Automatic Metadata Extraction via Image Processing Using Migne's Patrologia Graeca

Evagelos Varthis, Marios Poulos, Ilias Giarenis, Sozon Papavlasopoulos

A robust framework for segmentation and word-spotting on the Greek polytonic script of Migne's Patrologia Graeca

Evagelos Varthis[1*], Marios Poulos[1], Ilias Giarenis[2], Sozon Papavlasopoulos[1]

1 Ionian University, Library Science Department, Greece

2 Ionian University, History Department, Greece

*Corresponding author: evagelosvar@gmail.com

Abstract
A wealth of knowledge is kept behind libraries and cultural institutions in various digital forms without however the possibility of a simple term search, let alone of a substantial semantic search. One such important collection that contains knowledge, accumulated in the passage of the ages and remain inaccessible for the greater part, is Patrologia Graeca. So far, little research has been conducted to make this digital collection searchable to a certain degree, in order to retrieve and reveal its gathered knowledge in an efficient way. In this study, a novel approach is proposed which strives towards recognizing words from large printed corpora such as Patrologia Graeca. The proposed framework firstly applies an efficient segmentation process at word level and transforms the word-images of Greek polytonic script of the Patrologia Graeca into special compact shapes. Afterwards the contours of these shapes are extracted and compared with the contour of a similarly transformed query word-image in order to locate the specific word in the digitized documents. For the comparison, we use a series of three descriptors, Hu's invariant moments for discarding unlikely similar matches, Shape Context for the contour similarity and the Pearson's correlation coefficient for final pruning of the dissimilar words and additional verification. Comparative results are presented by using instead of Pearson's correlation coefficient the Long-Short Term Memory Neural Network engine of Tesseract Optical Character Recognition system. The described framework due to the simplicity and efficiency that provides, can be applied for massive creation of search indexes and consequently semantic enrichment of Patrologia Graeca. The framework has the potential to be applicable for other printed collections with proper configuration of the parameters. An additional and very significant consequence of our method's effectiveness and simplicity is that it can be used as a pre-stage to provide a large number of word-image and label pairs, These pairs can be used for training neural networks or common classifiers such as k-nearest neighbor or state vector machine.

keywords
Patrologia Graeca; Word Spotting; Shape Context; Time Series

I. INTRODUCTION

According to the works presented in [12][29][45] the methods of retrieving word-images through printed or handwritten texts can be classified into two categories:

- recognition-based retrieval
- recognition-free retrieval

The recognition-based retrieval requires a complete recognition of the characters by the use of Optical Character Recognition (OCR) on different sections of the document. On the other hand, the method of the free recognition searches on pages the words that are likely to exist in the documents, and retrieves these pages bypassing the character recognition.

The recognition-free retrieval is alternatively known in the literature as word spotting or

keyword spotting and has a rising potential for classifying documents mainly in areas where traditional OCR systems fail.

In recognition-free retrieval the input methods for the query image can be distinguished into two categories [12][45], depending on how the input is provided by the user:

- query-by-example (QBE)
- query-by-string (QBS) methods

In the QBE framework, a query image is selected to be searched in the document collection whereas in QBS method a text string is provided by the user to the system in order to spot the desired word [22]. Word spotting techniques are alternative solutions for OCRa nd can be applied mainly in areas that OCR fails due to unknown printed fonts or degraded printed texts in order to build efficiently, search indexes that help acquire the knowledge from scanned texts [12][19][45] [29].

The word spotting approach has the advantage of a small execution time compared to OCR, and the robustness to noisy documents [35][13]. Nowadays, word spotting is actively studied for handwritten texts [39] [34] due to the great difficulties that the OCR faces in handling and recognizing properly such scripts. However it is not less applicable for digitized printed texts such as Patrologia Graeca (PG) which also presents difficulties for the OCR because of its polytonic script and the degraded and noisy scanned pages [39].

It is worth to remind that PG is a very special collection that contains knowledge accumulated over a period of 1500 years by east Christian fathers presenting 4,287 works written by 658 authors as identified by Perseus Digital Library (PDL) [51].

PG texts cover a broad range of interests such as theology matters and doctrine issues as in Athanasius of Alexandria, connection with the modern psychology as in Macarius the Great, social events and reports about the everyday life in the medieval times, music, politics, history and many more.

The collection was published in 19th century (1857-66) by Jacques-Paul Migne and consists of 166 volumes bound as 161 with an enormous size roughly estimated at 122500 pages with Latin translations and accompanied scholarship.

PG has been digitized and is available on the Web by the large scale digitization project of Google [52][46] and others [36][51][53]. However, the access of PG for semantic navigation or simple searching is very limited. Google has applied OCR over the pages of PG with results ranging from acceptable to very bad as far as it concerns the recognition of words.

Unlike with Google, a different approach is adopted by Thesaurus Linguae Graeca (TLG) where, they converted in editable text approximately 20% of PG works with extensive writing for years. The most notable authors such as Basilius Caesariensis, Gregory of Nazianzus, Athanasius the Great of Alexandria e.t.c are included, however a vast amount of knowledge is still in images (nearly 140 authors and 1524 works have been transformed to editable text). A weakness that appears in the case of TLG is that the editable collections do not include the comments or the external references that PG often provides, with smaller fonts, in the footer section. Yet another important issue is that the material of TLG falls under specific license and the free distribution is prohibited, however for research issues is allowed under certain limitations [53]. Therefore, we consider for the aforementioned reasons, that the provision for searching via word spotting on the scanned PG pages will add to them significant semantic enrichment. Such type of service forms a powerful tool for the scholars to locate additional information or interesting interconnected knowledge. The objective of this study is to provide a simple however robust method for word-spotting on the scanned texts of PG as it is described in following sections. As far as we are aware such methodology has not been described in the literature, however a close similarity to a part of our framework can be considered to be the well known upper and lower word-profiles [35][33] as they are used in the third stage of our similarity method. The overview of our methodology is the following:

We apply to the pages segmentation at the word level, by using erosion with a horizontal kernel and converting each page into a set of word images. During the segmentation phase a specialized technique is applied to omit the punctuation marks and to preserve the diacritics of the words.

Afterwards we transform the Query Word Image (QWI) into a compact shape by applying erosion with a circle kernel on its characters. We extract the contour of QWI and apply in series two shape descriptors to test the similarity of QWI with the segmented word images of the candidate document that is going to be recognized. During the comparative phase the segmented words have also been transformed to compact shapes with the same erosion method we applied to the QWI. The two shape descriptors used for the word-image similarity are, HU's moments for cutting-off the very dissimilar word-images and Shape Context (SC) for contour matching. As final verification we apply Pearson's correlation coefficient (PCC) on the upper and lower profiles of the candidate words. Our method can be applied to PG as well as to other printed corpora by proper parameter tuning, since the framework remains essentially the same or, in some cases, even easier to be applied compared to the difficulties of the ancient Greek polytonic script. Additionally, it is worth pointing out that our system applies the word spotting method on PG texts for the Greek words while at the same time the Latin script translations exist in the pages, a fact that in itself is a challenging one.

This paper is organized as follows: In Section 2, we briefly summarize some of the related works for word-spotting on handwritten and printed texts (mainly for Greek polytonic texts) as well as the specific particularities of the PG corpus. In Section 3, we describe the background theory used for the matching of the transformed word-images, specifically, HU moments, and the SC . In Section 4 a detailed analysis of the methodology of the proposed framework system is presented and the challenges we met are discussed. The segmentation and similarity processes are described with the use of proper illustrations. Additionally the PCC is briefly presented. In Section 5 experimental results of our method are depicted. The benefits, the limitations and the possible extensions of our method to other areas are also discussed. Moreover, comparative results are presented for the case of using the Long Short Term Memory (LSTM) Neural Network (NN) engine of Tesseract OCR software.

II. RELATED WORKS FOR WORD-SPOTTING AND PARTICULARITIES OF PATROLOGIA GRAECA

A variety of word spotting techniques for document indexing and retrieval have been proposed for printed and handwritten documents and are discussed in the detailed and in-depth survey of Giotis et al. [12], who also provides an extensive list with the specific methods that have been applied. In [29] Murugappan et al. present a comprehensive study for word spotting techniques, focusing mainly on printed documents. In their study the word-spotting matching methods are categorized either at a pixel level or at a feature level. At a pixel level Statistical methods, Hausdorff methods and Coarse feature based methods are commonly used whereas at a feature level Primitive String methods and Geometric Feature Graph) methods are applied. Indicatively, Rios et.al. in [35] use simple features based on projection profile, upper word profile, lower word profile and column transitions to represent handwritten or printed word bitmap image. Gatos and Pratikakis in [11] detect the salient regions in the printed document, extract several word instances to capture variations in order to define a group of featured vectors that are used for matching with the template image. In [23] Konidaris et al. adopt a segmentation-free approach for word spotting based on SIFT descriptor through a two step process. Firstly, candidate areas of the query keyword image are located in order to narrow the search and secondly the query is compared for matching with

the candidate areas for the creation of the final bounding boxes.

As far as we are aware in the literature, the PG has not been explicitly studied for applying word spotting techniques, however Konidaris et.al in [22] investigate fragments of historical poly-tonic printed Greek texts during the period of Renaissance and Enlightenment (1471–1821) and employ a methodology that has the potential to be applied to other than Greek historical printed documents. Also in [39] Sfikas et.al present a study for polytonic texts of modern history and especially in official journals of Greek government.

It is worth to note that PG is printed by using the polytonic Greek script which has 24 upper-case and lower-case characters as well as diacritics over certain characters according to grammatical and syntactical rulers [9]. These diacritics help the reader on how the words are pronounced and emphasized. These diacritics are the accute, grave and circumflex accent, smooth and rough breathing, the subscript, the diaeresis and are shown briefly in Table 1

This particularity creates additional difficulties, mainly at the level of character segmentation and consequently in their recognition with traditional OCR systems[37][17][10]. Moreover, the text degradation such as stained paper, faded ink, connected characters as well as the scanning process that introduces skewing, low contrast, warping effects e.t.c. increases further the difficulty for any OCR. Additional works related to Greek polytonic script can be found by Katsouros et al. [17] where two set of features from each text line are extracted by using a sliding window and model each polytonic character with a multistate, left to right HMM while Gatos et al. [10] propose an OCR framework in which the divide each segmented text line in three horizontal zones and build five recognition modules for characters and accents. An alternative to the aforementioned image feature methods are Neural Networks due to the success they have in recognizing letters or numbers on images [1][28][42][44][48] as well as discovering patterns, however they need a large number of sample data in order to achieve good results. Moreover, as stated in [12], applying a NN for character recognition in a specific language script, does not necessarily fulfill the condition that this NN is applicable to another language script. An additional collection of new sample data is required to implement NN in the new language framework, and perhaps a different NN structure is required. NN applications on large image collections such as PG with the rich combination of letters and diacritics is therefore a tedious and laborious task, so the efficiency of its application is questionable and a case-by- case analysis is generally suggested. Unlike neural networks and their complexities, our proposed framework does a direct segmentation and spotting of the words of interest on the given PG image pages, a task that greatly simplifies the recognition procedure.

Table 1 Acient Greec script diacritics.

Diacritics				**Word examples**
Accents	Acute	´	ό	χριστοφόρου
	Grave	`	ὶ	χριστιανικαὶ
	Circumflex	~	ῖ	χριστοειδεῖς
Breathings	Smooth	᾽	ἀ	ἀληθείας
	Rough	῾	ἑ	ἑκάτερος
Subscript		˛	ᾳ	παιδείᾳ
Diaeresis		¨	ΐ	παιδοποιΐᾳ

III. Background Theory

In this section we briefly review the HU moments and the SC descriptor while as aforementioned, PCC is reviewed in Section 4 together with the description of the third stage

similarity process for a better readability. HU, SC and PCC constitute the basic tool-case of our proposed framework in order to apply efficient word-spotting on PG scanned texts

3.1 Hu moments

Hu's moments introduced in [15] are formed by six absolute orthogonal invariants and one skew invariant. These seven moments perform scale, rotation and translation invariance and also they are invariant to parallel projection. Their use for spotting patterns on images has been proven very effective and they have been used numerous times in literature for retrieving similar images [16][32].

The complete set of the seven algebraic equations are omitted because Hu's moments theory is a broadly known subject, however a detailed explanatory analysis can be found in [38].

3.2 Shape descriptors

3.2.1 Concise review for shape descriptors

Shape descriptors can be mainly classified into two categories: boundary-based and skeleton-based. In boundary-based representations, shapes are represented by a set of boundary points or by a set of boundary curves [8][40][6][4][27][41].

Unlike boundary-based descriptors, skeleton-based frameworks, seek to capture a structural representation of the shape by the use of a set of axial curves [2][3][24][47][25] while on the same time display insensitivity to articulation and deformation.

3.2.2 Shape context

In our framework we use a boundary based representation descriptor and specifically the SC algorithm. SC [4][31] is one of the most powerful 2D shape descriptors, however does not perform well for shapes with articulation parts [40][27]. Our method is not affected by this limitation of the SC since -as it is described in Subsection 4.5- the QWI under comparison is converted to a simple 2D compact shape without any articulation parts.

Specifically, SC is a rich local descriptor that uses a discrete uniform spacing set of points sampled from the internal or external contours on the shape. The SC describes the spatial locations of the other n−1 sampled points of the shape in a log polar histogram which is defined with 5 bins for log r and 12 for θ and is inherently tolerant to small perturbations of parts of the shape. For a point p_i on a shape P we compute the histogram of the remaining n-1 points as:

$$h_i(k) = \#\{q \neq p_i : (q - p_i) \in bin(k)\}$$

The cost of matching two points p_i and q_j between two shapes P, Q is denoted as $C_{i,j} = C(p_i, q_j)$. Shape contexts are distributions of histograms so χ^2 statistic is used and is given by:

$$C_{i,j} = \frac{1}{2}\sum_{k=1}^{K}\frac{[h_i(k) - h_j(k)]^2}{h_i(k) + h_j(k)}$$

Where $h_i(k)$ and $h_j(k)$ is the $K-bin$ normalized histogram of p_i and q_j. SC is translation invariant by definition since the distances are measured with respect to the sampled point. Scale invariance is obtained by normalizing the distances between points with the mean distance between all n^2 point pairs. Rotation invariance for each point can be achieved by

taking as a reference frame the tangent vector of the point, instead of the positive x-axis. Given the set of costs between all pairs of points on shape P and Q we want to minimize the total cost of matching subject to constrain that the matching be one-to-one, for a specific permutation $\pi(i)$.

$$H(\pi) = \sum_i C(p_i, q_{\pi(i)})$$

This corresponds to a bipartite matching problem which can be solved in $O(N^3)$ time using the Hungarian method [5]. After the matching of the points, a total shape distance metric is defined to capture the contour similarity as it is explained in detail in [31]. This metric is the weighted sum of the following three terms: a) shape context distance, b) image appearance distance and c) bending energy.

IV. Methodology

4.1 Segmentation at word level

In order to evaluate our proposed framework, we collected 46 pages of the work of Didymus Alexandrinus [7] residing in the 39th volume of the PG and applied to them segmentation at the word level. More specifically, we chose the pages with columns numbered in the range from 291 to 382, since each double column page of PG has a particular double numbering.

A comprehensive list of various segmentation methods can be found in .[18][50][49][26], however because the segmentation is not an easy task for the Greek polytonic script due to the existence of diacritics over the characters, modifying techniques are required as it is explained below.

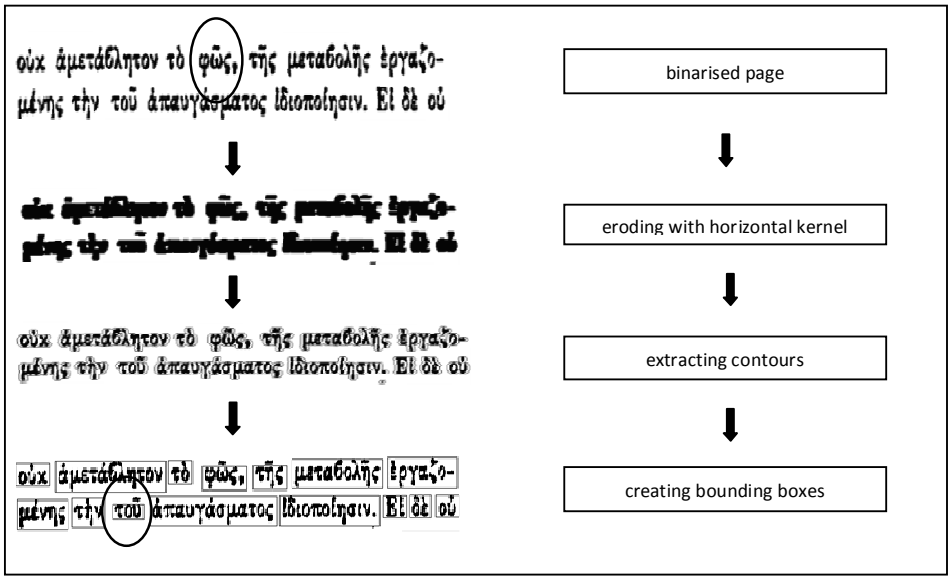

Figure 1. First stage of the segmentation work-flow.

4.2 First segmentation phase

Given the particularity of PG, we apply a simple -however efficient- segmentation method in which an erosion is applied to the characters, so that the characters get connected (forming in this way the desired connected components) having concurrently, the constrain not to fill

the empty space between the words, see Figure 1. The erosion is employed with the use of a horizontal nine pixel kernel ([1,1,1,1,1,1,1,1,1]) for the specific size of pages. The page size is with small variations around 2100x3200 pixels as it was extracted directly from the Google's scanned document and converted to 300 dpi png image. The kernel size varies according to the page size and it is analogous to the page size. For the case of 600 dpi resolution to get same results the suggested kernel size should be doubled.

Having obtained the connected components (CC) we find the contours of the connected regions and subsequently the bounding boxes as described in algorithm [43] and implemented in OpenCV library [30].

However, applying the above kernel is not enough for the proper word segmentation because some of the CCs with small area such as diacritical marks do not form connected components with the corresponding words in which they belong. This mostly happens in cases where the word has only low-case letters with no upper extensions. On the other hand most of the punctuation marks form with the words CCs and a separation is required. See Figure 2 for the word "φῶς," how the diacritic is separated from the desired CC while the punctuation mark is joined.

4.3 Second segmentation phase

In order to properly segment the words joined with the corresponding diacritics marks and discarding the punctuation marks, the following techniques are additionally applied to the bounding boxes we got during the first segmentation stage:

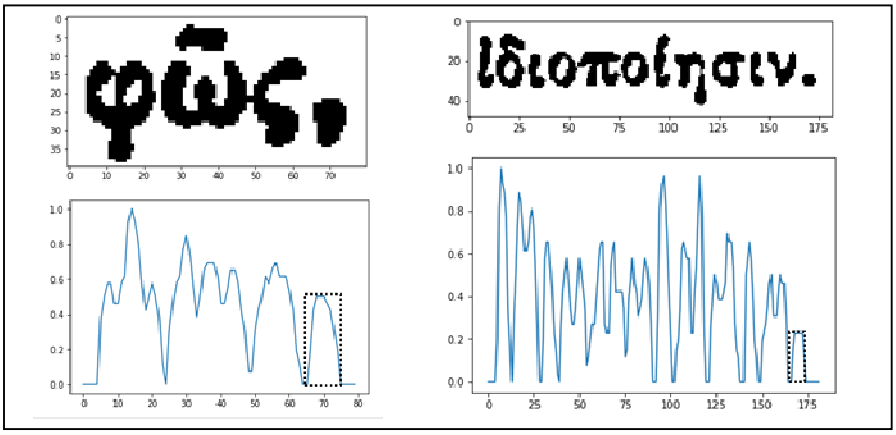

Figure 2. Second stage of the segmentation work-flow, elimination of the punctuation marks.

a) The elimination of punctuation marks from the aforementioned cropped regions is achieved by taking the ink histogram of the cropped region in order to check for a small fluctuation pattern to the tail of the histogram. The image is additionally cropped just before the fluctuation, resulting in the finalized segmented word. The punctuation marks which are successfully eliminated from the pages are: comma (,), semicolon (;), period (.) and upper period (.) , see Figure 2.

b) The joining of the diacritics with the word is achieved by enlarging the bounding box to the upper area by a specific number of pixels (which is depended on the page size) and removing the cropped characters from the upper bound of the newly created bounding box, see Figure 3. For the specific resolution of our pages ten pixels were the appropriate number of pixels to enlarge the bounding box. The above presented segmentation method although simple in implementation, is extremely effective for the PG pages.

The word segmentation error that appears in both the body text of interest and the comments section (a section for commentaries and references) is nearly 4% and mostly happens in the comments section of the pages where the characters and line spaces are slightly smaller, see Figure 4.

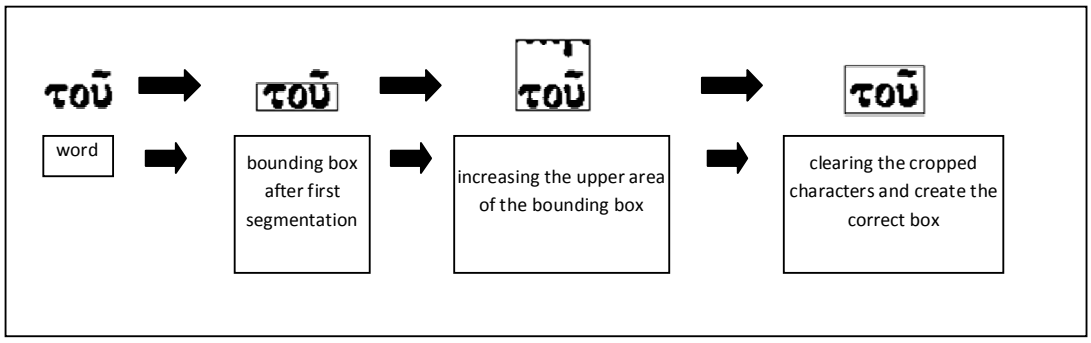

Figure 3. Second stage of the segmentation work-flow, joining of the diacritics.

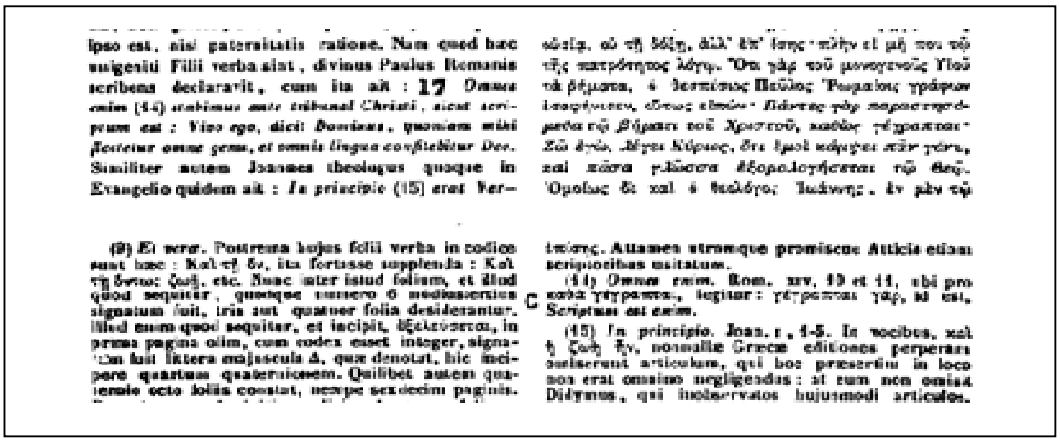

Figure 4. PG page with commentary section.

One limitation of the segmentation process is that it is not simultaneously applicable to the titles and subtitles of PG because the fonts spaces and sizes greatly differ (they are larger) compared with these of body text and comments, so the specific kernel sometimes fails to create the necessary CCs. A larger kernel size is required so as to segment properly the titles and subtitles. Here, it is worth to point out that, the title and subtitles appear only in the first page of any of the multipage PG work of specific author, so it is safe to say that the fail of the title and subtitle segmentation is a rare occurrence. One additional advantage of our segmentation technique is that the segmentation is insensitive to small page rotation and skew, since it bypasses the line segmentation and does direct word segmentation.

4.4 Input method for creating search indexes.

At this point, we have two options for finding the similar words. Either we choose an image from the set of the segmented word-images as query image (applying QBE), or construct a query image by simply writing a string-word as it is proposed by Konidaris et al. [22] (applying QBS). In the first case we have to create a simple database for storing the pairs string and query-image for the interested query-images. In the second case the transformed word image is created directly from the string whatever that string is. The second option is by

far more flexible compared to the first, however misleading minor details to the design may lead to failure of the system. For the evaluation of the framework we choose the first option for simplicity reasons and because the build of a QBS system is out of the scope of the current study.

4.5 Description of the similarity process

After the completeness of the segmentation process every page has been converted to a set of word images. In order to examine the similarity of a QWI with each of the segmented words we apply to them a Compact Shape Transformation (CTS) as it is shown in Figure 5. The transformation is achieved by applying erosion with a circle kernel with a nine pixel radius both on the QWI and the candidate segmented word. It is important to note that the size of the kernel is again analogous with the page resolution. Afterwards, the corresponding contours of the produced compact shapes are extracted.

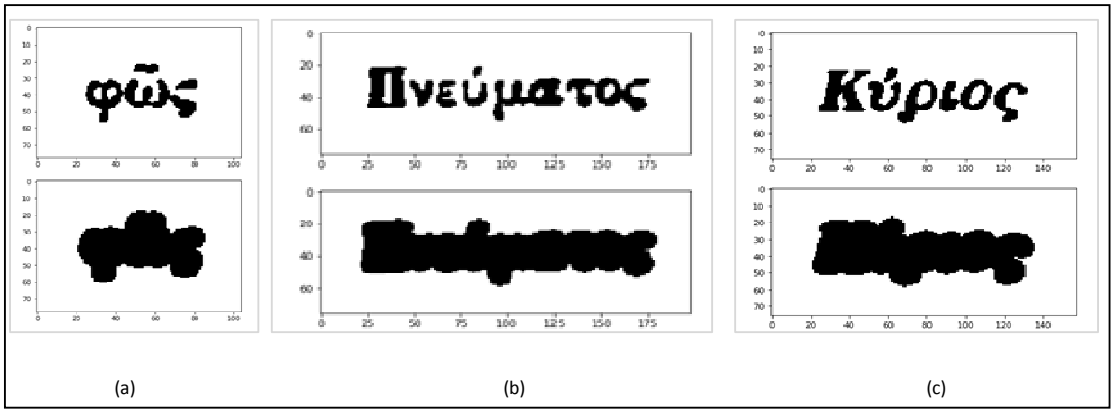

Figure 5. Compact shapes produced by applying erosion with circle kernel on the words.

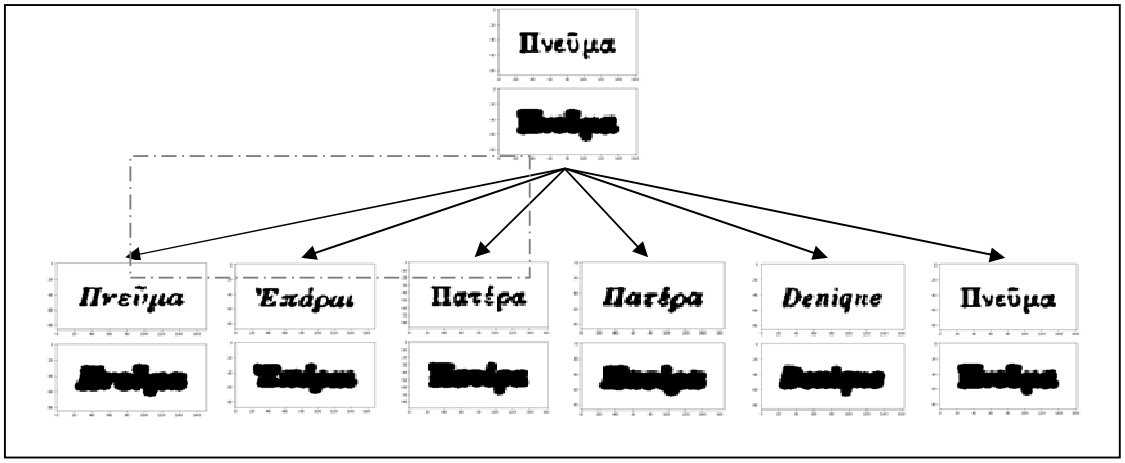

Figure 6. Examples of words transformed to nearly similar compact shapes.

The similarity process has three stages as it is shown in Figure 7. Firstly, the extracted contours are compared using HU's moments in order to narrow the search of the candidate word-images, pruning the very dissimilar contours. The words that successfully pass this stage are fed to the next stage. During the second stage the SC descriptor is applied to cut off more dissimilar images which as aforementioned is a very robust matching algorithm. It is

worth to remind that SC has been applied to the MNIST database giving 99,3 % recognition success for its handwritten digits comparable with the state of the art NNs. It has also been applied to retrieve similar trademarks as well as other shapes [4] with high precision and retrieval rates. The similar images Figure 6 obtained by SC are fed into the third stage in which the upper and lower profile of the words are extracted for the final verification.

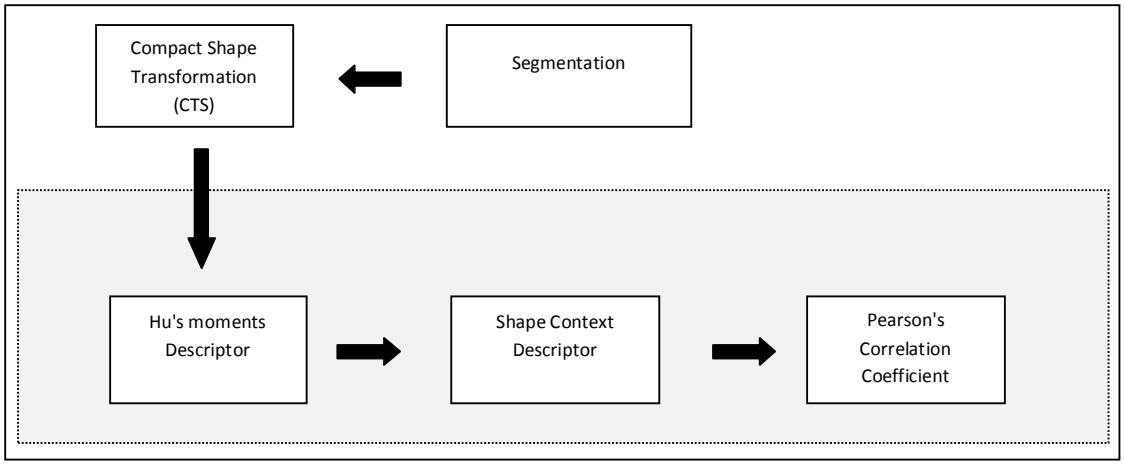

Figure 7. Similarity stages.

The upper and lower profiles of the words are extracted by using the distance from the bounding box to the horizontal dividing line of the word. This dividing line is computed by taking the vertical histogram of the word and finding the two largest maximums. The line is drawn in the middle of these maximums, see Figure 8b. In Figure 8c,d we see the produced upper profile for the word-pairs "*Πνεῦμα*-Πατέρα" and "*Πνεῦμα- Πνεῦμα*" which are properly aligned for comparison.

We model each of these profiles for each pair of words as time series of random variables X,Y respectively as:

$$X= \{X1, X2, ...\} \text{ and } Y= \{Y1, Y2, ...\}$$

We compare these times series for similarity by checking the linear dependence they have with the help of the PCC which is given by:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}$$

where $\mathrm{cov}(X,Y)$ is the covariance of the variables X, Y and σ_X, σ_Y the corresponding standard deviations of the variables X, Y.

PCC when applied to a sample -as it is actually our case- is commonly represented as

$$r_{XY} = \frac{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}\sqrt{\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2}}$$

where $\overline{X}$, $\overline{Y}$ the mean values of the corresponding samples. PCC is selected because it is a

very robust descriptor compared with others descriptors for testing similarity in time series as stated and evaluated in [21]. More specifically, as it is presented, has less sensitivity to small variation of signals compared to the Euclidean distance, Discrete Wavelet Transform, Discrete Fourier Transform, Mahalanobis Distance, Minkowski Distance and Dynamic Time warping.

It is important to note that prior to the application of the PCC to the compared images we delete any row or column of white pixels around the images and resize the images to have the same dimensions (height and width). Finally a border of only one white pixel is created around the bounding boxes of the words and then the upper and lower profiles are extracted. In so doing, the upper and lower profiles of the two words are properly aligned for comparison. Afterwards exponential smoothing is applied to the time series with a factor $a = 0.8$ and the smoothed time series are compared for linear dependence applying PCC. The exponential smoothing series $s = \{s_1, s_2 ...\}$ is given by:

$$s_t = a \cdot X_t + (1 - a) \cdot s_{t-1}, t > 0$$

where X_t is the current observation time series and s_{t-1} the previous smooth statistic.

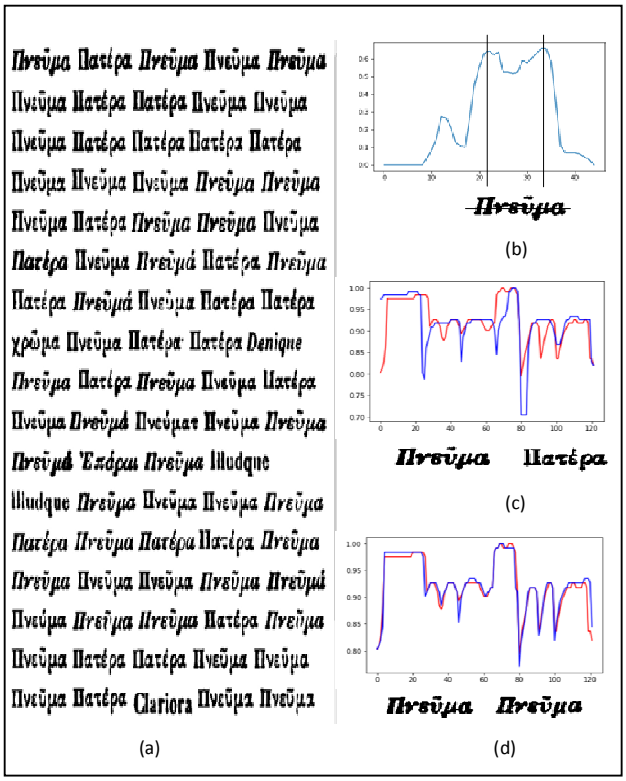

Figure 8. Third stage of similarity process.

V. Results and Discussion

The evaluation of the system is based on 46 pages of the work of Didymus Alexandrinus residing in the 39th volume of the PG as it is already mentioned in Section 4.1 because the specific PG pages have been converted by the TLG into editable text and serve as ground truth data. The raw data of TLG can be found in [20]. However, a direct evaluation of the produced word-spotting results is not an easy task and cannot be achieved by direct comparison with the texts of TLG. This happens because the TLG texts do not contain the

commentary sections of the PG scanned pages while on the other hand our spotting framework additionally finds the words that reside on the commentary page sections, resulting in a diversion as far as the population of each word is concerned. To solve this inconsistency we cropped manually the commentary sections from these 46 pages. The framework is evaluated for a set of seventeen words that carry significant semantic weight and additional one monosyllable conjunction word "γὰρ" (because) without semantic weight. This word however, has high occurrence due to the connectiveness that provides between two different sentences. This set of the chosen words is shown in Table 2 accompanied by the number of occurrences found in the TLG texts. Right next to them we present two popular evaluation measures, the Precision and Recall rates achieved at the end of the second stage (after applying SC) and after the third stage (after applying PCC). The results of the first stage are omitted because the main purpose of this stage acts only as pruning the very dissimilar images. Additionally, we present in the same table the Recall rates obtained by applying the Tesseract LSTM NN OCR engine [14] (a kind of Recurrent Neural Network (RNN)) in the place of PCC of the third stage, for a comparative display. The Tesseract uses the ancient Greek trained language data "grc". Also both PCC and Tesseract have as input the same segmentation words. It means that Tesseract does not apply any page or line segmentation but only the character segmentation of the already segmented word from our framework. The Precision and Recall rates are given as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Where TP is the number of true positive word recognitions, TN is the number of true negative word recognitions, FP is the false positive words recognitions and FN is the false negative word recognitions.

The set of the chosen words varies in the number of characters in order to study the behavior of our system for short, medium and long words. It is clearly shown in Table 2 that as the number of characters in the word increases the discrimination becomes stronger after the end of the second stage. This happens because when a long word is transformed with our CST method then it is easily distinguished by the SC descriptor due to the fact that it carries more variations to its shape. As it is shown in Table 2 after the 1st and 2nd stage the long words retrieved have a Precision percentage 100%.

It is also important to point out that CST accompanied by SC is very sensitive in discriminating the Latin from the Greek words. The retrieved words during the first two stages contain a very limited number of Latin words. One exception to this rule occurs when the query image represents a short word with two-three characters and in rare cases with more than three. However, these words in most cases do not carry any significant semantic value, they are mostly conjunctions or pronouns and generally undesired for word spotting. Two such exceptions are the words "Θεὸς" (God) and "Δαυὶδ" (David), however at the third similarity stage the dissimilar images are easily distinguished. In general, traditional OCR systems present difficulties in distinguishing the multiple languages in the scanned pages. Applying Tesseract on the whole PG page produces Greek text mixed with garbage text from the Latin script. Moreover, between the columns exist letters that further complicates the segmentation and recognition results. Consequently much effort is required and significant financial cost to distinguish the useful text from the garbage letters. During our experiments we also applied the framework on the complete set of PG pages without excluding the

commentary sections. Our system is able to capture the words in commentary sections that have a smaller font size compared with that of the main text, as well as different kind of font. The same word in the main text is printed with italic font while in the commentary section printed in normal font and vice versa.

The proposed techniques and methods under discussion applied with the use of the OpenCV 3.4 library i.e. HU's moments, SC descriptor and Contour extraction are implemented in specific OpenCV functions [30]. In the first similarity stage the threshold of HU's moments shape matching function is kept at 0.3 while in the second similarity stage the threshold for the SC descriptor is kept at 0.1. These values are fixed for all the query images used and during the whole evaluation. The selected values were chosen because they presented generally a good behavior and balance such as to disregard as dissimilar only a very small percentage of true positive images while on the other hand not to be computationally time consuming, especially for the SC descriptor. The system beyond the similar words that retrieves, also retrieves the slight variations of the query image mostly the variations in "grammatical case" which is expressed most of the times by the change of the last word character. For example the system captures the words "Παῦλον", "Παύλου" which are also desired instances of the word "Παῦλος" and can be accepted as true positive word instances, however we accepted them as false instances for the sake of the evaluation. In particular, the words "Παῦλον" and "Παύλου" are not easy to be distinguished even in the third stage because the PCC is insensitive to very small variations (nearly same shape of last character or different diacritic) so as to have a value near to one. Tesseract OCR on the other hand, since it applies character recognition is able to distinguish these words in the final verification. However for word spotting it is questionable if such discrimination it needed.

One other important issue that appears in Table 2 is that Tesseract and PCC discrimination shows a diversion for long words. PCC shows a stronger matching (number of successful recognized words and Recall rate) while Tesseract results are not as good. This happens because Tesseract is a character recognition system and as the number of characters in the word increase the word recognition error also increases. For smaller words PCC and Tesseract produce nearly similar Recall rates.

VI. Conclusion

In this study a framework for word spotting on the Greek polytonic script of PG has been presented and evaluation results are given. Specifically the Recall and Precision rates of the methodology are illustrated for a set of representative words that carry semantic weight and vary in the number of characters. The whole system consists of a segmentation and similarity parts subdivided in two and three stages respectively. The first segmentation stage is the initial creation of the word's bounding box while in the second stage we eliminate the punctuation marks, join properly the diacritics and create the final bounding boxes. The three stages of the similarity process are HU's moments and SC for pruning the very dissimilar images, and PCC for the final verification. The segmentation part achieves a high percentage of correct word segmentation nearly 96% .The average precision and recall rates for the examined words are 62,0% and 89,3% respectively after applying CST with the SC descriptor and 94,5 and 95,3% after applying PCC. The results obtained by PCC are also compared with those produced by applying the Tesseract LSTM NN OCR engine and show that our method outperforms slightly Tesseract mainly for the long words when Tesseract has as input the produced word-images of our segmentation method. A strong point of our framework is the efficiency in distinguishing the Greek words in the two-language scanned pages of PG. Additionally our system can be used as pre-stage for automatically creating candidate true or false labels of word images in order to train NNs or common classifiers. Moreover our system can be easily applied to other printed collections with Greek polytonic script since is

insensitive to specific fonts. This can be achieved by tuning accordingly the main system's parameters such as the horizontal kernel, the circle kernel, the HU's moments shape matching function threshold, SC threshold and the exponential smoothing factor.

Table 2 Precision and Recall rates.

Query words	Pearson's Decision Threshold of Upper or Lower Profile	1st Stage and 2nd Stage Similarity					3rd Stage Similarity					Tesseract LSTM engine	
		Set of 25703 segmented words			Precision	Recall	Set of Z words			Precision	Recall	Set of Z words	Recall
		TP	Z=TP+FP	TP+FN			TP'	TP'+FP'	TP'+FN'			TP''	
γὰρ	**0.60**	133	258	140	51.6	95.0	126	139	133	90.6	94.7	100	75.2
ἅγιον	**0.60**	19	65	20	29.2	95.0	18	25	19	72.0	94.7	16	84.2
Θεὸς	**0.60**	76	800	98	9.5	77.6	70	95	76	73.7	92.1	73	96.1
Ἰησοῦ	**0.60**	15	21	16	71.4	93.8	12	15	15	80.0	80.0	14	93.3
Δαυῒδ	**0.70**	11	881	11	1.2	100.0	11	11	11	100.0	100.0	8	72.7
κύριος	**0.50**	58	83	58	69.9	100.0	53	57	58	93.0	91.4	52	89.7
πνεῦμα	**0.50**	60	75	60	80.0	100.0	60	60	60	100.0	100.0	49	81.7
Παῦλος	**0.50**	26	29	31	89.7	83.9	25	26	26	96.2	96.2	26	100.0
Χριστοῦ	**0.50**	22	26	24	84.6	91.7	20	21	22	95.2	90.9	21	95.5
ἀγένητος	**0.50**	4	16	4	25.0	100.0	4	4	4	100.0	100.0	3	75.0
μονογενῆ	**0.70**	4	4	6	100.0	66.7	4	4	4	100.0	100.0	2	50.0
Κορινθίοις	**0.50**	15	19	19	78.9	78.9	13	13	15	100.0	86.7	15	100.0
ὑποστάσεως	**0.60**	9	28	14	32.1	64.3	8	8	9	100.0	88.9	9	100.0
ἀκατάληπτος	**0.70**	3	7	5	42.9	60.0	3	3	3	100.0	100.0	0	0.0
παντοκράτωρ	**0.50**	7	7	7	100.0	100.0	7	7	7	100.0	100.0	2	28.6
ἀπαυγάσματος	**0.60**	3	3	3	100.0	100.0	3	3	3	100.0	100.0	3	100.0
παντοκράτορος	**0.60**	2	4	2	50.0	100.0	2	2	2	100.0	100.0	1	50.0
ἐξομολογήσεται	**0.60**	2	2	2	100.0	100.0	2	2	2	100.0	100.0	2	100.0

References

[1] Akm Ashiquzzaman and Abdul Kawsar Tushar. 2017. Handwritten Arabic numeral recognition using deep learning neural networks. In 2017 IEEE International Conference on Imaging, Vision and Pattern Recognition, icIVPR 2017, 1–4. DOI:https://doi.org/10.1109/ICIVPR.2017.7890866

[2] Cagri Asian and Sibel Tari. 2005. An axis-based representation for recognition. In Proceedings of the IEEE International Conference on Computer Vision(ICCV), 1339–1346. DOI:https://doi.org/10.1109/ICCV.2005.32

[3] Xiang Bai, Longin Jan Latecki, and Wen Yu Liu. 2007. Skeleton pruning by contour partitioning with discrete curve evolution. IEEE Trans. Pattern Anal. Mach. Intell. 29, 3 (2007), 449–462. DOI:https://doi.org/10.1109/TPAMI.2007.59

[4] S. Belongie, Jitendra Malik, and J. Puzicha. 2002. Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. 24, 4 (2002), 509–522. DOI:https://doi.org/10.1109/34.993558

[5] C. Papadimitriou and K. Steiglitz. 1982. Combinatorial Optimization: Algorithms and Complexity. Prentice-Hall.

[6] Houssem Chatbri, Keisuke Kameyama, and Paul Kwan. 2016. A comparative study using contours and skeletons as shape representations for binary image matching. Pattern Recognit. Lett. 76, 1 (2016), 59–66. DOI:https://doi.org/10.1016/j.patrec.2015.04.007

[7] Didimus Alexandrinus. 1866. De Trinitate Libri Tres, col. 270. In Patrologia Graeca vol. 39. Migne.

[8] Aykut Erdem and Sibel Tari. 2010. A similarity-based approach for shape classification using Aslan skeletons. Pattern Recognit. Lett. 31, 13 (2010), 2024–2032. DOI:https://doi.org/10.1016/j.patrec.2010.06.003

[9] G. Horrocks. 2009. Greek: A History of the Language and its Speakers. John Wiley & Sons.

[10] B. Gatos, G. Louloudis, and N. Stamatopoulos. 2011. Greek polytonic OCR based on efficient character class number reduction. In Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 1155–1159. DOI:https://doi.org/10.1109/ICDAR.2011.233

[11] B. Gatos and I. Pratikakis. 2009. Segmentation-free word spotting in historical printed documents. In Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 271–275. DOI:https://doi.org/10.1109/ICDAR.2009.236

[12] Angelos P. Giotis, Giorgos Sfikas, Basilis Gatos, and Christophoros Nikou. 2017. A survey of document image word spotting techniques. Pattern Recognit. 68, (2017), 310–332. DOI:https://doi.org/10.1016/j.patcog.2017.02.023

[13] Angelos P. Giotis, Giorgos Sfikas, Christophoros Nikou, and Basilis Gatos. 2015. Shape-based word spotting in handwritten document images. In Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 561–565. DOI:https://doi.org/10.1109/ICDAR.2015.7333824

[14] Inc GitHub. 2017. Tesseract-OCR. tesseract-ocr/tesseract.

[15] Ming Kuei Hu. 1962. Visual Pattern Recognition by Moment Invariants. IRE Trans. Inf. Theory 8, 2 (1962), 179–187. DOI:https://doi.org/10.1109/TIT.1962.1057692

[16] Zhihu Huang and Jinsong Leng. 2010. Analysis of moment invariants on image scaling and rotation. In Innovations in Computing Sciences and Software Engineering, V7-476–480. DOI:https://doi.org/10.1007/978-90-481-9112-3-70

[17] Vassilis Katsouros, Vassilis Papavassiliou, Fotini Simistira, and Basilis Gatos. 2016. Recognition of Greek Polytonic on Historical Degraded Texts Using HMMs. In Proceedings - 12th IAPR International Workshop on Document Analysis Systems, DAS 2016, 346–351. DOI:https://doi.org/10.1109/DAS.2016.60

[18] Amandeep Kaur, Seema Baghla, and Sunil Kumar. 2015. Study of various character segmentation techniques for handwritten off-line cursive words: A review. Int. J. Adv. Sci. Eng. Technol. 3, 3 (2015), 154–158.

[19] Mariem Gargouri Kchaou, Slim Kanoun, and Jean Marc Ogier. 2012. Segmentation and word spotting methods for printed and handwritten Arabic texts: A comparative study. In Proceedings - International Workshop on Frontiers in Handwriting Recognition, IWFHR, 274–279. DOI:https://doi.org/10.1109/ICFHR.2012.266

[20] Khazarzar. 2000. Ruslan Khazarzar Library, Patrologia Section. Retrieved May 20, 2019 from http://khazarzar.skeptik.net/pgm/PG_Migne/.

[21] A. Kianimajd, M. G. Ruano, P. Carvalho, J. Henriques, T. Rocha, S. Paredes, and A. E. Ruano. 2017. Comparison of different methods of measuring similarity in physiologic time series. IFAC-PapersOnLine 50, 1 (2017), 11005–11010.

DOI:https://doi.org/10.1016/j.ifacol.2017.08.2479

[22] T. Konidaris, B. Gatos, K. Ntzios, I. Pratikakis, S. Theodoridis, and S. J. Perantonis. 2007. Keyword-guided word spotting in historical printed documents using synthetic data and user feedback. Int. J. Doc. Anal. Recognit. 9, 2–4 (2007), 166–177. DOI:https://doi.org/10.1007/s10032-007-0042-4

[23] Thomas Konidaris, Anastasios L. Kesidis, and Basilis Gatos. 2016. A segmentation-free word spotting method for historical printed documents. Pattern Anal. Appl. 19, 4 (2016), 963–976. DOI:https://doi.org/10.1007/s10044-015-0476-0

[24] Stelios Krinidis and Vassilios Chatzis. 2009. A skeleton family generator via physics-based deformable models. IEEE Trans. Image Process. 18, 1 (2009), 1–11. DOI:https://doi.org/10.1109/TIP.2008.2007351

[25] Vitaliy Kurlin. 2015. A homologically persistent skeleton is a fast and robust descriptor of interest points in 2d images. In Computer Analysis of Images and Patterns (CAIP), 606–617. DOI:https://doi.org/10.1007/978-3-319-23192-1_51

[26] Laurence Likforman-Sulem, Abderrazak Zahour, and Bruno Taconet. 2007. Text line segmentation of historical documents: A survey. Int. J. Doc. Anal. Recognit. 9, 2–4 (2007), 123–138. DOI:https://doi.org/10.1007/s10032-006-0023-z

[27] Haibin Ling and David W. Jacobs. 2007. Shape classification using the inner-distance. IEEE Trans. Pattern Anal. Mach. Intell. 29, 2 (2007), 286–299. DOI:https://doi.org/10.1109/TPAMI.2007.41

[28] Honey Mehta, Sanjay Singla, and Aarti Mahajan. 2016. Optical character recognition (OCR) system for Roman script & English language using Artificial Neural Network (ANN) classifier. In International Conference on Research Advances in Integrated Navigation Systems, RAINS 2016, 1–5. DOI:https://doi.org/10.1109/RAINS.2016.7764379

[29] Abirami Murugappan, Baskaran Ramachandran, and P. Dhavachelvan. 2011. A survey of keyword spotting techniques for printed document images. Artif. Intell. Rev. 32, 2 (2011), 119–136. DOI:https://doi.org/10.1007/s10462-010-9187-5

[30] OpenCv. 2014. OpenCV Library. OpenCV Website. Retrieved May 10, 2019 from https://docs.opencv.org/3.4.5/modules.html

[31] Ronald Poppe and Mannes Poel. 2006. Comparison of silhouette shape descriptors for example-based human pose recovery. In FGR 2006: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, 541–546. DOI:https://doi.org/10.1109/FGR.2006.32

[32] Qing Chen, Emil Petriu, and Xiaoli Yang. 2004. A comparative study of Fourier descriptors and Hu's seven moment invariants for image recognition. 103–106. DOI:https://doi.org/10.1109/ccece.2004.1344967

[33] Tony M. Rath and R. Manmatha. 2007. Word spotting for historical documents. Int. J. Doc. Anal. Recognit. 9, 2–4 (2007), 139–152. DOI:https://doi.org/10.1007/s10032-006-0027-8

[34] George Retsinas, Giorgos Sfikas, Georgios Louloudis, Nikolaos Stamatopoulos, and Basilis Gatos. 2018. Compact deep descriptors for keyword spotting. In Proceedings of International Conference on Frontiers in Handwriting Recognition, ICFHR. DOI:https://doi.org/10.1109/ICFHR-2018.2018.00062

[35] Israel Rios, Alceu S. Britto, Alessandro L. Koerich, and Luiz. E. S. Oliveira. 2010. Evaluation of different feature sets in an OCR free method for word spotting in printed documents. In Proceedings of the 2010 {ACM} Symposium on Applied Computing (SAC),Sierre, Switzerland, March 22-26, 2010, 52–56. DOI:https://doi.org/10.1145/1774088.1774099

[36] Bruce Robertson and Federico Boschetti. 2017. Large-Scale Optical Character Recognition of Ancient Greek. Mouseion 14, 3, LVIII–Series III (2017), 341–359. DOI:https://doi.org/http://dx.doi.org/10.3138/mous.14.3-3

[37] Bruce Robertson and Federico Boschetti. 2017. Large-Scale Optical Character Recognition of Ancient Greek. Mous. J. Class. Assoc. Canada, LVIII-Series III 14, 3 (2017), 341–359. DOI:https://doi.org/http://dx.doi.org/10.3138/mous.14.3-3

[38] Michael Schlemmer, Manuel Heringer, Florian Morr, Ingrid Hotz, Martin Hering Bertram, Christoph Garth, Wolfgang Kollmann, Bernd Hamann, and Hans Hagen. 2007. Moment invariants for the analysis of 2D flow fields. IEEE Trans. Vis. Comput. Graph. 13, 6 (2007), 1743–1750. DOI:https://doi.org/10.1109/TVCG.2007.70579

[39] Giorgos Sfikas, Angelos P. Giotis, Georgios Louloudis, and Basilis Gatos. 2015. Using attributes for word spotting and recognition in polytonic Greek documents. In Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 686–690. DOI:https://doi.org/10.1109/ICDAR.2015.7333849

[40] Yahya Sirin and M. Fatih Demirci. 2017. 2D and 3D shape retrieval using skeleton filling rate. Multimed. Tools Appl.

76, 6 (2017), 7823–7848. DOI:https://doi.org/10.1007/s11042-016-3422-2

[41] Anuj Srivastava, Shantanu H. Joshi, Washington Mio, and Xiuwen Liu. 2005. Statistical shape analysis: Clustering, learning, and testing. IEEE Trans. Pattern Anal. Mach. Intell. 27, 4 (2005), 590–602. DOI:https://doi.org/10.1109/TPAMI.2005.86

[42] Shriansh Srivastava, J. Priyadarshini, Sachin Gopal, Sanchay Gupta, and Har Shobhit Dayal. 2019. Optical character recognition on bank cheques using 2D convolution neural network. In Advances in Intelligent Systems and Computing. 589–596. DOI:https://doi.org/10.1007/978-981-13-1822-1_55

[43] Satoshi Suzuki and Keiichi A. Be. 1985. Topological structural analysis of digitized binary images by border following. Comput. Vision, Graph. Image Process. 30, 1 (1985), 32–46. DOI:https://doi.org/10.1016/0734-189X(85)90016-7

[44] J. Ignacio Toledo, Manuel Carbonell, Alicia Fornés, and Josep Lladós. 2019. Information extraction from historical handwritten document images with a context-aware neural model. Pattern Recognit. 86, (2019), 27–36. DOI:https://doi.org/10.1016/j.patcog.2018.08.020

[45] S Vijayarani and A Sakila. 2016. A Survey on Word Spotting Techniques for Document Image Retrieval. Int. J. Eng. Appl. Sci. Technol. 1 1, 8 (2016), 38–42.

[46] Luc Vincent. 2007. Google book search: Document understanding on a massive scale. In Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 819–823. DOI:https://doi.org/10.1109/ICDAR.2007.4377029

[47] Aaron D. Ward and Ghassan Hamarneh. 2010. The groupwise medial axis transform for fuzzy skeletonization and pruning. IEEE Trans. Pattern Anal. Mach. Intell. 32, 6 (2010), 1084–1096. DOI:https://doi.org/10.1109/TPAMI.2009.81

[48] LeCun Yann, Cortes Corinna, and Burges Chris. 2018. MNIST handwritten digit database. New York University. DOI:https://doi.org/10.1016/j.biotechadv.2011.08.021.Secreted

[49] Nida M. Zaitoun and Musbah J. Aqel. 2015. Survey on Image Segmentation Techniques. Procedia Comput. Sci. Conf. Commun. Manag. Inf. Technol. 65, (2015), 797–806. DOI:https://doi.org/10.1016/j.procs.2015.09.027

[50] Hui Zhang, Jason E. Fritts, and Sally A. Goldman. 2008. Image segmentation evaluation: A survey of unsupervised methods. Comput. Vis. Image Underst. 110, 2 (2008), 260–280. DOI:https://doi.org/10.1016/j.cviu.2007.08.003

[51] Perseus Project Homepage,. URL: http://www.perseus.tufts.edu/hopper/opensource.

[52] Levy, Google's two revolutions, Newsweek (Dec. 2004). URL http://www.newsweek.com/googles-two-revolutions-123507.

[53] TLG License Agreement, URL: http://www.tlg.uci.edu/subscriptions/site_license.php.