A System for Processing and Recognition of Greek Byzantine and Post-Byzantine Documents

Panagiotis Kaddas[1,2(✉)], Konstantinos Palaiologos[1,3], Basilis Gatos[1], Vassilis Katsouros[4], and Katerina Christopoulou[1,5]

[1] Computational Intelligence Laboratory, Institute of Informatics and Telecommunications, NCSR "Demokritos", 15310 Athens, Greece
{pkaddas,k.palaiologos,bgat,achristopoulou}@iit.demokritos.gr
[2] Department of Informatics and Telecommunications, University of Athens, 15784 Athens, Greece
[3] Hellenic Institute, Royal Holloway, University of London, Egham Hill, Egham TW20 0EX, Surrey, UK
[4] Institute for Language and Speech Processing, Athena Research Center, Athens, Greece
vsk@athenarc.gr
[5] School of Environment, Geography and Applied Economics, Department of Economics & Sustainable Development, Harokopio University, 17676 Athens, Greece

Abstract. Processing and recognition of Greek Byzantine and Post-Byzantine (old Greek) Documents has been proven to be a tedious task in the domain of Historical Document Image Processing. Several unique characteristics of these documents (existence of character ligatures, abbreviations, lack of clear word division, existence of symbols or punctuations in an arbitrary position) impose significant difficulties for current processing and recognition tools. In this work, we introduce a system for processing and recognition of old Greek documents and give details about all the components that comprise it. These include an image pre-processing, a text line segmentation and a recognition module. In order to test the proposed system, we introduce and provide publicly a new dataset of old Greek Documents that includes text line images and the corresponding transcription. Using this dataset, we evaluate the embedded recognition engine of the proposed system which is the open-source Calamari-OCR engine employing a variety of configurations. The best result corresponded to a character error rate less than 1.5% which is acceptable and promising. Finally, we also achieved promising results when comparing the embedded OCR engine with other recognition methods already proposed for the recognition of old Greek Documents.

Keywords: Document Analysis System · Deep Neural Networks · Calamari-OCR · Greek Byzantine and Post-Byzantine Documents · Text Line Recognition

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
G. A. Fink et al. (Eds.): ICDAR 2023, LNCS 14190, pp. 366–376, 2023.
https://doi.org/10.1007/978-3-031-41685-9_23

1 Introduction

Old Greek Documents are an important source of historical information for scholars related to our cultural heritage conservation. In this paper, we focus on processing and recognition of Greek Byzantine and Post-Byzantine (old Greek) documents dated from the 12th to the 16th century. As it can be observed in the sample of Fig. 1, old Greek documents of this period have some unique characteristics such as the existence of character ligatures (neighboring character maybe joined together), abbreviations, lack of clear word division, existence of symbols or punctuations in an arbitrary position, which impose significant difficulties for current optical character recognition tools.

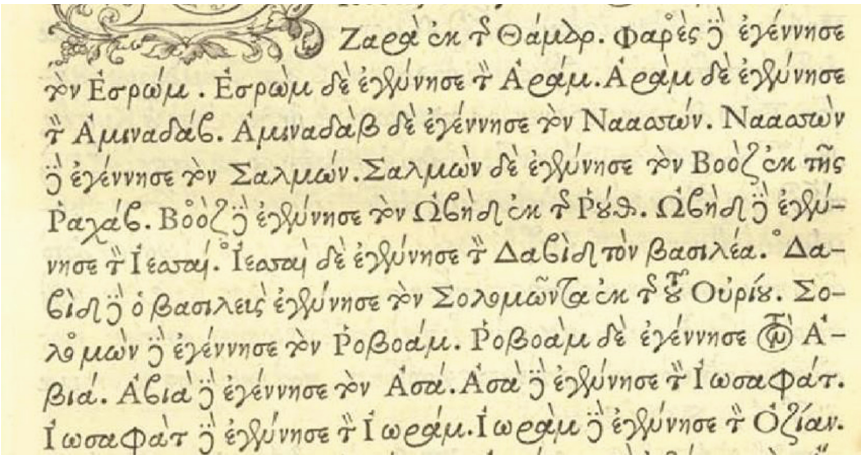


Fig. 1. A sample of Greek early printing document (*grecs du roi* typeface, Greek New Testament, published by Robert Estienne in 1550).

In this work, we introduce a system for processing and recognition of old Greek documents. We give details about all the components that comprise it focusing both on the automatic procedures for image pre-processing, text line segmentation and recognition, as well as on the semi-automatic procedures for correcting the text line segmentation and recognition result. The embedded recognition engine of the proposed system is the open-source, TensorFlow-based Calamari-OCR engine [1] that uses an advanced deep neural network. In order to test the recognition engine, we introduce and provide publicly a new dataset of old Greek Documents [2] that includes text line images and the corresponding transcription. By employing a variety of configurations on the recognition engine, we demonstrate that we can achieve very promising results using a small number of images for training. The best result obtained corresponds to a character error rate less than 1.5%. Finally, we also achieved promising results when comparing the embedded OCR engine with other recognition methods already proposed for the recognition of old Greek Documents.

The rest of the paper is organized as follows. In Sect. 2, the related work is presented, Sect. 3 introduces the proposed system, Sect. 4 demonstrates our experimental results and Sect. 5 presents the conclusion of this work.

2 Related Work

Processing and recognition of old Greek Documents has not attracted lot of attention in the literature. There are some approaches that follow more traditional image processing techniques based on feature extraction and some more recent techniques based on Convolutional Neural Networks (CNN).

In approaches [3, 4], the document image is first binarized, enhanced and skeletonized. Next, the open and closed cavities of the skeletonized characters are detected and a feature extraction step is applied in order to provide the input for the recognition process. Finally, the individual cavities are recognized on the basis of their features. At the feature extraction step, all segments that belong to a protrusion of an isolated character's cavity are calculated. For the classification step, decision trees, the K-NN classifier and support vector machines (SVMs) are employed. The corpus used for the experiments originates from the Sinaitic Codex Number Three, the Book of Job collection written by three different writers.

CNNs are used in approaches [5, 6]. In [5], a convolutional recurrent neural network architecture is proposed that comprises octave convolution and recurrent units which use effective gated mechanisms. The proposed architecture has been evaluated on three newly created collections from Greek historical handwritten documents as well as on standard datasets like IAM and RIMES. In [6], the focus is on the effort to automate transcription of Greek paleographic manuscripts dating from the 10^{th} to the 16^{th} century. To this end, two datasets with a parallel corpus of transcriptions were introduced and the experiments were done using an AI powered handwritten text recognition tool based on the Transkribus tool [7].

3 Proposed System

In this Section, we give details about all the components that comprise the proposed system for processing and recognition of old Greek documents. This includes the image pre-processing component that performs image binarization, as well as the text line detection and the text line recognition components applied both in an automatic and a semi-automatic way.

3.1 Image Pre-processing

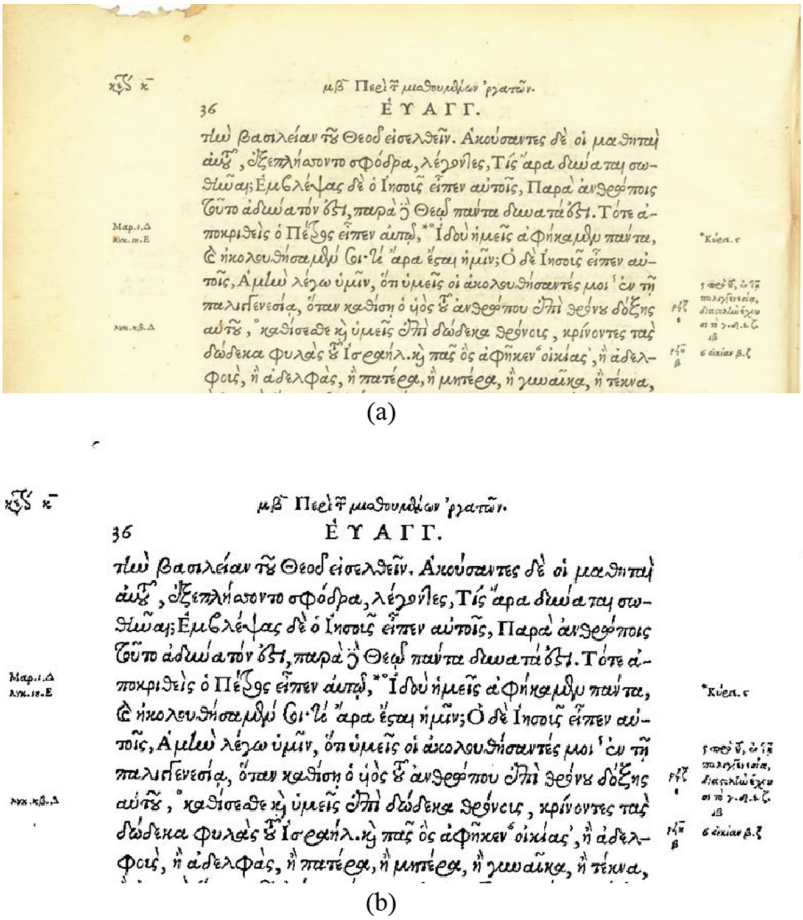
Image pre-processing includes image binarization that refers to the conversion of the grayscale or color image to a binary image. Having the binary version of the image mainly helps our system to re-define the text line polygons that are automatically extracted (see Sect. 3.2) in order to exclude non-text areas or to fully include text areas that lie in the polygon limits.

The binarization method used in the proposed system is fully described in [8] and consists of five distinct steps: a preprocessing procedure using a low-pass Wiener filter, a rough estimation of foreground regions using Niblack's approach, a background surface calculation by interpolating neighboring background intensities, a thresholding by combining the calculated background surface with the original image and, finally,

a postprocessing step that improves the quality of text regions and preserves stroke connectivity. An example of the binarization pre-processing step is demonstrated in Fig. 2.

3.2 Text Line Segmentation

Image pre-processing is followed by the step of text line segmentation, which is a necessary procedure in order to obtain precise and accurate recognition results. Automatic text line segmentation was carried out by the use of a variation of the well-known YOLOv5 [9] Deep Neural Network model (YOLOv5-OBB[1]). In order to automatically edit detection results acquired from YOLOv5-OBB and to efficiently apply OCR, the



(a)

(b)

Fig. 2. Example of the binarization pre-processing method. (a) original image (b) resulting binary image.

[1] https://github.com/hukaixuan19970627/yolov5_obb

detected polygons are sorted using Density-based spatial clustering (DBSCAN) in order to preserve the correct reading order of the text lines.

At a first step, the automatic text line detection procedure results to a set of polygons that surround each text line of the document, as shown in Fig. 3. Then, the user can correct the segmentation results. The system provides the user with the following functions:

- To correct a polygon by moving the desired points in the right position.
- To add a new polygon before or after a polygon.
- To delete a polygon.
- To connect a polygon with another polygon.

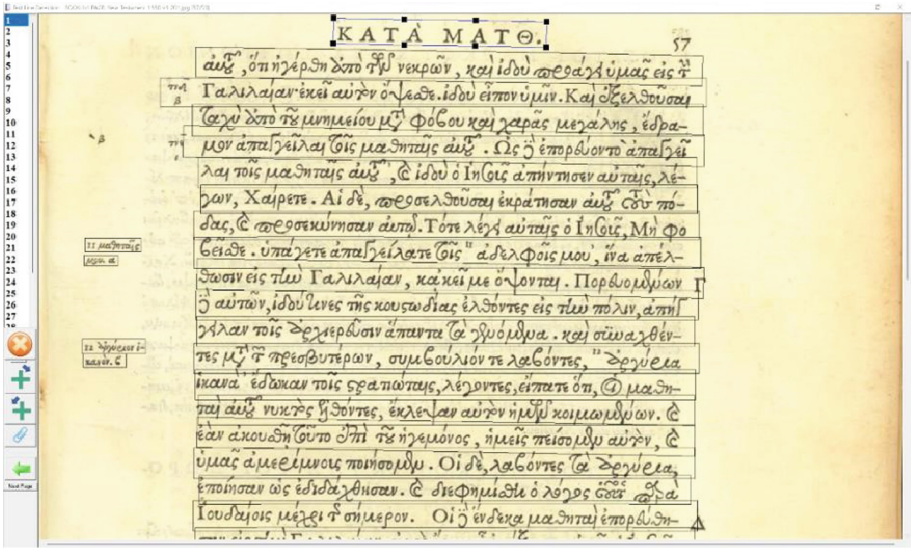


Fig. 3. The result of the text line segmentation shown in the proposed system.

In Fig. 4(a) one can see the result of the text line segmentation of the system, while in Fig. 4(b) the corrected polygon by the user. In Fig. 5 we present another example of the functionality of the system where the user can connect two parts of the same text line and create a polygon for the whole line.

When the procedure of text line detection and correction is completed, the system creates text line images with the image parts inside each polygon. These images will be then used as input for the text line recognition module.

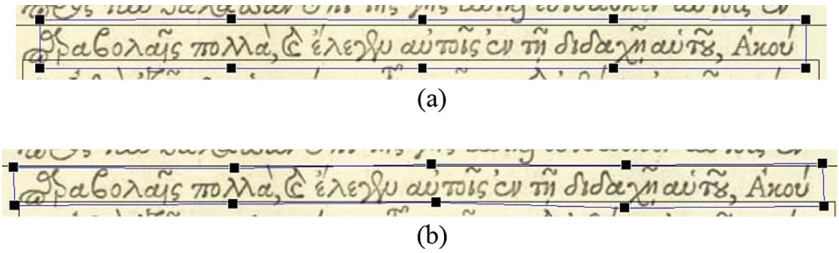


Fig. 4. An example of the text line detection (a) and the correction of the line polygon by the user (b).

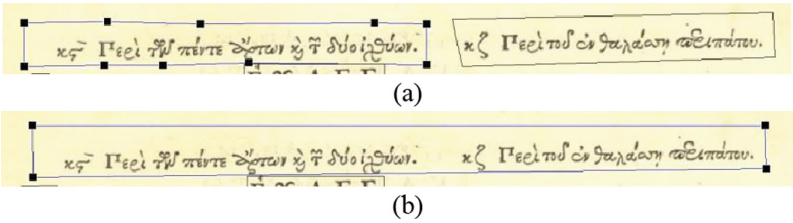


Fig. 5. An example of two polygons (a) that should be connected in order to form one text line (b).

3.3 Text Line Recognition

Having completed the text line segmentation, the user can move to the next step which is the text line recognition. This is first done automatically by the open-source, TensorFlow-based Calamari-OCR engine [1] that uses advanced deep neural network. As it is explained in Sect. 4, the best configuration for the recognition engine is selected after experimentation with a new database for old Greek documents.

At a next step, the user can correct the predicted OCR lines (see Fig. 6). A part of the image is presented with the text line of the prediction enclosed in a red box, thus placing the text line into context in relation to the surrounding text. What follows is a detailed image of the text line, which helps the user to focus on the correction. The predicted text appears in a box while a virtual keyboard appears below in order to help the user to make corrections. The user is able to navigate to the next or previous lines using the arrows.

The user checks the text thoroughly for errors and can perform corrections with the following ways. Using the cursor, he/she can select the erroneous character and either type the correct one, or he/she can choose to use the virtual keyboard below, containing a comprehensive list of Greek polytonic characters (see Fig. 7).

At a next phase, the corrected text together with the corresponding corrected text line polygon are used for network re-training.

Fig. 6. The user interface provided for text line recognition correction.

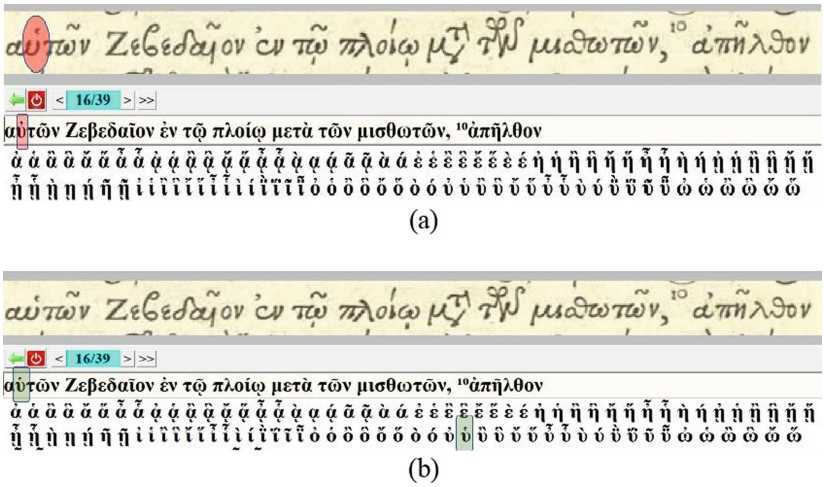


Fig. 7. Correction of recognition results: (a) the user spots the error, (b) the user corrects the error by selecting the correct character using the virtual keyboard.

4 Experimental Results

In order to test the proposed system, we introduce and provide publicly a new dataset of old Greek Documents [2] (see Fig. 1). This dataset consists of 57 pages containing the complete Gospel of Matthew, from the third edition of the Greek New Testament published in 1550 by Robert Estienne (1503–1559). Estienne, also known in his activities with the name Stephanus, was appointed "Royal Typographer" during the rule of King

of France François I (1494–1547). As a typographer and scholar, he published a number of classical texts including Greek and Latin translations of the Bible. His first edition of the Greek New Testament was in 1546 and the text was based on that printed by Erasmus in Basel in 1516. For his most significant edition, known as *Editio Regia* or the "Royal Edition" printed in 1550, Estienne used fifteen additional Byzantine manuscripts and presented for the first time a textual apparatus listing the variant readings of the different manuscripts he examined. The edition printed in large folio size using the *grecs du roi* typeface. This typeface, which attempts to imitate the Greek handwriting of this period, includes a large number of ligatures and abbreviations. Produced by Claude Garamont on the basis of the Greek minuscule style of the calligrapher Angelos Vergikios (1505–1569) from Crete, who was active copying Greek manuscripts in Venice and France, it became the most widely used Greek typeset for European printers. The dataset consists of 2045 text lines, 1431 used for training, 204 for validation and 410 for test. The text lines were produced automatically by our system based on [9] and then corrected by a user as it is described in Sect. 3.2. Also, the corresponding transcription was corrected following the procedure described in Sect. 3.3.

In order to test the recognition accuracy of the proposed system, we evaluated the embedded recognition engine which is the open-source Calamari-OCR engine [1] using a wide variety of configurations. In Table 1, we present selected results that show high performance and are associated with the predefined network architecture and the application or not of data augmentation during training. The evaluation presented in Table 1 uses as metrics the Character Error Rate (CER) and Word Error Rate (WER). Concerning the network architecture, we present results for:

def: The default Calamari with one BiLSTM layer and.

htr +: An adaptation of the standard network structure of the Transkribus platform [7].

def is the default Calamari network follows a *CONV1- > MAXPOOL- > CONV2 > MAXPOOL- > BiLSTM* scheme, where: *CONV1* is Convolutional Layer with 40 filters, stride 1 and 3x3 receptive field, followed by *ReLU* activation. *CONV2* is similar to *CONV1* but with 60 filters. *MAXPOOL* is a 2x2[2] max pooling layer and *BiLSTM* is a Bidirectional Long Short-term Memory layer with 200 hidden nodes. After the *BiLSTM* layer, Dropout is applied with a skip ratio of 0.5.

The *htr +* Calamari network follows a *CONV1- > CONV2- > MAXPOOL1- > CONV3 > MAXPOOL2- > BiLSTM1- > BiLSTM2- > BiLSTM3* scheme, where *CONV1* is a Convolutional Layer with 8 filters, stride 2x4 and 2x4 receptive field, followed by leaky *ReLU* activation. *CONV2* is a Convolutional Layer with 32 filters, stride 2x4 and 1x1 receptive field, followed by leaky *ReLU* activation. *MAXPOOL1* is a 2x4 max pooling layer with stride 2x4. *CONV3* is a Convolutional Layer with 64 filters, stride 1x1 and 3x3 receptive field, followed by leaky *ReLU* activation. *MAXPOOL2* is a 2x1 max pooling layer with stride 2x1. *BiLSTM1*, *BiLSTM2* and *BiLSTM3* are Bidirectional Long Shortterm Memory layers with 256 hidden nodes respectively. After each BiLSTM layer, Dropout is applied with a skip ratio of 0.5. For both architectures, CTC Loss is calculated as a scoring function.

[2] When notation AxB is used, A is for the horizontal axis of a layer (x-width) and B for the vertical axis (y-height)

Calamari also includes network architecture *deep3* which is not included in our results because of lower performance. Data augmentation includes padding, distortions, blobs, and multiscale noise.

Table 1. Experimental results (CER% / WER%) on the new dataset of old Greek Documents [2].

Network architecture:	*htr +*	*def*
Augmentation:		
NO	5.12 / 23.61	2.08 / 11.82
YES	3.66 / 18.06	**1.45 / 8.53**

As it can be observed in Table 1, the *def* architecture outperforms the *htr +* architecture while data augmentation improves the results significantly. The best results correspond to CER of 1.45% and WER of 8.53% which are acceptable and promising having in mind the small number of images included in the training set.

In order to compare the embedded OCR engine using best configuration (*def* architecture + augmentation in training) with other recognition methods applied on old Greek Documents, we also trained and tested on the datasets presented in [4]. Table 2 presents the results for these 4 datasets compared to the approach of [4] using best settings (Deslanting) as well as to the approaches of de Sousa Neto et al. [10] and Puigcerver [11]. As is can be observed in Table 2, the embedded OCR engine of the proposed system is comparable or outperforms existing approaches taking into account the CER performance.

Table 2. Comparative experimental results (CER% / WER%) on the datasets presented in [4].

Dataset	Tsochatzidis et al. [4]	de Sousa Neto et al. [10]	Puigcerver [11]	OCR Engine embedded in the proposed system
χφ53	**6.77 / 30.09**	7.85 / 34.63	10.45 / 30.20	8.04 / 35.64
χφ79	6.51 / 28.51	7.75 / 33.13	10.33 / 28.55	**5.14 / 28.25**
χφ114	7.71 / 34.30	8.03 / 36.72	10.19 / 34.58	**7.01 / 41.79**
Eparchos	**4.53 / 20.03**	4.95 / 21.91	5.18 / 22.21	32.48

5 Conclusions

In this work, we present a system for processing and recognition of Greek Byzantine and Post-Byzantine Documents. This includes modules for (i) image pre-processing for binarization, (ii) text line segmentation that is first done automatically by employing a

variation of the well-known YOLOv5 [9] Deep Neural Network model (YOLOv5OBB) and then corrected manually using a user-friendly interface and (iii) text line recognition that is provided by an advanced deep neural network using the open-source Calamari engine [1]. The correction of the OCR is done in an efficient way and by using a virtual keyboard. Moreover, we introduce and provide publicly a new dataset of old Greek Documents [2] that includes text line images and the corresponding transcription. This dataset helps us (i) to find the best configuration of the recognition network and (ii) to assess the accuracy the embedded OCR engine. Therefore, it proved to be acceptable and promising for the case of old Greek Documents since it resulted to a character error rate less than 1.5% by using a relatively small set of images for training. Promising results were also achieved when comparing the embedded recognition engine with other recognition methods already proposed for the recognition of old Greek Documents. In all cases, the proposed system is comparable or outperforms existing approaches taking into account the CER performance. Future work includes testing of other deep neural network architectures for the task of text line segmentation and recognition and also the creation of a larger dataset that can be used for training.

Acknowledgments. This research has been partially co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call "RESEARCH-CREATE-INNOVATE", project Culdile (Cultural Dimensions of Deep Learning, project code: T1EΔK-03785) and the Operational Program Attica 2014–2020, under the call "RESEARCH AND INNOVATION PARTNERSHIPS IN THE REGION OF ATTICA", project reBook (Digital platform for re-publishing Historical Greek Books, project code: ATTP4–0331172).

References

1. Wick, C., Reul, C., Puppe, F.: Calamari - A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition. Digit. Humanit. Q. **14**(1) (2020)
2. https://zenodo.org/record/7876098#.ZEvjNtJBxNh
3. Ntzios, K., Gatos, B., Pratikakis, I., Konidaris, T., Perantonis, S.J.: An old Greek handwritten OCR system based on an efficient segmentation-free approach. Int. J. Doc. Anal. Recogn. (IJDAR) **9**(2–4), 179–192 (2007). special issue on historical documents
4. Gatos, B., Ntzios, K., Pratikakis, I., Petridis, S., Konidaris, T., Perantonis, S.J.: An efficient segmentation-free approach to assist old Greek handwritten manuscript OCR. Pattern Anal. Appl. (PAA) **8**(4), 305–320 (2006)
5. Tsochatzidis, L., Symeonidis, S., Papazoglou, A., Pratikakis, I.: HTR for Greek historical handwritten documents. J Imaging **7**, 260 (2021)
6. Platanou, P., Pavlopoulos, J., Papaioannou, G.:. Handwritten paleographic greek text recognition: a century-based approach. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 6585–6589. European Language Resources Association, Marseille (2022)
7. https://readcoop.eu/transkribus/
8. Gatos, B., Pratikakis, I., Perantonis, S.J.: Adaptive degraded document image binarization. Pattern Recogn. **39**, 317–327 (2006)
9. https://github.com/ultralytics/yolov5

10. de Sousa Neto, A.F., Bezerra, B.L.D., Toselli, A.H., Lima, E.B.: HTR-Flor: a deep learning system for offline handwritten text recognition. In: Proceedings of the 33rd SIBGRAPI Conference on Graphics, Patterns and Images, pp. 54–61. Recife/Porto de Galinhas (2020)
11. Puigcerver, J.: Are multidimensional recurrent layers really necessary for handwritten text recognition? In: Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition, pp. 67–72. Kyoto (2017)