth information related to the task of writer identification.

2 Description of the original dataset

The reference dataset used in this work is composed of images depicting Greek papyri dating back to the 6th century. Writer identification is ensured by the presence of notary subscriptions, whose analysis can be found in [29] and in a forthcoming article on Notaries in Aphrodito by Isabelle Marthot-Santaniello. Images have been checked, selected, cataloged and cropped (see [22]) to keep only the part written by the notary and not by potential parties and witnesses.

Table 1. Distribution of papyri by authors

Authors	Number of papyri
Amais, Anouphis, Daueit, Kyros2, Victor2	1
Ieremias, Kollouthos, Konstantinos, Psates	2
Philotheos	3
Andreas, Apa Rhasios	4
Dioscorus, Hermauos, Kyros3, Menas, Theodosios	5
Isak	8
Kyros1	9
Pilatos, Victor1	10
Dios	15
Abraamios	21

Almost all the documents used are part of the richest archive of the Byzantine period belonging to Dioscorus of Aphrodito, which collects more than 700 texts (see [26] and [15]). All the images of this archive are accessible in [1]. The only exception is for the writer Menas, who lived in the same period and practises the same profession but living in Hermopolis, a city roughly 150 km north of Aphrodito. His dossier is about to be published by Isabelle Marthot-Santaniello while the images are provided by BNUS, Strasbourg (Menas 1 and 2) and the British Library (Menas 3, 4 and 5).

The basis of the present dataset is the GRK-Papyri, used for the task of identifying the writers and composed of 50 images distributed unequally among the 10 writers present [22]. Other papyri have been added to this starting set, increasing the number of papyri of the above writer, as well as introducing new writers. Summarizing, a number of papyri images equal to 122 was reached for a total number of 23 writers, further increasing the imbalance in the number of documents per author. It is useful to note that not all the images contain the same amount of text: in fact some of them include few fragments of text, while others contain larger sequences of lines.

Table 1 reports detailed information on the distribution of papyri by writers. Metadata on each papyrus image (publication name of the papyrus, reference on the writer attribution, copyright) will be available in a dedicated section of d-scribes.org.

The papyri appear to be very degraded by time: in fact for most of them there are missing pieces or holes that extend almost everywhere, causing the loss of the writing trace (see Fig.1).

In addition to the characteristics already listed, there are also different contrasts, different lighting conditions and in some cases even reflections due to the presence of glass used to preserve them. Moreover, there are both color and grayscale images (as shown in Fig.2). All the samples are in JPG format, with a resolution that varies both in height and in width.

To achieve the goal of enhancing image quality, we applied a procedure based on the use of standard tools, such as LabelImg and GIMP, and python scripts

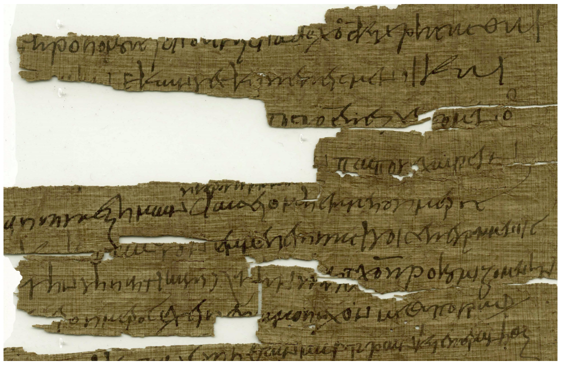

Fig. 1. Examples of original papyri.

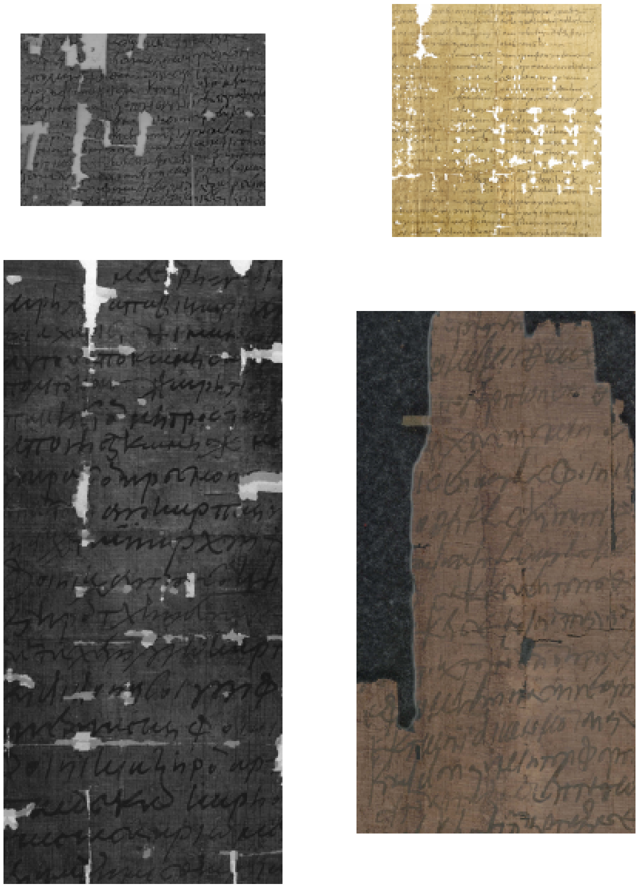

Fig. 2. Examples of original papyri.

with specific libraries for images processing: details are reported in the next Section.

3 Image Enhancement

Given the heavy degradation of images in this dataset, applying image enhancement techniques is needed in order to improve the performance of any computational method related to handwriting analysis.

3.1 Background smoothing

The first step tries to uniform as much as possible the image background. For this purpose, it was necessary to fill all the holes in the image with the same color as the papyrus; this operation allows the elimination of any unwanted colors, which could adversely affect the following phases of image enhancement. This procedure, carried out for both color and grayscale images, was performed using the GIMP program (GNU Image Manipulation Program) which allows us to perform manipulations on digital images in a very simple and fast way. The result obtained is showed in the Fig. 3.

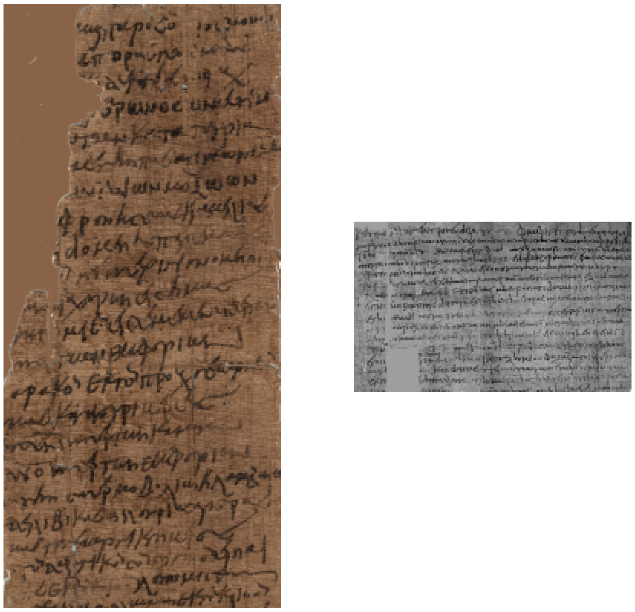

Fig. 3. An example of unwanted colors elimination.

3.2 Line resizing

The next phase consists of various operations that allow us to adequately prepare images for the labeling, i.e. for the selection and the extraction of individual lines in each papyrus, to be assigned to the actual writer. As previously mentioned, each image has its own dimension, which is different from that of all the other images. However, in order to be processed and compared properly, all the lines must have similar dimensions: therefore it was necessary to resize all the lines to an appropriate size.

For each papyrus image, the heights of the individual lines were taken and a resize was made in proportion to the maximum line height identified among all the papyri images. Once the new height relative to an image has been identified, the new width is calculated in such a way as to preserve the original aspect ratio (ratio between width and height).

3.3 Image rotation

Before proceeding with the labeling phase, some rotations were made (with the GIMP program) on papyrus images whose text was too skewed (see Fig.4). The rotation was necessary for using the LabelImg tool, through which it is possible to correctly label the lines of the papyri. Among all the available images, only few of them required this rotation. Furthermore, color images were converted in grayscale ones, using Python libraries (see Fig.5): thus, all the images included in our dataset are in grayscale format.

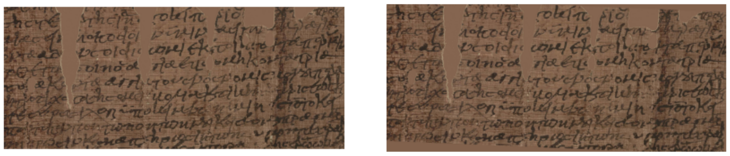

Fig. 4. An example of an image before and after rotation.

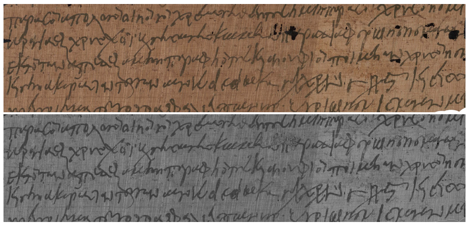

Fig. 5. An example of an image before and after grayscale conversion.

3.4 Row labeling

The row labeling operation was manually performed with LabelImg tool, taking care of not including part of the image where there is no writing. The tool allows us to create rectangles (bounding box) surrounding the rows to be selected (see Fig.6). Following this operation, LabelImg generates a file with XML extension, containing all the details of the bounding boxes such as, for example, the start and end coordinates of every one.

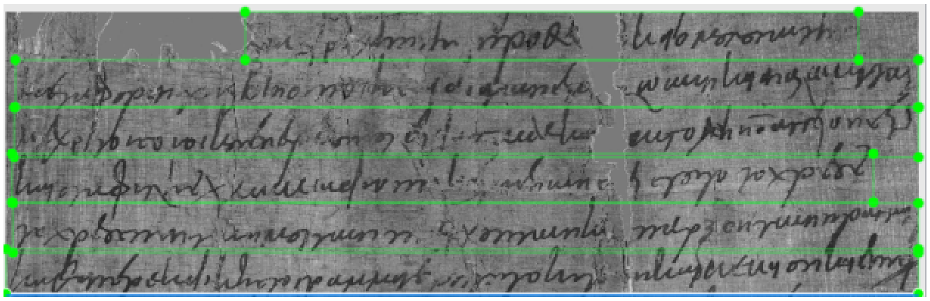

Fig. 6. Image divided into bounding box rows.

Using both the information included in this XLM file and the python libraries, it is possible to cut out each selected element and save it separately: the result of this process is a sequence of images, each containing a single row, for each original papyrus image. Moreover, since the images obtained according to the above process exhibit a very large width variation, two different procedures were designed for generating images of similar width, namely Pre-processing 1 end Pre-processing 2.

- Pre-processing 1: in this case the line image with the lowest width among all the available ones was identified (this width value is equal to 1232 pixels in our case) and all the other images were divided according to this width. This implies that, the final part of a line image may be discarded if its width is lower than the above value (see Fig.7).
- Pre-processing 2: in this second option all the lines have been cut assuming a predetermined threshold value for the width (500 pixels). The result is a larger number of line images, possibly reducing the loss of useful writing (see Fig.8).

4 Description of the final dataset and conclusion

In this paper, we have proposed a new dataset of handwriting on Ancient Greek papyri for the task of writer identification: the PapyRow Dataset. The samples, derived directly from the extended GRK-Papyri dataset, are obtained with

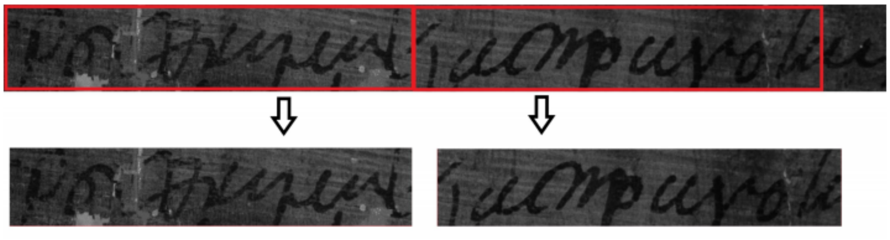

Fig. 7. Image cutting according to Pre-processing 1.

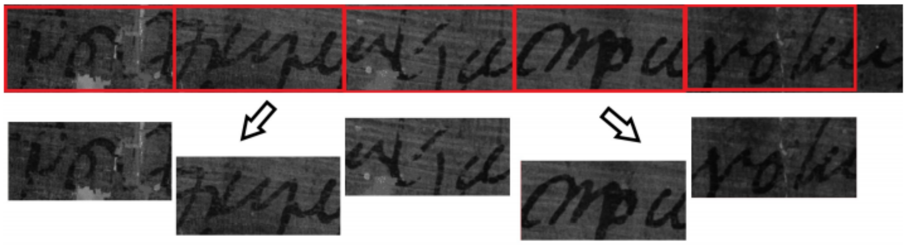

Fig. 8. Image cutting according to Pre-processing 2.

some enhancement image operation. In particular, as the papyri appear to be very degraded by time, with missing pieces, holes, different contrasts and different lighting conditions, we have employed a background smoothing. Moreover, since image dimensions were very different, they have been resized preserving the original aspect ratio, but allowing the line height to be very similar across all images. Image rotation have been also applied for correcting too skewed images and a format conversion have been performed on color images to obtain a database composed of all grayscale images.

Finally, we have used a rows segmentation procedure to generate the PapyRow Dataset. This dataset, including 6498 samples, is freely available for non-commercial research along with the confirmed ground-truth information related to the task of writer identification, at link:
http://webuser.unicas.it/fontanella/papyri/
The distribution of the samples among all the writers is reported in the histogram of Fig.9.

The row segmentation have been performed for two main reasons: on one hand, for increasing the number of samples available for the training phase of the considered classifiers and, on the other hand, to discard the parts of the original images that contain no text.

Moreover, as previously noticed, not all the images contain the same amount of text: in fact some of them include few fragments of text, while others contain larger sequences of lines.

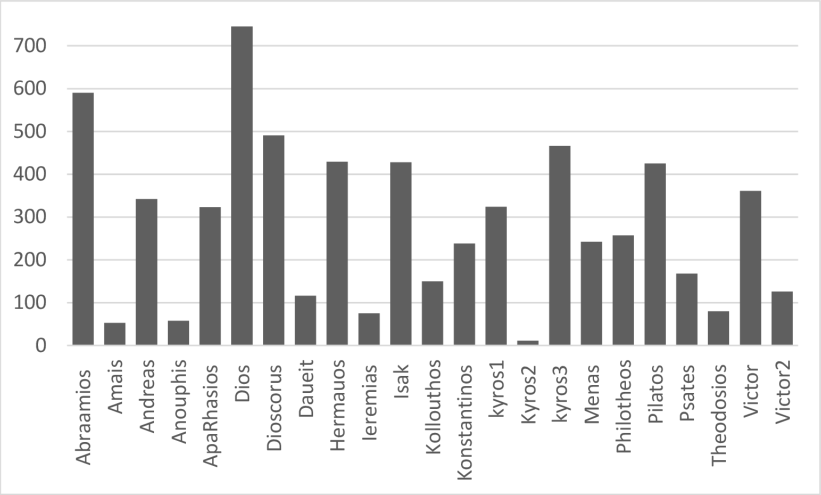

Fig. 9. Sample distribution among all the writers.

References

1. Bipab (ed.): The Bank of Papyrus Images of Byzantine Aphrodite BIPAb (00), (http://bipab.aphrodito.info//
2. Bria, A., Cilia, N., De Stefano, C., Fontanella, F., Marrocco, C., Molinara, M., Scotto di Freca, A., Tortorella, F.: Deep transfer learning for writer identification in medieval books. In: 2018 IEEE International Conference on Metrology for Archaeology and Cultural Heritage. pp. 455–460 (2018)
3. Bulacu, M., Schomaker, L.: Text-independent writer identification and verification using textural and allographic features. IEEE Trans. Pattern Anal. Mach. Intell. **29**(4), 701–717 (2007)
4. Cilia, N., De Stefano, C., Fontanella, F., Scotto di Freca, A.: A ranking-based feature selection approach for handwritten character recognition. Pattern Recognition Letters (2018)
5. Cilia, N., De Stefano, C., Fontanella, F., M., M., Scotto di Freca, A.: Minimizing training data for reliable writer identification in medieval manuscripts. Lecture Notes in Computer Science, ICIAP Proceedings **11808**, 198–208 (2019)
6. Cilia, N., De Stefano, C., Fontanella, F., M., M., Scotto di Freca, A.: An end-to-end deep learning system for medieval writer identification. Pattern Recognition Letters **129**, 137–143 (2020)
7. Cilia, N., De Stefano, C., Fontanella, F., M., M., Scotto di Freca, A.: An experimental comparison between deep learning and classical machine learning approaches for writer identification in medieval documents. Journal of Imaging **6(9)**(89) (2020)
8. Cilia, N., De Stefano, C., Fontanella, F., M., M., Scotto di Freca, A.: What is the minimum training data size to reliably identify writers in medieval manuscripts? Pattern Recognition Letters **129**, 198–204 (2020)
9. Cilia, N., De Stefano, C., Fontanella, F., Marrocco, C., M., M., Scotto di Freca, A.: A page-based reject option for writer identification in medieval books. Lecture Notes in Computer Science, ICIAP Proceedings **11808**, 198–208 (2019)

10. Cilia, N., De Stefano, C., Fontanella, F., Marrocco, C., M., M., Scotto di Freca, A.: A two-step system based on deep transfer learning for writer identification in medieval books. Lecture Notes in Computer Science, CAIP Proceedings **11679**, 305–316 (2019)
11. Dahllof, M.: Scribe Attribution for Early Medieval Handwriting by Means of Letter Extraction and Classification and a Voting Procedure for Larger Pieces. In: Proceedings of the 22nd International Conference on Pattern Recognition. pp. 1910–1915. IEEE Computer Society (2014). https://doi.org/10.1109/ICPR.2014.334
12. Djeddi, C., Al-Maadeed, S., Gattal, A., Siddiqi, I., Ennaji, A., El Abed, H.: Icfhr 2016 competition on multi-script writer demographics classification using" quwi" database. ICFHR Proceedings IEEE **01**, 602–606 (2016)
13. Fiel, S., Kleber, F., Diem, M., Christlein, V., Louloudis, G., Nikos, S., B., G.: Icdar 2017 competition on historical document writer identification (historical-wi). Proceedings of the 2017 International Conference on Document Analysis and Recognition (ICDAR) pp. 1377–1382 (2017)
14. Fornes, A., Dutta, A., Gordo, A., J., L.: The icdar 2011 music scores competition: Staff removal and writer identification. Proceedings of the 2011 International Conference on Document Analysis and Recognition, ICDAR p. 1511–1515 (2011)
15. Fournet, J. (ed.): Les archives de Dioscore d'Aphrodité cent ans après leur découverte, histoire et culture dans l'Égypte byzantine. Actes du Colloque de Strasbourg. Études d'archéologie et d'histoire ancienne, Paris (2008)
16. Joutel, G., Eglin, V., Bres, S., Emptoz, H.: Curvelets based feature extraction of handwritten shapes for ancient manuscripts classification. In: Document Recognition and Retrieval XIV, San Jose, California, USA, January 30 - February 1. pp. 65000D 1–12 (2007)
17. Kleber, F., Fiel, S., Diem, M., R., S.: Cvl-database: An off-line database for writer retrieval, writer identification and word spotting. Proceedings of the 2013 International Conference on Document Analysis and Recognition, ICDAR p. 560–564 (2013)
18. Liang, Y., Fairhurst, M.C., Guest, R.M., Erbilek, M.: Automatic handwriting feature extraction, analysis and visualization in the context of digital palaeography. IJPRAI **30**(4), 1653001 1–26 (2016)
19. Louloudis, G., Gatos, B., N., S.: Icfhr 2012 competition on writer identification challenge 1: Latin/greek documents. Frontiers in Handwriting Recognition (ICFHR) p. 829–834 (2012)
20. Louloudis, G., Gatos, B., N., S., Papandreou, A.: Icdar 2013 competition on writer identification. Proceedings of the 2013 International Conference on Document Analysis and Recognition, ICDAR p. 1397–1401 (2013)
21. Louloudis, G., Stamatopoulos, N., B., G.: Icdar 2011 writer identification contest. Proceedings of the 2011 International Conference on Document Analysis and Recognition, ICDAR p. 1475–1479 (2011)
22. Mohammed, H., Marthot-Santaniello, I., V., M.: Grk-papyri: A dataset of greek handwriting on papyri for the task of writer identification. Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR) pp. 726–731 (2019)
23. Papavassiliou, V., Stafylakis, T., Katsouros, V., Carayannis, G.: Handwritten document image segmentation into text lines and words. Pattern Recognition **43**(1), 369–377 (2010)
24. Pintus, R., Yang, Y., Gobbetti, E., Rushmeier, H.E.: A talisman: Automatic text and line segmentation of historical manuscripts. In: 2014 Eurographics Workshop

on Graphics and Cultural Heritage, GCH 2014, Darmstadt, Germany, October 6-8, 2014. pp. 35–44 (2014)

25. Pintus, R., Yang, Y., Rushmeier, H.E.: ATHENA: automatic text height extraction for the analysis of text lines in old handwritten manuscripts. JOCCH **8**(1), 1:1–1:25 (2015)
26. Ruffini, G. (ed.): Life in an Egyptian Village in Late Antiquity: Aphrodito Before and After the Islamic Conquest. Cambridge; New York: Cambridge University Press (2018)
27. Shaus, A., Gerber, Y., Faigenbaum-Golovin, S., Sober, B., E., P., I., F.: Forensic document examination and algorithmic handwriting analysis of judahite biblical period inscriptions reveal significant literacy level. PLoS ONE **15(9): e0237962** (2020)
28. Van Minnen, P. (ed.): The Future of Papyrology. in Bagnall RS, ed., The Oxford Handbook of Papyrology, Oxford: Oxford Univ. Press (2009)
29. Worp, K., Diethart, J. (eds.): Notarsunterschriften im Byzantinischen Ägypten (1986)
30. Yosef, I.B., Beckman, I., Kedem, K., Dinstein, I.: Binarization, character extraction, and writer identification of historical Hebrew calligraphy documents. IJDAR **9**(2-4), 89–99 (2007)