OCR Workflow: Facing Printed Texts of Ancient, Medieval and Modern Greek Literature

Christina Tzogka[1⋆], Fotini Koidaki[2*], Stavros Doropoulos[1*], Ioannis Papastergiou[1*], Efthymios Agrafiotis[1*], Katerina Tiktopoulou[2*], and Stavros Vologiannidis[3*]

[1] DataScouting, 30 Vakchou Street, 54629 Thessaloniki, Greece
{ctzogka, doro, ipapaste, agrafiotis}@datascouting.com
[2] School of Philology, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
{koidaki, atiktopo}@lit.auth.gr
[3] Department of Computer, Informatics and Telecommunications Engineering, International Hellenic University, Terma Magnisias, 62124 Serres, Greece
svol@ihu.gr

Abstract. Book digitization is being increasingly enhanced, as it facilitates not only the dissemination and preservation of cultural heritage but also the analysis of large amounts of textual data as well as the extraction and discovery of knowledge in a faster, dynamic and interactive way. Quite often, OCR, as the core technology of book digitization, has to address major difficulties related to the condition of the primary source or to scanning issues. The main contribution of this paper is to provide an extensive study on Tesseract, an open-source OCR system, including image pre-processing and text post-processing methods, that overcome a variety of image handling problems. Additionally, a re-trained Greek language model, based on individual fonts training plus pairs of image-text training, is being provided. Finally, this paper proposes a pipeline of methods, including text line detection, that result in enhanced accuracy for Greek Literature documents, even when they consist of distorted pages, due to scanning issues or damaged physical material.

Keywords: OCR · Scanned Document · Deep Learning · Training · Text Line

1 Introduction

New technologies have introduced new ways of researching, reading and preserving written texts. The researcher can now search for information and extract knowledge through a large amount of textual documents, while the student and reader can acquire textual content quickly and effortlessly by creating his personalized and editable digital library [1]. At the same time, the role of the library is being modernized. The latter is now transformed into an intelligent entity capable of managing, researching, classifying, analyzing information and dealing even with the most complex bibliographic needs of researchers.

In order to be able to take advantage of the opportunities offered by the digitization technologies in researching, reading and preserving Greek Literature, a necessary condition is the conversion of the scanned literary documents into computer editable text [2]. This paper presents the scientific findings as well as the

⋆ The authors sincerely acknowledge the valuable contribution of the digital humanities research team, Maria Georgoula, Valando Landrou, Markia Liapi and Marianna Sylivrili, who as project partners supported post-processing phase, by providing the ground-truth data and valuable feedback on the OCR results.

workflow for the development of a comprehensive OCR system trained to analyse scans of books and recognise Greek polytonic characters of modern and ancient texts.

Moreover, it utilizes and evolves parts of a remarkable research conducted in the context of the "*Exploitation of Cultural Assets with computer-assisted Recognition, Labeling and meta-data Enrichment (ECARLE)*" project. The ECARLE project concerns the development of an integrated Software as a Service (SaaS) that can take scans of printed documents as input and export them as editable electronic texts, enriched with semantic metadata about the structure of the document, its publication, as well as semantic elements of cultural interest (i.e. person names, place names, dates, book titles etc.). The scientific interests of ECARLE project focus on the publications of the 19th century, since this is the century during which typography emerged and flourished in Greece producing printed artefacts of major cultural importance. To this end, the structure of the paper reflects the research progression which is organized in the following five main sections: a. OCR Challenges, b. State Of The Art Review, c. Tesseract Training, d. Proposed Method, e. Experimental Evaluation and f. Future Challenges.

The main contribution of this paper is i. to provide an extensive study on the OCR system, including image pre-processing techniques that address a variety of image quality issues and text post-processing, ii. to provide a fine-tuned Greek language model based on individual font training as well as pairs of image-text training, iii. to propose a pipeline of methods that enhance the accuracy of the digitisation of Greek literary documents, even when they include worn (e.g distorted/skew) pages, due to scanning issues or damaged physical material. The evaluation of the already available and the re-trained OCR models was performed by copy-editors, specialising in the subject–matter in order to create ground-truth data which will be used for evaluation and further optimization.

2 OCR Challenges

This section makes a detailed description of the kinds of challenges that the proposed OCR system succeeds to address. Before inserting an image into the OCR system, it is necessary to pre-process this image, since it may contain imperfections [3], which complicate the process and reduce the quality of the generated text. Furthermore, imperfections make it difficult to identify characters due to their deviation and variability.

- ◇ Deviation: The letters in an image differ significantly from the structure expected by the OCR system. This may be due to physical deterioration, or printing errors, or even scanning imperfections.
- ◇ Variability: Identical characters that are represented in one or more images, show variability, i.e. the same letter is represented by more than one glyph which substitute each other as stylistic alternates in the same font.

The reasons for these defects could be categorized as physical material problems and problems during the scanning.

- Problems of physical material: The physical material, depending on the storage way, usage and age, may have deteriorated significantly compared to its initial condition. Examples of such alterations are the fading of the letters, the thinning of the paper in which the text on back of the sheet from the front is visible or even stains, marks and watermarks added by the readers or the owner etc.
- Scanning issues: The physical material, depending on the procedure and equipment used during the scanning, may not be accurately represented in its digital

facsimile. This may be due to faults in the placement of the object in the scanner, lenses with limited sharpness, incorrect or in-homogeneous illumination and refractions. The lesion may include warping[4], skewing [5] as well as the appearance of shadows, color and contrast gradients.
However, in addition to the scanning failures mentioned above, quite often there are some expected scanning challenges. For instance, while a book is flipped during scanning, the degree of distortion changes due to the way a book is located on the mounting surface, based on its geometry.

These problems are both local and universal. In fact, their intensity may differ from page to page. For instance, pages in the center of the book, that do not have much exposure to environmental conditions and fluctuations, may have variations in wear rate. Similar variations in wear rate are also expected to exist within the page itself.

3 State Of The Art Review

As mentioned above, text recognition is undoubtedly a crucial and demanding process in order to achieve great performance. The accuracy [6, 7] of the digital text produced depends on the extent of the physical deterioration suffered by the source materials as well as the condition of the scanned images. For this reason the texts that will be used during experiments and evaluation are carefully selected on the basis of combined criteria concerning external and internal factors. According to the relevant international literature, the workflow for the conversion of digitized material into computer-editable digital text is analyzed in the following stages: i. image pre-processing, ii. extract properties and character recognition, iii. post-processing. The state-of-the-art OCR engines [8] that have been used in related works are the following:

1. ABBYY Finereader (www.abbyy.com) OCR engine clearly defines the state of the art for layout analysis and OCR, supporting about 200 recognition languages.
2. OCRopus [9] is an open-source engine (based on LSTM neural networks) with significant recognition capabilities compared to glyph-based approaches. Additionally, this method provides the ability to train new models by just providing input image in pair with their ground-truth text, on line level.
3. Tesseract [10] is an open-source OCR engine (based on LSTM neural networks) which provides a wide variety of trained mixed models. Moreover, Tesseract supports language modelling and training either on specific fonts or on image samples in pair with their ground-truth text.

In order to achieve best possible performance, the open-source OCR software, Tesseract, was selected. Our decision was mainly based on its popularity in the open-source state-of-the-art OCR engines literature as well as on its various capabilities on further training and fine-tuning that result in great performance.

3.1 OCR System's Features

An OCR system includes a set of parameters, machine learning techniques and training profiles so that they can identify as diverse items as possible. Taking into account the most usual user requirements, the most common parameters refer to the system's ability to recognize a variety of languages, fonts, document types and text alignments so that it can meet the most likely needs of its users. Although this generalisation may satisfy the objectives of the most use cases of such software, it often creates problems:

1. It is infeasible to include all the possible use-cases in the scope of the software and particularly the specialized and not frequently encountered. For example, obsolete typography and archaic writing systems have been disregarded by the OCR models, which results in undefined user stories and unsatisfied user needs. Frequently, the existence of these scenarios is unknown to the system designer due to their rarity, as a result of which they are not even included in the strategy for selecting the optimal parameters and training forms of the systems.
2. Generalization can reduce system performance since it's goal is to maximize performance in as many usage scenarios as possible rather than in specific scenarios. For example, an OCR system that has been trained to recognize a large amount of fonts is quite likely to make mistakes by confusing different characters because of similar typographic features. Knowing that it is impossible to cover all the software's usage scenarios, the designers of these OCR systems (e.g. OCRopus, Tesseract) allow their users to customize them according to their needs. The configuration of OCR systems is mainly divided into three categories:
   (a) **Change OCR settings:** This category includes various variables that regulate the operation of the OCR system. These variables may be the set of recognizable languages, the architecture of the recognition model, the page segmentation etc.
   (b) **Fine-tune training models:** Each OCR system uses at its core a set of machine learning models [11] (e.g CNN, LSTM, etc.). The parameters of the models (e.g. number of neurons, layers, activation functions) have been selected by the designers based on the general usage scenarios. In OCR systems, such as Tesseract, the user is given the opportunity to fine-tune these models, in order to define a set of parameters that could be more efficient than the predefined parameters for the specific task.
   (c) **Train the system:** Many language models provided by the OCR engines are font-created and they have not been trained in real texts. Training procedure requires ground-truth data in order the model to learn to recognise not only characters, but also words and strings. The texts on which the pretrained models have been trained are generic, i.e they do not include all possible letter combinations of a language. Most OCR systems (e.g.OCRopus, Tesseract) allow their users to choose the languages, characters and corresponding texts/images that will be used to train the model.

4 Tesseract Training

Tesseract is an open-source OCR software that supports various operating systems (Windows, Linux, etc.) and includes a large number of pre-trained languages.

4.1 Default Model

Tesseract provides two packages of polytonic Greek, that divide the Greek polytonic writing system between the ancient and the modern one:

– Ancient Greek (GRC) is based on a vocabulary derived from texts before 1453, which is the year of conquest of Constantinople by the Ottoman troops.
– Modern Greek (ELL) includes texts after 1453 with a strong emphasis on modern Greek.

Tesseract gives three different models (i.e tessdata) for each language, one normal, one optimized for accuracy (best) and one optimized for speed (fast). The best model was used as a basis for optimizing the results for both Greek models.

Evaluation: Both OCR models for the polytonic Greek were evaluated individually, without any image pre-processing in the evaluation data. In both cases the system recognised the punctuation marks with great success but at the same time there was observed an extremely large failure in unknown fonts, since the Tesseract Greek models have been trained in a number of the most common fonts (e.g. Arial, Dejavu, Coutier-New etc.). Furthermore, there were phenomena of text fragmentation, line separation into two or more lines, line creation, etc.

The evaluation (by copy-editors see section 4.3) revealed that the quality of the output of the default Tesseract model was extremely low (see Version-1 in Figures below 3 and 4), as the text has been altered and disfigured in such an extent that it was difficult to correct or even read. In fact, it was quite difficult for the human eye even to match the digital text with the text of the scanned page. Consequently, we noticed that it was necessary to train the system with machine-learning-assisted correction algorithms.

4.2 Training Default Model (grc-ecarle)

The corpus that we created for the training of the system includes content collected from various websites that contain texts written in the Greek archaistic form, also known as katharevousa (i.e polytoniko.org) as well as from monasteries' webpages which publish texts in katharevousa. Finally, a corpus of 5000 lines was built whose content was mainly Ancient Greek texts (especially historical texts of Thucydides). The training focused on "GFS Didot Classic" font, since it was observed to prevail all over the corpus. The training setup was based on Tessdoc/TrainingTesseract while the total number of epochs equals 15,000. The corpus was split into training (90%) and evaluation (10%). In order to identify the font characters for which the system had not been trained, we decided to train both the Greek models included by Tesseract (grc and ell).

The exported trained language model was further trained based on pairs of ground-truth data and image samples (i.e approximately 20 scanned pages) from the available documents. The total number of epochs equals 5,000. The final trained model will, hereafter, referred to as "grc-ecarle" and will be used throughout the proposed method that is presented below.

Evaluation: The evaluation results of the new trained model, were significantly better (see Version-2 in Figures 3 and 4), since the training focused on training the system on a single font, as it was described above. Particularly, the evaluation was performed on image samples, that have not undergone any pre-processing, showed a tendency of the system to be able to easily recognize the morphology of the characters as well as the punctuation. This is mainly due to the fact that Tesseract uses LSTM neural networks that have the ability to learn large sequences of letters. In this way, issues concerning punctuation, due to wear, are being fixed. Therefore, LSTMs is how we handled and finally cured issues concerning the misrecognition of:

- capital letters and punctuation (e.g comma, colon)
- numbers and common symbols (; "" «»+ = & #)
- glyphs that are above or below the line height, such as $\beta, \theta, \rho, \gamma$

Grc-ecarle trained model's weakness: The trained model ("grc-ecarle") has shown excellent results in addressing some crucial problems. However, the evaluation of the very first outputs revealed the weaknesses of the system and confirmed the need to create ground-truth data for evaluation and for further accuracy improvement. The weaknesses of the system concerned the identification of some characters, that could be further improved as well as the following transcript failures:

- horizontal and vertical line segregations
- addition of both noisy lines, which contain only symbols and numbers, and empty lines (no OCRed text at all)
- inverted content of consecutive lines

It is worth mentioning that the image processing technique (Leptonica) used by Tesseract could not adequately address the image distortion and skew problems. Therefore, it became obvious that further image pre-processing was required to improve the results.

4.3 Copy-Editors Post-Process

The post-processing phase is the last step of the OCR training workflow and involves the OCR error correction. Instead of an automated post-processing system, human assisted post-processing was preferred as more efficient for our dataset, since the polytonic accent system requires expert editors. This kind of process aims to create validation data, which are essential in order to improve the system and to evaluate its performance.

5 Proposed Method

In order to address the grc-ecarle train model's weakness and further improve the system, we constructed a pipeline of processes (Figure 1).

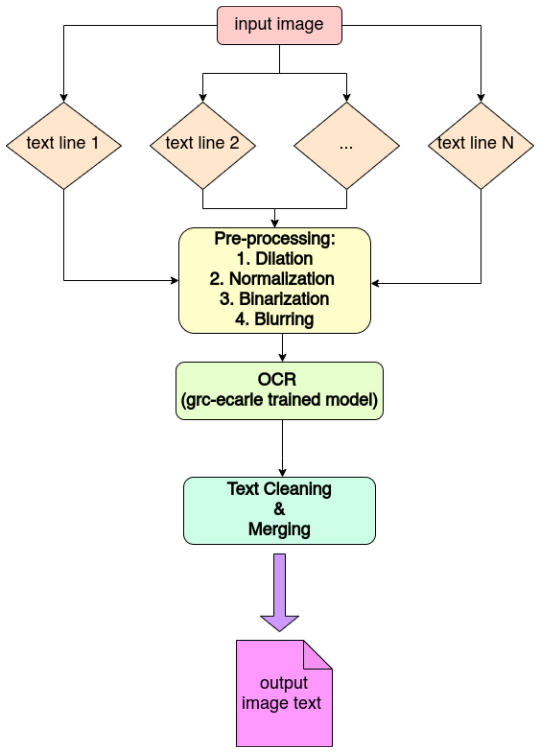
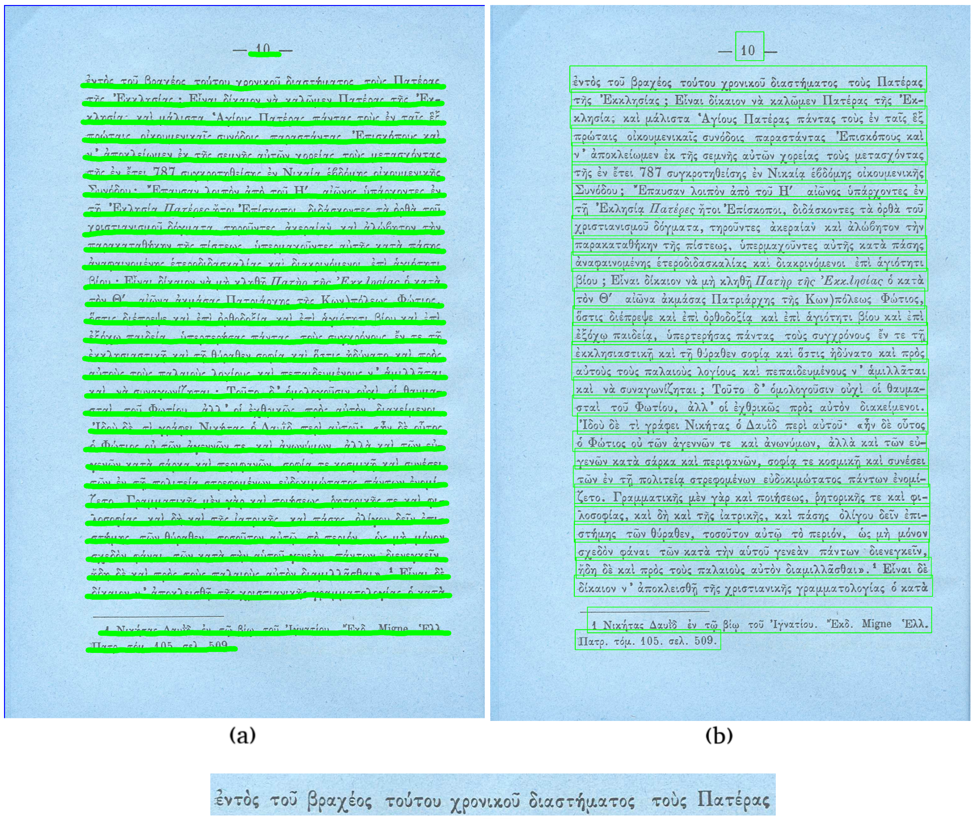


Fig. 1: Proposed Method's Pipeline.

5.1 Text-line Detection and Cropping

Initially, we attempted to split every page into lines, so that a possible distortion couldn't affect the sequence of the lines. The first step of the OCR pipeline corresponds to text-line detection and cropping, in which the scanned images are transformed into text lines and the content is represented in lines, instead of pages. The open-source software dhSegment [12] was used for the purposes of text line detection [13] in association with the baseline detection model.

The open-source software dhSegment locates the area occupied by the respective text line. More precisely, it applies an underlining technique, without providing the exact bounding box of a line (Figure 2-a). This gap is filled by an algorithm designed to generate bounding box coordinates (Figure 2-b) with high precision, so that the corresponding text line could be cropped (Figure 2-c). In fact, this algorithm calculates the line width by estimating the distance (step 2.(b) - algorithm 1) at which the line ends and subtracting it from the current line height.



— 10 —

ἐντὸς τοῦ βραχέος τούτου χρονικοῦ διαστήματος τοὺς Πατέρας
τῆς Ἐκκλησίας ; Εἶναι δίκαιον νὰ καλῶμεν Πατέρας τῆς Ἐκ-
κλησίας καὶ μάλιστα Ἁγίους Πατέρας πάντας τοὺς ἐν ταῖς ἑξ
πρώταις οἰκουμενικαῖς συνόδοις παραστάντας Ἐπισκόπους καὶ
ν' ἀποκλείωμεν ἐκ τῆς σεμνῆς αὐτῶν χορείας τοὺς μετασχόντας
τῆς ἐν ἔτει 787 συγκροτηθείσης ἐν Νικαίᾳ ἑβδόμης οἰκουμενικῆς
Συνόδου ; Ἔπαυσαν λοιπὸν ἀπὸ τοῦ Η' αἰῶνος ὑπάρχοντες ἐν
τῇ Ἐκλησίᾳ *Πατέρες* ἤτοι Ἐπίσκοποι, διδάσκοντες τὰ ὀρθὰ τοῦ
χριστιανισμοῦ δόγματα, τηροῦντες ἀκεραίαν καὶ ἀλώβητον τὴν
παρακαταθήκην τῆς πίστεως, ὑπερμαχοῦντες αὐτῆς κατὰ πάσης
ἀναφαινομένης ἑτεροδιδασκαλίας καὶ διακρινόμενοι ἐπὶ ἁγιότητι
βίου ; Εἶναι δίκαιον νὰ μὴ κληθῇ *Πατὴρ τῆς Ἐκκλησίας* ὁ κατὰ
τὸν Θ' αἰῶνα ἀκμάσας Πατριάρχης τῆς Κων)πόλεως Φώτιος,
ὅστις διέπρεψε καὶ ἐπὶ ὀρθοδοξίᾳ καὶ ἐπὶ ἁγιότητι βίου καὶ ἐπὶ
ἐξόχῳ παιδείᾳ, ὑπερτερήσας πάντας τοὺς συγχρόνους ἔν τε τῇ
ἐκκλησιαστικῇ καὶ τῇ θύραθεν σοφίᾳ καὶ ὅστις ἠδύνατο καὶ πρὸς
αὐτοὺς τοὺς παλαιοὺς λογίους καὶ πεπαιδευμένους ν' ἁμιλλᾶται
καὶ νὰ συναγωνίζηται ; Τοῦτο δ' ὁμολογοῦσιν οὐχὶ οἱ θαυμα-
σταὶ τοῦ Φωτίου, ἀλλ' οἱ ἐχθρικῶς πρὸς αὐτὸν διακείμενοι.
Ἰδοὺ δὲ τί γράφει Νικήτας ὁ Δαυῒδ περὶ αὐτοῦ· «ἦν δὲ οὗτος
ὁ Φώτιος οὐ τῶν ἀγεννῶν τε καὶ ἀνωνύμων, ἀλλὰ καὶ τῶν εὐ-
γενῶν κατὰ σάρκα καὶ περιφανῶν, σοφίᾳ τε κοσμικῇ καὶ συνέσει
τῶν ἐν τῇ πολιτείᾳ στρεφομένων εὐδοκιμώτατος πάντων ἐνομί-
ζετο. Γραμματικῆς μὲν γὰρ καὶ ποιήσεως, ῥητορικῆς τε καὶ φι-
λοσοφίας, καὶ δὴ καὶ τῆς ἰατρικῆς, καὶ πάσης ὀλίγου δεῖν ἐπι-
στήμης τῶν θύραθεν, τοσοῦτον αὐτῷ τὸ περιόν, ὡς μὴ μόνον
σχεδὸν φάναι τῶν κατὰ τὴν αὐτοῦ γενεὰν πάντων διενεγκεῖν,
ἤδη δὲ καὶ πρὸς τοὺς παλαιοὺς αὐτὸν διαμιλλᾶσθαι». [1] Εἶναι δὲ
δίκαιον ν' ἀποκλεισθῇ τῆς χριστιανικῆς γραμματολογίας ὁ κατὰ

[1] Νικήτας Δαυῒδ ἐν τῷ βίῳ τοῦ Ἰγνατίου. Ἔκδ. Migne Ἑλλ.
Πατρ. τόμ. 105. σελ. 509.

ἐντὸς τοῦ βραχέος τούτου χρονικοῦ διαστήματος τοὺς Πατέρας

Fig. 2: This is a random sample where (a) shows text line detection (via dhSegment), (b) shows the bounding box generation per text line (via algorithm 1) and (c) shows the algorithm's output which corresponds to the first text line.

A scanning problem that the proposed algorithm below successfully addresses is the division of a line into 2 or even 3 individual lines. The algorithm detects cases like this, using the predefined threshold, and restores the line by attaching its fragments into a single line (step 3 - Algorithm 1). Additionally, the estimated line width is being normalized in order to handle lines with significantly smaller/higher line width compared to the average line width of the current page (step 4 - Algorithm

1). Finally, it removes lines that appeared extremely small length compared to a threshold value that was defined taking into account the average line length (step 6 - Algorithm 1). The algorithm finally estimates the actual coordinates of the bounding box (per line) in order to generate the cropped text line that will feed the OCR system (steps 7 and 8 - Algorithm 1).

Consequently, after applying dhSegment an additional set of sub-processes is being implemented. These sub-processes result in the proposed algorithm, that consist of the following steps:

Algorithm 1: Text line generation

```
1. extracting full list of initial bounding boxes $(x_{start}, y_{end}, x_{end}, y_{end})$ per line
   and sorting them by ascending $y_{end}$ (second item of initial bounding box).
   (a) $(x_{start}, x_{end})$ correspond to (left, right) points of the bounding box
   (b) $y_{start}$ is not estimated through dhSegment sub-processes, since the output
       delivered from its implementation corresponds to a thick line
   (c) $y_{end}$ corresponds to bottom point of the bounding box
2. contracting list with line width per text-line
   (a) we re-defined $y_{end}$ as the maximum of the two $y_{end}$ points per bounding
       box $(x_{start}, y_{end}, x_{end}, y_{end})$
   (b) we estimated line width $(y_{diff})$ per line by subtracting the current $y_{end}$
       from the previous one
3. merging consecutive lines if their (width) difference is under a threshold value
   (minLineDiff)
4. filtering $y_{diff}$ by estimating mean and standard deviation in order to
   normalize lines that have extremely small or large width
5. defining parameters for $y_{start}$ estimation and text line cropping
   (a) minLineLength = minimum text line length applied to throw away lines
       that do not contain valid text (e.g. symbols generated by noise)
   (b) $y_{endExtend}$ = small region extended $y_{end}$ to avoid cutting lines that may
       have been scanned askew
   (c) minStartY = minimum value of $y_{start}$ applied when line is at the top of
       the page and a minimum starting height should be picked
   (d) $y_{startEntend}$ = small region extended $y_{start}$ to avoid cutting lines that may
       have been scanned askew
6. removing lines whose length is under the minLineLength threshold
7. estimating the actual bounding boxes via a loop:
   (a) $y_{end}$ is considered to be the maximum from the two $y_{end}$ points in initial
       box $(x_{start}, y_{end}, x_{end}, y_{end})$, plus $y_{endExtend}$ (extend box to down
       direction)
   (b) if line is at the top of the page (i.e $y_{end}$ - $y_{diff} < 0$) then
       $y_{start} = minStartY$
   (c) if line is not at the top of the page (i.e $y_{end} - y_{diff} > 0$) then
       $y_{start} = y_{end} - y_{diff} - y_{startEntend}$ (extend box to up direction)
8. extracting final text line bounding box and cropping it
```

5.2 Image Pre-processing

The next step after the text line detection and cropping, is applying a set of image pre-processing techniques [14] to each text line. The image pre-processing step aims to address issues related to the image clarity. There are many techniques that contribute cumulatively to the above goal, each of which tries to identify and eliminate a specific problem. As mentioned above, the image processing techniques (Leptonica) used by Tesseract internally could not fix all the problems within an image. Therefore, further image pre-processing of the inputs is required. The proposed method involves the following techniques which are applied hierarchically:

- Dilation [15] technique is used to remove shadow and noise around the letters. In the dilation process the letters are expanded until the smallest faded letters filled with color. Dilation is based on a rule where the state of every pixel in the output image is determined by the corresponding pixel and its neighbors in the input image.
- Image Normalization [16] process changes the range of pixel intensity values and makes the image more familiar or normal to the human eye. This technique is often used to increase contrast allowing this way the noise removal. In order to achieve this all the high or extremely low frequency contents are being removed from the image.
- Binarization process converts the image from colored or grayscale to black-and-white in order to eliminate shadows, images and dots that do not correspond to characters in the page, since these phenomena mislead the OCR recognition system. Optimal performance of an OCR system requires high contrast between the characters of an image and the rest of the background.
- Blurring is a process that removes pixel gradations that correspond to noise. This is being achieved by convolving the image with a low-pass filter kernel that removes high frequency content (e.g. noise, edges) from the image. As a result of this technique, the edges in the image are a bit blurred leading to higher OCR accuracy.

5.3 OCRed Text Cleaning and Merging

At this point, input images have been properly processed and are in the ideal condition for the OCR process. The next step is to start the OCR process, after defining all the necessary parameters. More precisely, since input images are aligned to text lines, it is necessary to use the appropriate page segmentation mode. By default Tesseract expects a text page when it segments an input image. Therefore, when the input corresponds to a small text region, a different segmentation mode should be tried. A complete list of supported modes is available in the official documentation. For proposed method's purposes we have chosen the mode that treats the image as a single text line.

After completing the above-mentioned image processing and page segmentation, the latest trained "grc-ecarle" language model is being utilized. In the new OCR output every scanned page is transformed in a continuous text line. Therefore, a technique that will attach the individual text lines into a single textual unit is being applied. Meanwhile, a text cleaning process aims to the best possible clarity by removing the following:

- ◇ Blank lines and lines with noise, i.e the percentage of symbols and numbers in line exceeds a threshold (e.g 50%).
- ◇ Lines with extremely long words, i.e sequences of characters longer than a cutoff (e.g tokenMaxLength = 21), which are rare in the Greek language and probably correspond to noise.

It is worth noting that all the above-mentioned threshold values are being set based on the language's characteristics. It's quite possible these values may not result in such an optimal OCR output for another language.

6 Experimental Evaluation

This section presents the evaluation results for each of the three models tested. These models correspond to: a.) Version-1 (Tesseract's default Ancient Greek model), b.) Version-2 (grc-ecarle trained model) and c.) Version-3 (proposed method).

Table 1 presents the main features of the 3 versions of OCR models that were implemented during the experiments. Specifically, the following are mentioned:

- the mode of the input sample, either as a text line or as a full page
- the implementation of image pre-processing techniques to the OCR input
- the specific trained language model used for generating OCR output
- the issues that were observed to degrade the accuracy of the final results

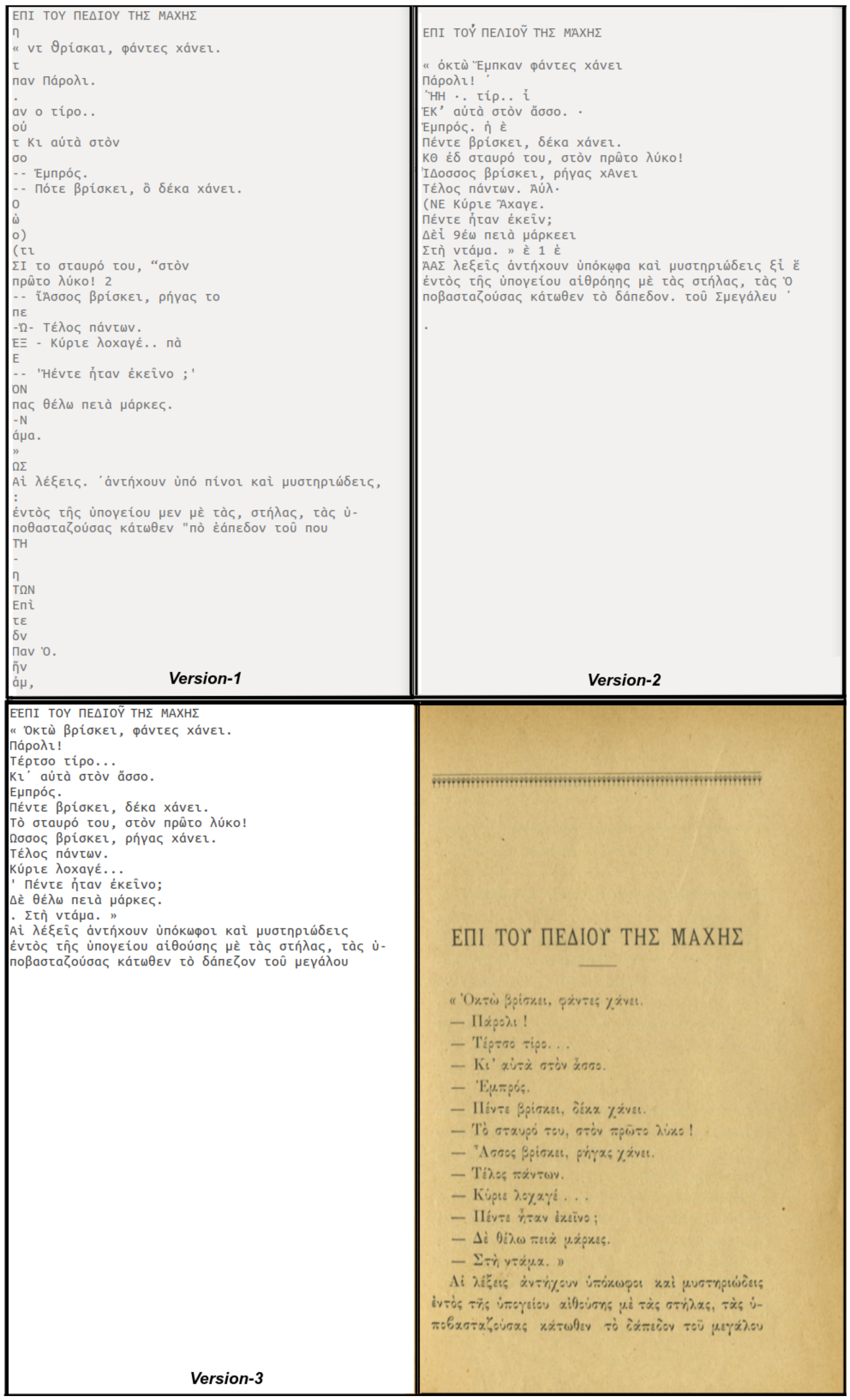
Version	**Input Mode**	**Image Pre-process**	**Trained Model**	**Issues**
Version-1	Full page	No	grc	- Extremely low character accuracy. - Line separation into more lines. - Inserting empty and "noisy" lines (i.e sequence of symbols and numbers).
Version-2	Full page	No	grc-ecarle	- Better character accuracy but still remain many cases of misspelled characters/numbers. - Less empty/"noisy" lines.
Version-3	Text line	Yes	grc-ecarle	- Enhancing character accuracy. - Fixing text fragmentation issues. - Optimal result.

Table 1: Model versions. Features and issues.

For each of the above models the performance was tested for a random page sample. Figures 3 and 4 show the progress made in OCR between the different versions of the implemented models, for two different samples. The first sample could be interpreted as normal (without distortion) by a human, but it turned out to involve several difficulties for the OCR system. The second sample is obviously distorted due to material/scanning problems. It is worth noting that in both Figure 3 and 4 we have interfered with Version-1 OCRed text. In fact, the latter text consists of too many lines, compared to the other Versions, while a significantly large percentage of them corresponds to empty lines. Therefore, we decided to remove a few empty lines in order to fit the example in figure's dimensions.

The superiority of the proposed method (Version-3) is visible to human eyes. Nevertheless, it is necessary to estimate the character accuracy of each OCRed text in Figures 3 and 4 and present it to a bar chart (Figure 5), in order to confirm that the proposed model outperforms the rest of implemented models. Character accuracy is being estimated after ground-truth text generation (post-processing phase), making use of the open-source tool, ocreval [17]. The latter bar chart confirms the ability of Version-3 to offer the optimal results, approaching to a large extent (98.56% and 97.11% for each image, respectively) the ground-truth text. Meanwhile, in Figure 3 there is a large accuracy gap between Version-1, which does not exceed 50%, and Version-3, which approaches 99%.

Finally, in order to give a comprehensive overview of the performance of each model, a evaluation set was constructed, consisting of 102 image samples. All image samples have passed the post-processing step and the corresponding ground-truth



```
ΕΠΙ ΤΟΥ ΠΕΔΙΟΥ ΤΗΣ ΜΑΧΗΣ
η
« ντ θρίσκαι, φάντες χάνει.
τ
παν Πάρολι.
.
αν ο τίρο..
ού
τ Κι αύτὰ στὸν
σο
-- Έμπρός.
-- Πότε βρίσκει, ὃ δέκα χάνει.
Ο
ὠ
ο)
(τι
ΣΙ το σταυρό του, “στὸν
πρῶτο λύκο! 2
-- ἴΆσσος βρίσκει, ρήγας το
πε
-Ὠ- Τέλος πάντων.
ΈΞ - Κύριε λοχαγέ.. πὰ
Ε
-- 'Ἠέντε ἦταν έκεῖνο ;'
ΌΝ
πας θέλω πειὰ μάρκες.
-Ν
άμα.
»
ΩΣ
Αἰ λέξεις. ΄άντήχουν ὑπό πίνοι καὶ μυστηριώδεις,
:
ἐντὸς τῆς ὑπογείου μεν μὲ τὰς, στήλας, τὰς ὑ-
ποθασταζούσας κάτωθεν "πὸ ἐάπεδον τοῦ που
ΤΗ
-
η
ΤΩΝ
Επὶ
τε
δν
Παν Ὀ.
ἥν
άμ,
```

```
ΕΠΙ ΤΟΫ ΠΕΛΙΟΫ ΤΗΣ ΜΆΧΗΣ

« όκτὼ Ἕμπκαν φάντες χάνει
Πάρολι! ΄
῍ΗΗ ·. τίρ.. ἰ
ΈΚ’ αὐτὰ στὸν ἄσσο. ·
Έμπρός. ἡ ὲ
Πέντε βρίσκει, δέκα χάνει.
ΚΘ έδ σταυρό του, στὸν πρῶτο λύκο!
῍ΙΔοσσος βρίσκει, ρήγας χΑνει
Τέλος πάντων. Άὐλ·
(ΝΕ Κύριε Ἄχαγε.
Πέντε ἦταν έκεῖν;
Δὲἶ 9έω πειὰ μάρκεει
Στὴ ντάμα. » ὲ 1 ἐ
ἈΑΣ λεξεῖς ἀντήχουν ὑπόκῳφα καὶ μυστηριώδεις ξἶ ἔ
ἐντὸς τῆς ὑπογείου αἰθρόησ μὲ τὰς στήλας, τὰς Ὀ
ποβασταζούσας κάτωθεν τὸ δάπεδον. τοῦ Σμεγάλευ ΄

.
```

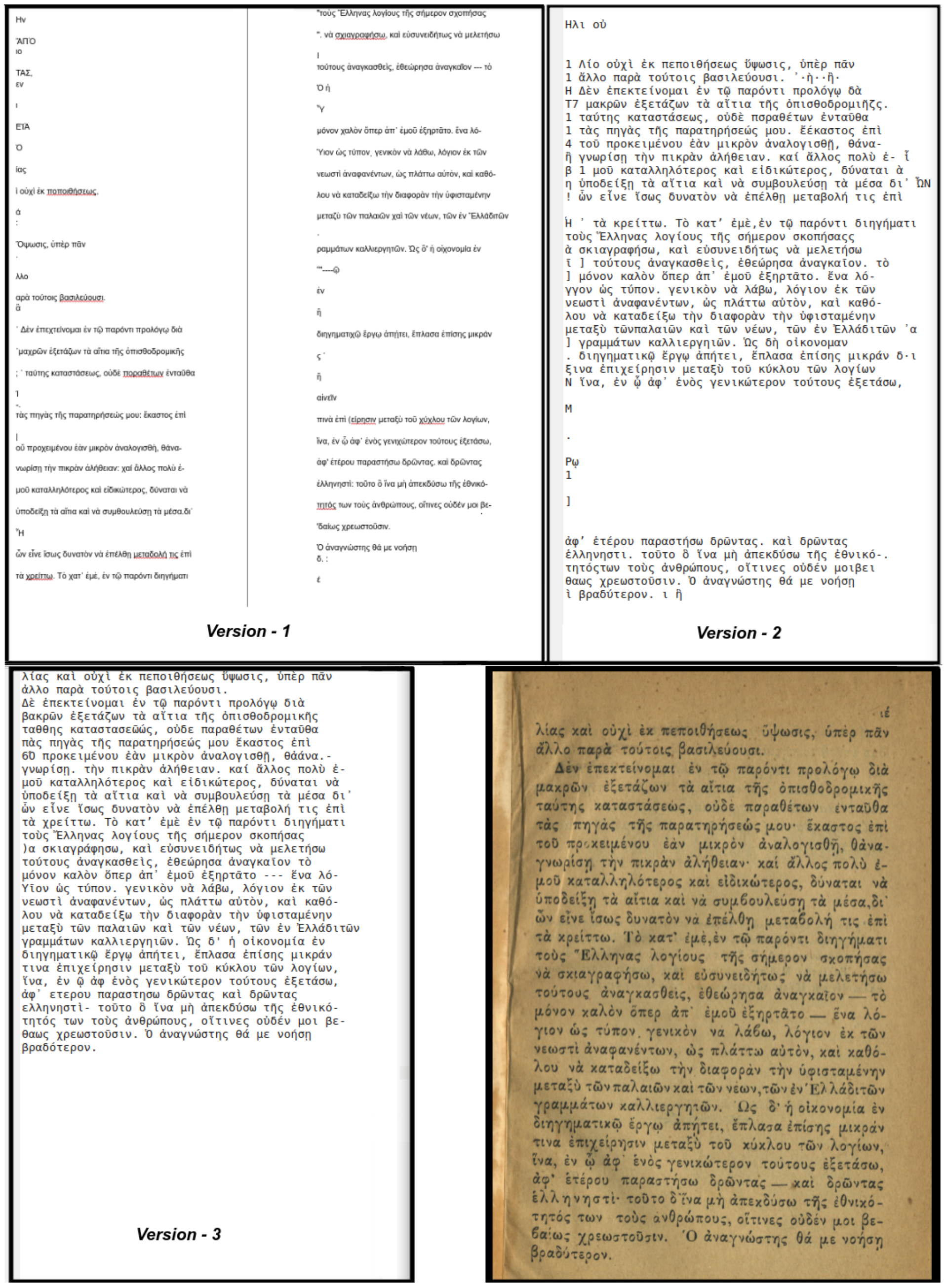
```
ΈΕΠΙ ΤΟΥ ΠΕΔΙΟΫ ΤΗΣ ΜΑΧΗΣ
« Ὀκτὼ βρίσκει, φάντες χάνει.
Πάρολι!
Τέρτσο τίρο...
Κι΄ αὐτὰ στὸν ἄσσο.
Εμπρός.
Πέντε βρίσκει, δέκα χάνει.
Τὸ σταυρό του, στὸν πρῶτο λύκο!
Ωσσος βρίσκει, ρήγας χάνει.
Τέλος πάντων.
Κύριε λοχαγέ...
' Πέντε ἦταν ἐκεῖνο;
Δὲ θέλω πειὰ μάρκες.
. Στὴ ντάμα. »
Αἱ λέξεῖς ἀντήχουν ὑπόκωφοι καὶ μυστηριώδεις
ἐντὸς τῆς ὑπογείου αἰθούσης μὲ τὰς στήλας, τὰς ὑ-
ποβασταζούσας κάτωθεν τὸ δάπεζον τοῦ μεγάλου
```

ΕΠΙ ΤΟΥ ΠΕΔΙΟΥ ΤΗΣ ΜΑΧΗΣ

« Ὀκτὼ βρίσκει, φάντες χάνει.
— Πάρολι !
— Τέρτσο τίρο. . .
— Κι' αὐτὰ στὸν ἄσσο.
— Ἐμπρός.
— Πέντε βρίσκει, δέκα χάνει.
— Τὸ σταυρό του, στὸν πρῶτο λύκο !
— Ἄσσος βρίσκει, ρήγας χάνει.
— Τέλος πάντων.
— Κύριε λοχαγέ . . .
— Πέντε ἦταν ἐκεῖνο ;
— Δὲ θέλω πειὰ μάρκες.
— Στὴ ντάμα. »

Αἱ λέξεις ἀντήχουν ὑπόκωφοι καὶ μυστηριώδεις ἐντὸς τῆς ὑπογείου αἰθούσης μὲ τὰς στήλας, τὰς ὑ-ποβασταζούσας κάτωθεν τὸ δάπεδον τοῦ μεγάλου

Fig. 3: Model versions. The improved output of the third version of a random normal image sample.



```
Ην

ἈΠΌ
ιο

ΤΑΣ,
εν

ι

ΕΊΑ

Ό

ίας

ὶ οὐχὶ ἐκ ποποιθήσεως.

ά
:

Ὄψωσις, ὑπὲρ πᾶν
.

λλο

αρὰ τούτοις βασιλεύουσι.
ἄ

᾿ Δὲν ἐπεχτείνομαι ἐν τῷ παρόντι προλόγῳ διὰ

᾿μαχρῶν ἐξετάζων τὰ αἴτια τῆς ὀπισθοδρομικῆς

; ᾿ ταύτης καταστάσεως, οὐδὲ πoραθέτων ἐνταῦθα

῏Ι
-.
τὰς πηγὰς τῆς παρατηρήσεώς μου: ἕκαστος ἐπὶ

|
οῦ προχειμένου ἐὰν μικρὸν ἀναλογισθῇ, θάνα-

νωρίσῃ τὴν πικρὰν ἀλήθειαν: χαί ἄλλος πολὺ ἐ-

μοῦ καταλληλότερος καὶ εἰδικώτερος, δύναται νὰ

ὑποδείξῃ τὰ αἴτια καὶ νὰ συμθουλεύσῃ τὰ μέσα.δι᾿

῞Η

ὧν εἶνε ἴσως δυνατὸν νὰ ἐπέλθῃ μεταδολή τις ἐπὶ

τὰ χρείττω. Τὸ χατ᾿ ἐμέ, ἐν τῷ παρόντι διηγήματι

"τοὺς Ἕλληνας λογίους τῆς σήμερον σχοπήσας

". νὰ σχιαγραφήσω, καὶ εὐσυνειδήτως νὰ μελετήσω

ι
τούτους ἀναγκασθείς, ἐθεώρησα ἀναγκαῖον --- τὸ

Ὁ ἡ

῎Υ

μόνον χαλὸν ὅπερ ἀπ᾿ ἐμοῦ ἐξηρτᾶτο. ἕνα λό-

῾Υιον ὡς τύπον, γενικὸν νὰ λάθω, λόγιον ἐκ τῶν

νεωστὶ ἀναφανέντων, ὡς πλάττω αὐτὸν, καὶ καθό-

λου νὰ καταδεῖξω τὴν διαφορὰν τὴν ὑφισταμένην

μεταξὺ τῶν παλαιῶν χαὶ τῶν νέων, τῶν ἐν Ἑλλάδιτῶν
.
ραμμάτων καλλιεργητῶν. Ὡς δ' ἡ οἰχονομία ἐν

""----ῳ

ἐν

ῆ

διηγηματιχῷ ἔργῳ ἀπήτει, ἔπλασα ἐπίσης μικράν

ς ῾

ῆ

αἰνεῖν

πινὰ ἐπὶ (εἴρησιν μεταξὺ τοῦ χύχλου τῶν λογίων,

ἵνα, ἐν ῷ ἀφ' ἑνὸς γενιχώτερον τούτους ἐξετάσω,

ἀφ' ἑτέρου παραστήσω δρῶντας. καὶ δρῶντας

ἐλληνηστὶ: τοῦτο ὃ ἵνα μὴ ἀπεκδύσω τῆς ἐθνικό-

τητός των τοὺς ἀνθρώπους, οἵτινες οὐδὲν μοι βε-

'δαίως χρεωστοῦσιν.

Ὁ ἀναγνώστης θὰ με νοήσῃ
δ. :

ἑ
```

```
Ηλι οὐ


1 Λίο οὐχὶ ἐκ πεποιθήσεως ὕψωσις, ὑπὲρ πᾶν
1 ἄλλο παρὰ τούτοις βασιλεύουσι. ᾿·ὴ··ῆ·
Η Δὲν ἐπεκτείνομαι ἐν τῷ παρόντι προλόγῳ δὰ
Τ7 μακρῶν ἐξετάζων τὰ αἴτια τῆς ὀπισθοδρομιῆζς.
1 ταύτης καταστάσεως, οὐδὲ πσραθέτων ἐνταῦθα
1 τὰς πηγὰς τῆς παρατηρήσεώς μου. ἕέκαστος ἐπὶ
4 τοῦ προκειμένου ἐὰν μικρὸν ἀναλογισθῇ, θάνα-
ῆ γνωρίσῃ τὴν πικρὰν ἀλήθειαν. καί ἄλλος πολὺ ἐ- ἵ
β 1 μοῦ καταλληλότερος καὶ εἰδικώτερος, δύναται ὰ
η ὑποδείξῃ τὰ αἴτια καὶ νὰ συμβουλεύσῃ τὰ μέσα δι᾿ ὯΝ
! ὧν εἶνε ἴσως δυνατὸν νὰ ἐπέλθῃ μεταβολή τις ἐπὶ

Ἡ ᾿ τὰ κρείττω. Τὸ κατ’ ἐμὲ,ἐν τῷ παρόντι διηγήματι
τοὺς Ἕλληνας λογίους τῆς σήμερον σκοπήσαςς
ὰ σκιαγραφήσω, καὶ εὐσυνειδήτως νὰ μελετήσω
ῑ ] τούτους ἀναγκασθεὶς, ἐθεώρησα ἀναγκαῖον. τὸ
] μόνον καλὸν ὅπερ ἀπ᾿ ἐμοῦ ἐξηρτᾶτο. ἕνα λό-
γγον ὡς τύπον. γενικὸν νὰ λάβω, λόγιον ἐκ τῶν
νεωστὶ ἀναφανέντων, ὡς πλάττω αὐτὸν, καὶ καθό-
λου νὰ καταδείξω τὴν διαφορὰν τὴν ὑφισταμένην
μεταξὺ τῶνπαλαιῶν καὶ τῶν νέων, τῶν ἐν Ἑλλάδιτῶν ᾿α
] γραμμάτων καλλιεργηιῶν. Ὡς δὴ οἰκονομαν
. διηγηματικῷ ἔργῳ ἀπήτει, ἔπλασα ἐπίσης μικράν δ·ι
ξινα ἐπιχείρησιν μεταξὺ τοῦ κύκλου τῶν λογίων
Ν ἵνα, ἐν ᾧ ἀφ᾿ ἑνὸς γενικώτερον τούτους ἐξετάσω,

Μ

.

Ρῳ
1

]


ἀφ’ ἑτέρου παραστήσω δρῶντας. καὶ δρῶντας
ἑλληνηστι. τοῦτο ὃ ἵνα μὴ ἀπεκδύσω τῆς ἐθνικό-.
τητόστων τοὺς ἀνθρώπους, οἵτινες οὐδέν μοιβει
θαως χρεωστοῦσιν. Ὁ ἀναγνώστης θά με νοήσῃ
ὶ βραδύτερον. ι ῆ
```

```
λίας καὶ οὐχὶ ἐκ πεποιθήσεως ὕψωσις, ὑπὲρ πᾶν
ἀλλο παρὰ τούτοις βασιλεύουσι.
Δὲ ἐπεκτείνομαι ἐν τῷ παρόντι προλόγῳ διὰ
βακρῶν ἐξετάζων τὰ αἴτια τῆς ὀπισθοδρομικῆς
ταθθης καταστασεῶώς, οὐδε παραθέτων ἐνταῦθα
πὰς πηγὰς τῆς παρατηρήσεώς μου ἕκαστος ἐπὶ
6Ū προκειμένου ἐὰν μικρὸν ἀναλογισθῇ, θάάνα.-
γνωρίσῃ. τὴν πικρὰν ἀλήθειαν. καί ἄλλος πολὺ ἐ-
μοῦ καταλληλότερος καὶ εἰδικώτερος, δύναται νὰ
ὑποδείξῃ τὰ αἴτια καὶ νὰ συμβουλεύσῃ τὰ μέσα δι᾿
ὧν εἶνε ἴσως δυνατὸν νὰ ἐπέλθῃ μεταβολή τις ἐπὶ
τὰ χρείττω. Τὸ κατ’ ἐμὲ ἐν τῷ παρόντι διηγήματι
τοὺς Ἕλληνας λογίους τῆς σήμερον σκοπήσας
)α σκιαγράφησω, καὶ εὐσυνειδήτως νὰ μελετήσω
τούτους ἀναγκασθεὶς, ἐθεώρησα ἀναγκαῖον τὸ
μόνον καλὸν ὅπερ ἀπ᾿ ἐμοῦ ἐξηρτᾶτο --- ἕνα λό-
Υῖον ὡς τύπον. γενικὸν νὰ λάβω, λόγιον ἐκ τῶν
νεωστὶ ἀναφανέντων, ὡς πλάττω αὐτὸν, καὶ καθό-
λου νὰ καταδείξω τὴν διαφορὰν τὴν ὑφισταμένην
μεταξὺ τῶν παλαιῶν καὶ τῶν νέων, τῶν ἐν Ἑλλάδιτῶν
γραμμάτων καλλιεργηιῶν. Ὡς δ' ἡ οἰκονομία ἐν
διηγηματικῷ ἔργῳ ἀπήτει, ἔπλασα ἐπίσης μικράν
τινα ἐπιχείρησιν μεταξὺ τοῦ κύκλου τῶν λογίων,
ἵνα, ἐν ῷ ἀφ ἑνὸς γενικώτερον τούτους ἐξετάσω,
ἀφ᾿ ετερου παραστησω δρῶντας καὶ δρῶντας
ελληνηστὶ- τοῦτο ὃ ἵνα μὴ ἀπεκδύσω τῆς ἐθνικό-
τητός των τοὺς ἀνθρώπους, οἵτινες οὐδέν μοι βε-
θαως χρεωστοῦσιν. Ὁ ἀναγνώστης θά με νοήσῃ
βραδότερον.
```

ιέ

λίας καὶ οὐχὶ ἐκ πεποιθήσεως ὕψωσις, ὑπὲρ πᾶν ἄλλο παρὰ τούτοις βασιλεύουσι.

Δὲν ἐπεκτείνομαι ἐν τῷ παρόντι προλόγῳ διὰ μακρῶν ἐξετάζων τὰ αἴτια τῆς ὀπισθοδρομικῆς ταύτης καταστάσεως, οὐδὲ παραθέτων ἐνταῦθα τὰς πηγὰς τῆς παρατηρήσεώς μου· ἕκαστος ἐπὶ τοῦ προκειμένου ἐὰν μικρὸν ἀναλογισθῇ, θὰ ἀναγνωρίσῃ τὴν πικρὰν ἀλήθειαν· καί ἄλλος πολὺ ἐμοῦ καταλληλότερος καὶ εἰδικώτερος, δύναται νὰ ὑποδείξῃ τὰ αἴτια καὶ νὰ συμβουλεύσῃ τὰ μέσα, δι᾿ ὧν εἶνε ἴσως δυνατὸν νὰ ἐπέλθῃ μεταβολή τις ἐπὶ τὰ κρείττω. Τὸ κατ᾿ ἐμέ, ἐν τῷ παρόντι διηγήματι τοὺς Ἕλληνας λογίους τῆς σήμερον σκοπήσας νὰ σκιαγραφήσω, καὶ εὐσυνειδήτως νὰ μελετήσω τούτους ἀναγκασθείς, ἐθεώρησα ἀναγκαῖον — τὸ μόνον καλὸν ὅπερ ἀπ᾿ ἐμοῦ ἐξηρτᾶτο — ἕνα λόγιον ὡς τύπον γενικὸν νὰ λάβω, λόγιον ἐκ τῶν νεωστὶ ἀναφανέντων, ὡς πλάττω αὐτὸν, καὶ καθόλου νὰ καταδείξω τὴν διαφορὰν τὴν ὑφισταμένην μεταξὺ τῶν παλαιῶν καὶ τῶν νέων, τῶν ἐν Ἑλλάδι τῶν γραμμάτων καλλιεργητῶν. Ὡς δ᾿ ἡ οἰκονομία ἐν διηγηματικῷ ἔργῳ ἀπήτει, ἔπλασα ἐπίσης μικράν τινα ἐπιχείρησιν μεταξὺ τοῦ κύκλου τῶν λογίων, ἵνα, ἐν ᾧ ἀφ᾿ ἑνὸς γενικώτερον τούτους ἐξετάσω, ἀφ᾿ ἑτέρου παραστήσω δρῶντας — καὶ δρῶντας ἑλληνηστί· τοῦτο δ᾿ ἵνα μὴ ἀπεκδύσω τῆς ἐθνικότητός των τοὺς ἀνθρώπους, οἵτινες οὐδὲν μοι βεβαίως χρεωστοῦσιν. Ὁ ἀναγνώστης θά με νοήσῃ βραδύτερον.

Fig. 4: Model versions. The improved output of the third version of a random distorted image sample.

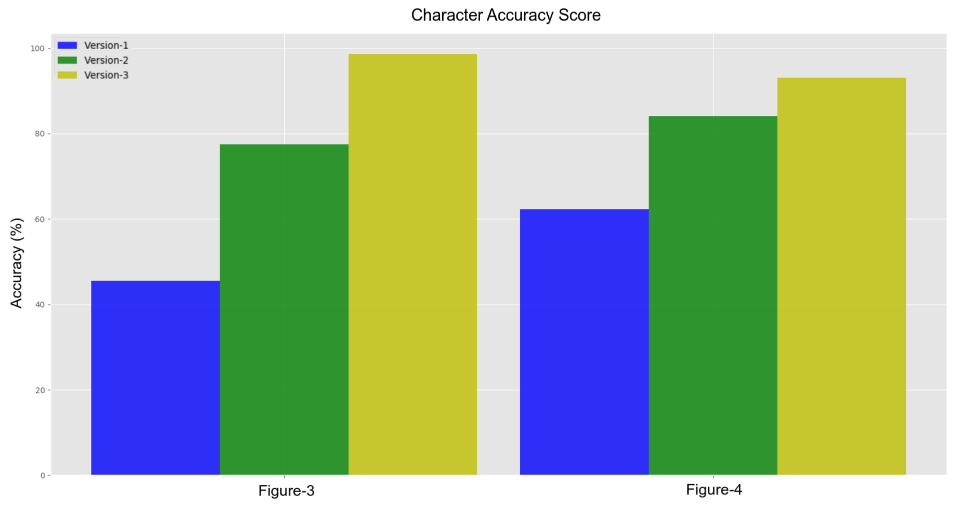


Fig. 5: Model versions. Bar chart of character accuracy for the outputs of Figure 3 and Figure 4.

text has already been produced. Each one of the implemented versions was evaluated on the specific set and the character accuracy was calculated, using ocreval tool. Table 2 summarizes the results of the evaluation, highlighting the superiority of the proposed model over the rest models for a numerous evaluation set.

Evaluation set: 102 image samples			
Character Accuracy (%)	Version-1	Version-2	Version-3
	62.30	84.04	93.06

Table 2: Results of versions' evaluation on a set of 102 image samples

7 Future Challenges

There are several challenges for future research. Firstly, the proposed method selects Tesseract open-source OCR engine as well as further training in order to achieve the optimal OCR accuracy for Greek Literature documents. To this end, further accuracy comparison with other current approaches could be incorporated. Furthermore, a sophisticated heuristic or machine learning approach could apply layout recognition that would involve the division of a whole document into front/body/back sections as well as the detection of headers, chapters, paragraphs, page numbers, etc., within a document's page.

References

[1] Vandegrift, M, Varner, S.: Evolving in Common: Creating Mutually Supportive Relationships Between Libraries and the Digital Humanities. In: Journal of Library Administration**53**(1) 67–78 (2013)

[2] Springmann, U., Lüdeling. A.: OCR of Historical Printings with an Application to Building Diachronic Corpora: A Case Study Using the RIDGES Herbal Corpus. In ArXiv.org (2017), `arxiv.org/abs/1608.02153`. Last accessed 2 Nov 2020

[3] Rahnemoonfar, M., Antonacopoulos. A.: Restoration of Arbitrarily Warped Historical Document Images Using Flow Lines. In: Proceedings of the 2011 International Conference on Document Analysis and Recognition, pp. 905–909. IEEE Computer Society, Washington (2011)

[4] Roy, S., Adhikari, G., Dasgupta T., Pradhan T.: An Adaptive Warp Correction Algorithm for Handwritten Text Images with Non-Linear Baselines. In: Proceedings of the 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1–7. IEEE Computer Society, Washington (2018)

[5] Zhang, Y., Zhong, J., Yu H., Kong. L.: Research on Deskew Algorithm of Scanned Image. In: Proceedings of the 2018 IEEE International Conference on Mechatronics and Automation (ICMA 2018), pp. 397–402. IEEE Computer Society, Washington (2018)

[6] Traub, M. C., van Ossenbruggen. J., Hardman. L.: Impact Analysis of OCR Quality on Research Tasks in Digital Archives. In: Research and Advanced Technology for Digital Libraries Lecture Notes in Computer Science **9316** 252–263 (2015)

[7] Strange, C., McNamara, D., Wodak, J., Wood, I.: Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers. In: Digital Humanities Quarterly **8**(1) (2014), `http://www.digitalhumanities.org/dhq/vol/8/1/000168/000168.html`. Last accessed 4 Oct 2017

[8] Reul, C., Springmann, U., Wick, C., Puppe, F.: State of the Art Optical Character Recognition of 19th Century Fraktur Scripts Using Open Source Engines. In ArXiv.org (2018), `https://arxiv.org/abs/1810.03436`. Last accessed 2 Nov 2020

[9] Breuel, T. M.: The OCRopus open source OCR system. In: Proceedings of SPIE - International Society for Optical Engineering, `https://www.deepdyve.com/lp/spie/the-ocropus-open-source-ocr-system-5xECzD6Gu0`. Last accessed 4 Dec 2020

[10] Smith, R.: An Overview of the Tesseract OCR Engine. In: 9th International Conference on Document Analysis and Recognition (ICDAR 2007), pp. 629–633, IEEE Computer Society, Washington (2007)

[11] Dojčinović, N., Mihajlović, I., Joković, J., Marković, V., Milovanović, B.: Neural Network Based Optical Character Recognition System. In: Proceedings of the 11th Symposium on Neural Network Applications in Electrical Engineering, pp. 111–114. IEEE Computer Society, Washington (2012)

[12] Ares, O. S., Seguin, B., Kaplan, F.: dhSegment: A generic deep-learning approach for document segmentation. In: 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 7–12. IEEE Computer Society, Washington (2018)

[13] Guo, Y., Sun, Y., Bauer, P., Allebach, J. P., Bouman, C. A.: Text Line Detection Based on Cost Optimized Local Text Line Direction Estimation. In: SPIE Proceedings on Color Imaging XX: Displaying, Processing, Hardcopy, and Applications, pp. 939507. SPIE, California (2015)

[14] Bieniecki, W., Grabowski, S., Rozenberg, W.: Image Preprocessing for Improving OCR Accuracy. In: Proceedings of 2007 International Conference on Perspective Technologies and Methods in MEMS Design, 2007, pp. 75–80. IEEE Computer Society, Washington (2017)

[15] Raid, A.M., Khedr, W.M., El-dosuky, M.A., Aoud, M.: Image Restoration Based on Morphological Operations. In International Journal of Computer Science, Engineering and Information Technology **4**(3) 9–21 (2014)

[16] Sane, P., Agrawal, R.: Pixel Normalization from Numeric Data as Input to Neural Networks: For Machine Learning and Image Processing. In: Proceedings of 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), pp. 2221–2225. IEEE Computer Society, Washington (2017)

[17] Santos, E. A.: OCR Evaluation Tools for the 21st Century. In: Proceedings of the Workshop on Computational Methods for Endangered Languages (2019), `https://journals.colorado.edu/index.php/computel/article/view/345`. Last accessed 4 Dec 2020.