A Classification-free Word-Spotting System

Nikos Vassilopoulos[a] and Ergina Kavallieratou[a*]

[a] Dept. Information and Communication Systems Engineering
University of the Aegean
Samos, Greece

ABSTRACT

In this paper, a classification-free Word-Spotting system, appropriate for the retrieval of printed historical document images is proposed. The system skips many of the procedures of a common approach. It does not include segmentation, feature extraction or classification. Instead it treats the queries as compact shapes and uses image processing techniques in order to localize a query in the document images. Our system was tested on a historical document collection with many problems and a Google book, printed in 1675. Moreover, some comparative results are given for a traditional word spotting system.

Keywords: Historical Documents, Document Image Retrieval, Word-spotting system

1. INTRODUCTION

The Word Spotting procedure is inspired by speech processing and it was introduced in Document Image Processing in order to facilitate the information retrieval in cases that the perfect results of OCR cannot be achieved. Those can be the cases e.g. of documents that are degraded or they include languages or symbols too rare to warrant to worth OCR training. In the case of historical documents both can happen at the same time.

A classical Word Spotting methodology can include all or any of the procedures shown in Fig. 1. Many nice works have been proposed in the past for historical document retrieval, printed or handwritten, based on this common approach or parts of it[1-3].

However, all of the above mentioned works use the segmentation stage, mostly up to word level. In the cases that the quality of the paper, the ink or the scanning is not in a perfectly well condition, the segmentation procedure can reduce the success rate of Word Spotting. Thus, several free-segmentation Word Spotting approaches have also been proposed, lately. B. Gatos and I. Pratikakis[4] apply segmentation-free word spotting to printed Historical Documents by localizing salient areas and matching extracted features for several skews and scales. R. Farrahi Moghaddam and M. Cheriet[5] present, at the same conference, another line and word segmentation-free methodology, that is based on connected component feature extraction, DTW and Euclidean Distance. Y. Leydier et al.[6] also present a segmentation free word retrieval technique using zones of interest and guides on which they perform cohesive matching. Moreover, they allow the synthesis of the query. Finally, M. Rusiňol et al.[7] present a segmentation free word spotting technique, appropriate for heterogeneous document collections, using feature extraction on patch level.

In our case, we were called to develop a retrieval system for the Archive of the Government Gazette of the Principality of Samos, a Greek island, ex-autonomous regime under the suzerainty of the Ottoman Empire. At the General State Archives records (GSA) of Samos lies the complete set of copies of the Government Gazette of the Principality of Samos from the first year of the registration (1894) until the end of the Principality of Samos regime (1912). The Gazette was the official organ of the Administration of the Principality of Samos and therein were published laws, decrees, circulars, court actions and deeds like auctions. Apart from this official part, at that time, were also published the speeches of the liege lords, the minutes of the General Assemblies of Plenipotentiaries (i.e. the local parliament) and short reports on various topics. In total, there is one volume per year (19 volumes) and the amount of pages can vary from 250 to 750 pages per volume. Nowadays, this material is found in the GSA of Samos and this is the only existing full hard copy of this archive. Moreover, there is an already digitalized version that we were invited to use

[*] Further author information: (Send correspondence to E.K.)
N.V.: E-mail: nvasilopoulos@aegean.gr,
E.K.: E-mail: kavallieratou@aegean.gr, Telephone: +30 22730 82263

Document Recognition and Retrieval XX, edited by Richard Zanibbi, Bertrand Coüasnon,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 8658, 86580F · © 2013 SPIE-IS&T
CCC code: 0277-786/13/$18 · doi: 10.1117/12.2002975

in order to build a system that will perform automatic retrieval every time that a Samian citizen wishes to look for something in that archive. The bad quality of the scanned archive (fig.2) prohibited the use of the common approach of figure 1.

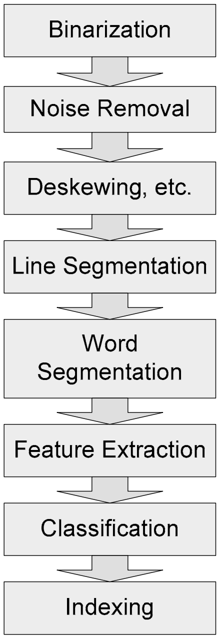


Figure 1. The modules of a traditional Word-Spotting approach.

The contribution of the present work consists of:

- The trial of a simplified word spotting system that is based mainly on picture processing techniques instead of pattern recognition, skipping segmentation and clustering and using the words as compact shapes.
- The creation of a small ground truth set of old Greek documents with common problems, available to the scientific community,
- A system that will make easier the research and study of the rare Archive of the Government Gazette of the Principality of Samos.

The proposed methodology is presented in section 2, while in section 3, retrieval results are presented for the above mentioned collection, as well as for a Google book in order to give more objective results. Some comparative results are also given for a traditional system, similar to the one presented in figure 1. Finally some conclusions are drawn in section 4.

2. THE PROPOSED METHODOLOGY

The above mentioned archive presents some special characteristics due to its scanning that it took place several years ago by non-specialists:

- the resolution is low, just 200 dpi,
- the pages, newspaper size, are scanned two at the time (fig.2),
- unevenly lighted image (fig.2),
- lightly skewed part of the image (fig.2),
- old printing of bad quality (fig.3),

- the language in use (fig.3) is an older version of Greek with a lot of accents, that are not used at the moment and it is difficult to find appropriate OCR software.

The bad quality of printing and scanning (fig.3) proved to be a very problematic situation during the application of the known techniques, even after trying to improve them. The low performance of each task, due to the special problems, was accumulated to the whole giving a lower final result. The necessity to keep the system as simple as possible with a minimum number of modules was soon realized.

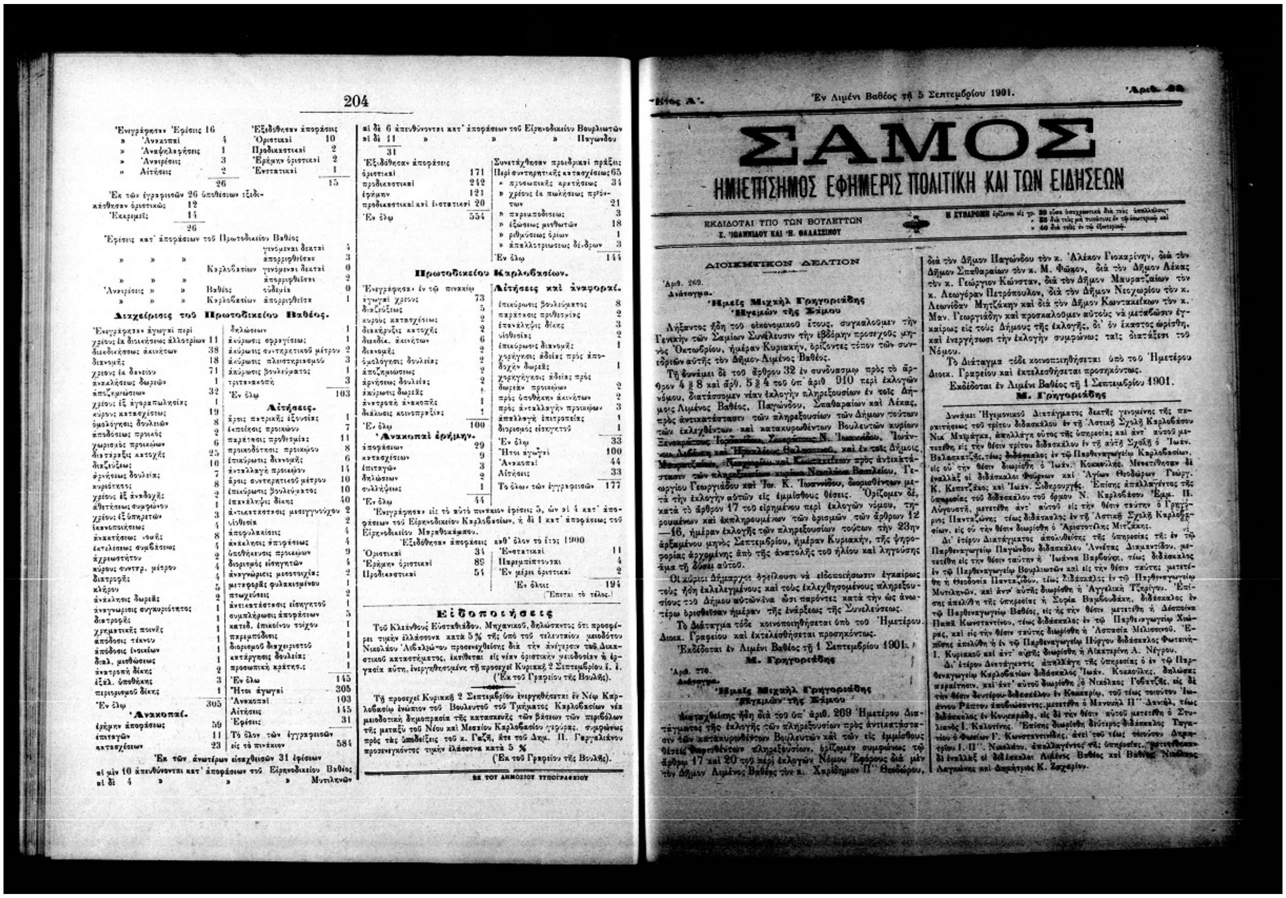

204
Ἔτος Α'.
Ἐν Λιμένι Βαθέος τῇ 5 Σεπτεμβρίου 1901.
ΣΑΜΟΣ
ΗΜΙΕΠΙΣΗΜΟΣ ΕΦΗΜΕΡΙΣ ΠΟΛΙΤΙΚΗ ΚΑΙ ΤΩΝ ΕΙΔΗΣΕΩΝ
ΕΚΔΙΔΟΤΑΙ ΥΠΟ ΤΩΝ ΒΟΥΛΕΥΤΩΝ
Σ. ΙΩΑΝΝΙΔΟΥ ΚΑΙ Η. ΘΑΛΑΣΣΙΝΟΥ
ΔΙΟΙΚΗΤΙΚΟΝ ΔΕΛΤΙΟΝ

Figure 2. Page from the Government Gazette of the Principality of Samos.

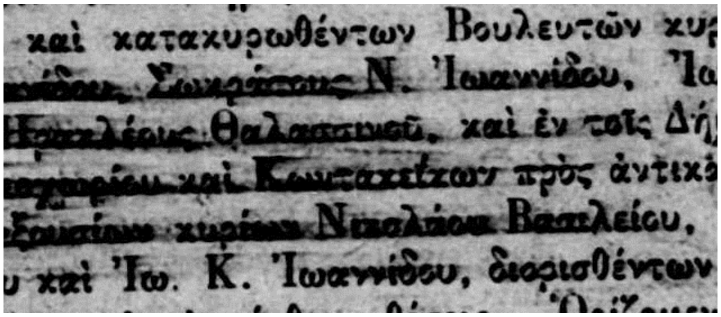

καὶ κατακυρωθέντων Βουλευτῶν
Σωκράτους Ν. Ἰωαννίδου,
Θαλασσινοῦ, καὶ ἐν τοῖς
καὶ Κωνσταντίνου πρὸς ἀντικα
κυρίων Νικολάου Βασιλείου,
καὶ Ἰω. Κ. Ἰωαννίδου, διορισθέντων

Figure 3. Detail from the archive.

The proposed system appears in the figure 4. It consists of simple procedures of image processing. First, adaptive thresholding is applied to the image using as threshold the mean of the 9x9 neighborhood. Next, the main body of the text, although no segmentation is performed, is estimated by the query, using the technique mentioned in Kavallieratou et al.[8]. Unfortunately this limits the retrieval in the words that fit the size of the query.

Instead of extracting feature vectors from the query and the document images, the whole query is kept and the document image is scanned to find the specific query, without applying any segmentation. In order to get rid of unnecessary details and smooth small differences in skew and scale, the query is transformed into more compact shape by normalization. The normalization consists of applying opening to the image with an elliptical structuring element of the size of [word main body] x [word main body*0.5]. The same normalization is applied to all document images. This normalization that takes about 5 secs / image 3300x4500 pixels (fig.2), it can be applied once to all document images and be kept stored for the future queries. Several examples of the normalization procedure are shown in fig.5.

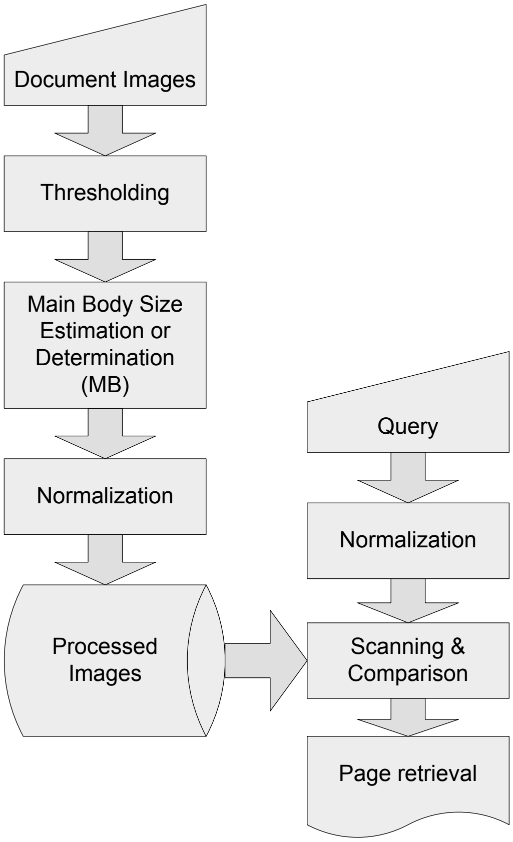


Figure 4. The proposed approach.

Each query could be selected by the user or uploaded by an image file. Although synthesis was also considered, as it is described in Y. Leydier et al.[6], the results were worse due to the bad quality of printing that results no standard relative position, in vertical direction and, the characters and the accents, that are not all present in the modern Greek. After the normalization, the query image is applied to every pixel of the image (left-upper corner) and is compared with the corresponding part of the image, using the Sum of Squared Differences (SSD) matching algorithm:

$$h[m,n] = \sum_{k,l} (q[k,l] - I[m+k, n+l])^2 ,$$

where *I* is the page image and *q* the query image.

In Sum of Squared Differences (SSD), the differences are squared and aggregated. Finally, if the similarity is large enough (please see §3) the corresponding page is retrieved.

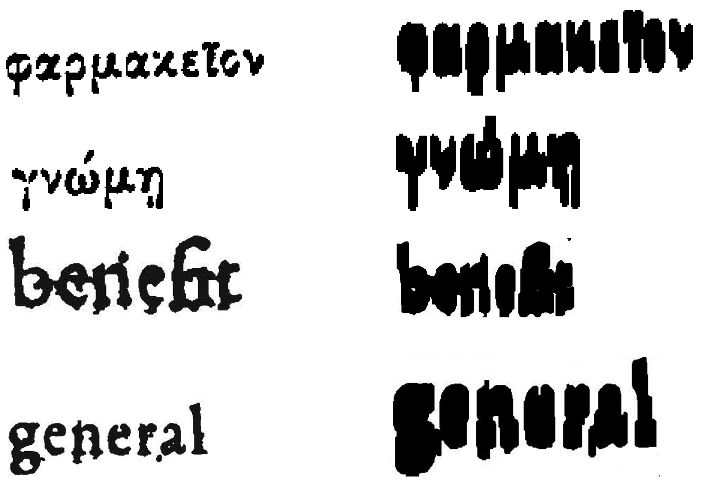


Figure 5. Examples of the normalization procedure.

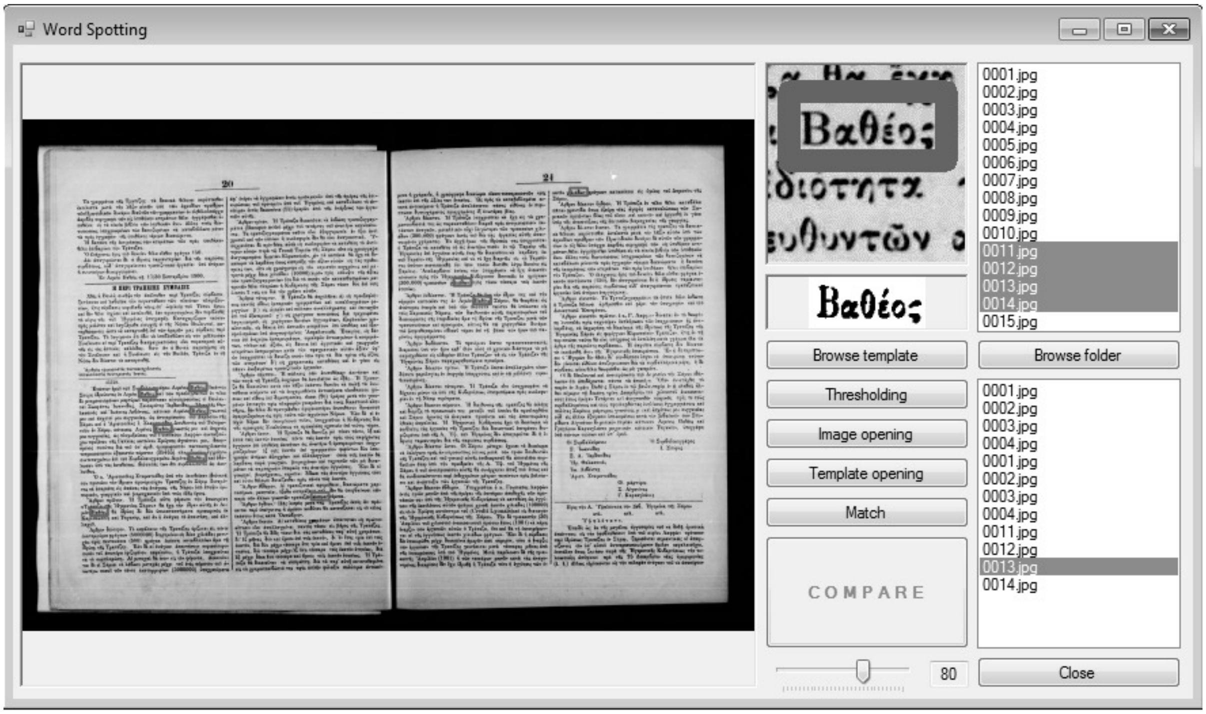
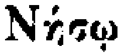
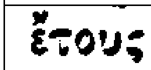
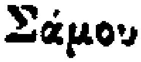
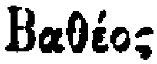
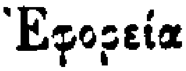
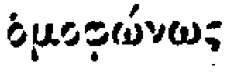
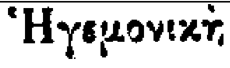
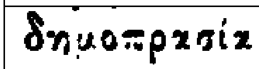
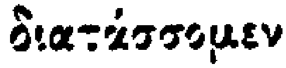
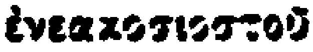


Figure 6. A screenshot of the implemented system.

Visual C# and OpenCV were used to implement the thresholding, the opening filter and the SSD matching algorithm. The system shown in figure 6 can look for a query (image of text), selected by the user or uploaded by file, in a collection or the selected images of it (upper list). The modules of the methodology, described above, can be applied separately or all together (compare button). The similarity accuracy, mentioned in section 3, can be selected by the user (slider fig.6) and the retrieved images will be presented in the lower list.

3. EXPERIMENTAL RESULTS

Table 1. The queries for the greek archive, their translation and occurrences in the ground truth sample.

Queries	**Translation**/note	occurences
 Νήσω	**Island**, in dative	10
 ἔτους	**Year** (gen), often met in dates	33
 Σάμου	the **name of the island** (gen)	58
 Βαθέος	the **name of the capital** (gen)	73
 Ἐφορεία	**Tax office**	5
 ὁμοφώνως	**unanimously** often met in decisions	16
 Ἡγεμονική	**Hegemonic**, reference to the island	10
 δημοπρασία	**auction**, often published	18
 διατάσσομεν	**we order**, often met in decisions	21
 ἐνεακοσιοστοῦ	**900th**, often met in dates	21

In order to perform experiments on the proposed system, ground truth results were extracted from 15 scanned images of the Greek archive, that is 30 document pages, by human reader, for 10 queries. Words of different sizes (in characters) were included. In table 1, the queries are shown, next to an explanatory note and the amount of times they occur in the ground truth sample.

Moreover, a google book, *A sermon preach'd before the king*[†], was also used in order to extract more objective results with OCR text provided by Google. The book consists of 44 pages and it was published in 1675. In this case we used the book binarised images as provided by Google. A sample image is shown in figure 7. The book, although carefully binarised, includes a lot of noise and an older version of alphabet symbols (fig.8).

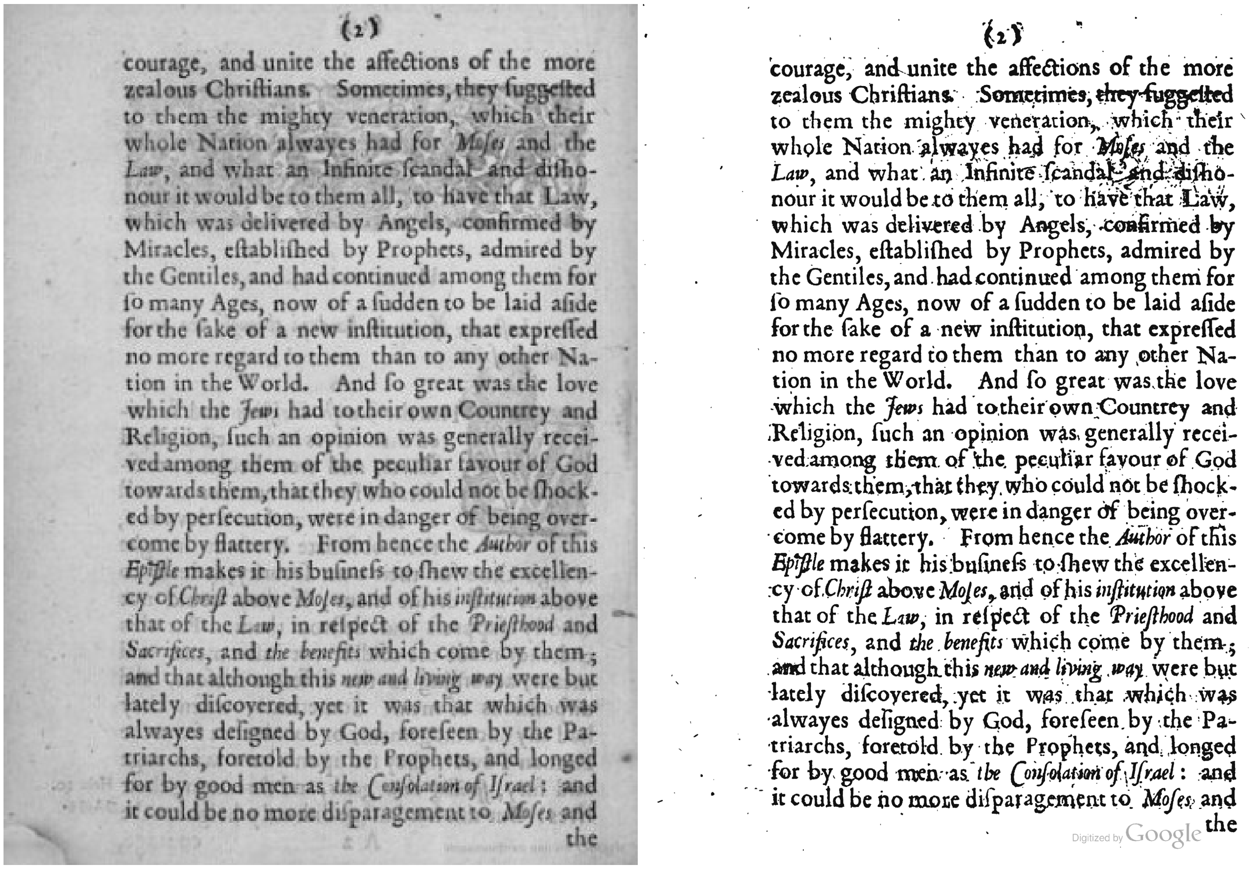
In order to evaluate the proposed system, precision (1), recall (2) and F measure (3) were used:

$$precision = \frac{correctly \quad retrieved \quad words}{total \quad retrieved \quad words} \qquad (1)$$

[†] http://books.google.gr/books?id=-SY3AAAAMAAJ&printsec=frontcover&dq=preach+king&hl=el&sa=X&ei=jIlQT6SkOcrP0QX79Mz0Cw&redir_esc=y#v=onepage&q=preach%20king&f=false

$$recall = \frac{correctly \quad retrieved \quad words}{existed \quad words} \tag{2}$$

$$F = \frac{2 * recall * precision}{recall + precision} \tag{3}$$

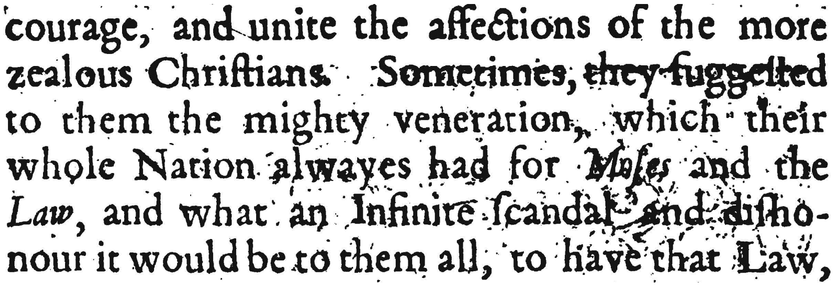


(2)

courage, and unite the affections of the more zealous Chriſtians. Sometimes, they ſuggeſted to them the mighty veneration, which their whole Nation alwayes had for *Moſes* and the *Law*, and what an Infinite ſcandal and diſhonour it would be to them all, to have that Law, which was delivered by Angels, confirmed by Miracles, eſtabliſhed by Prophets, admired by the Gentiles, and had continued among them for ſo many Ages, now of a ſudden to be laid aſide for the ſake of a new inſtitution, that expreſſed no more regard to them than to any other Nation in the World. And ſo great was the love which the *Jews* had to their own Countrey and Religion, ſuch an opinion was generally received among them of the peculiar favour of God towards them, that they who could not be ſhocked by perſecution, were in danger of being overcome by flattery. From hence the *Author* of this *Epiſtle* makes it his buſineſs to ſhew the excellency of *Chriſt* above *Moſes*, and of his *inſtitution* above that of the *Law*, in reſpect of the *Prieſthood* and *Sacrifices*, and *the benefits* which come by them; and that although this *new and living way* were but lately diſcovered, yet it was that which was alwayes deſigned by God, foreſeen by the Patriarchs, foretold by the Prophets, and longed for by good men as *the Conſolation of Iſrael*: and it could be no more diſparagement to *Moſes* and the

Figure 7. Sample page in original and binary from Google books.

The results for the Greek documents are presented in table 2, while the comparative results for the Google book can be seen in table 3. In the case of Google book, it should be mentioned, that from the list of occurrencies in Google typed text, the occurrences in italics have been excluded, while the occurrences that include the query have been added e.g. teachers for teacher, etc.

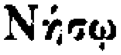
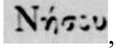
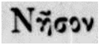
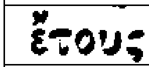
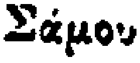
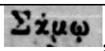
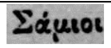
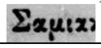
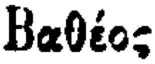
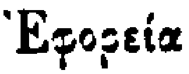
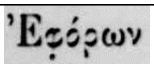
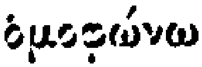
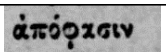
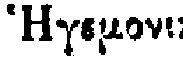
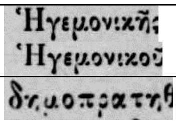
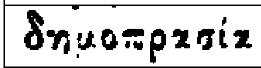
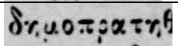
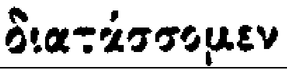
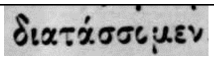
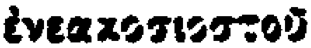


courage, and unite the affections of the more zealous Chriſtians. Sometimes, they ſuggeſted to them the mighty veneration, which their whole Nation alwayes had for *Moſes* and the *Law*, and what an Infinite ſcandal and diſhonour it would be to them all, to have that Law,

Figure 8. Detail from the Google book.

Since Greek is an inflectional language, for each noun several similar word forms can be found (dative, genitive, accusative, etc.). This is obvious in the results like Νήσῳ (dat.), Σάμου (gen), ηγεμονική (acc.), etc., however they did not considered correct, although it could be useful in many cases. This could be one of the reasons that the results in English text are higher, plus the fact that the Google documents are scanned in 600 dpi while the resolution of the Greek documents is just 200 dpi.

Table 2. Experimental Results for Greek documents.

Queries	CPU time/ page (sec)	Similarity > 85%			False Positives
		Precis.%	Rec.%	F meas.%	
	4.73	22.5	90	36	, 
	2.59	100	62.85	77.19	-
	11.23	85.71	41.37	55.81	,  
	5.04	100	30.13	46.31	-
	13.88	83.33	100	90.90	
	15.06	94.44	100	97.14	
	126.98	18.03	100	30.55	
	11.57	83.33	83.33	83.33	
	11.90	94.73	100	97.29	
	9.74	100	28.57	44.44	-

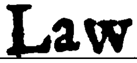
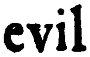
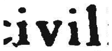
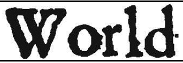
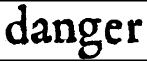
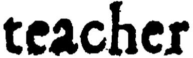
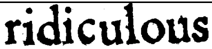
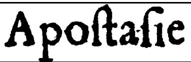
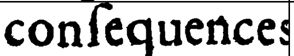
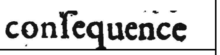
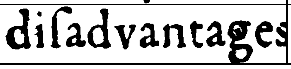
The computational cost shown in tables 2 & 3 under *CPU time/per page (sec)* concerns the system implemented as mentioned in §2. In our first experiments in Matlab, the computational cost could reach double or triple depending on the query size. In those cases an extra trick was introduced in order to reduce the computational cost. The image were scanned every 2-3 pixels, instead of every pixel, depending on the image resolution, in order to reduce the computational cost. However, this didn't were considered necessary in the implemented system (fig.6) since the computational cost is much lower. This solution is kept in mind for larger collections.

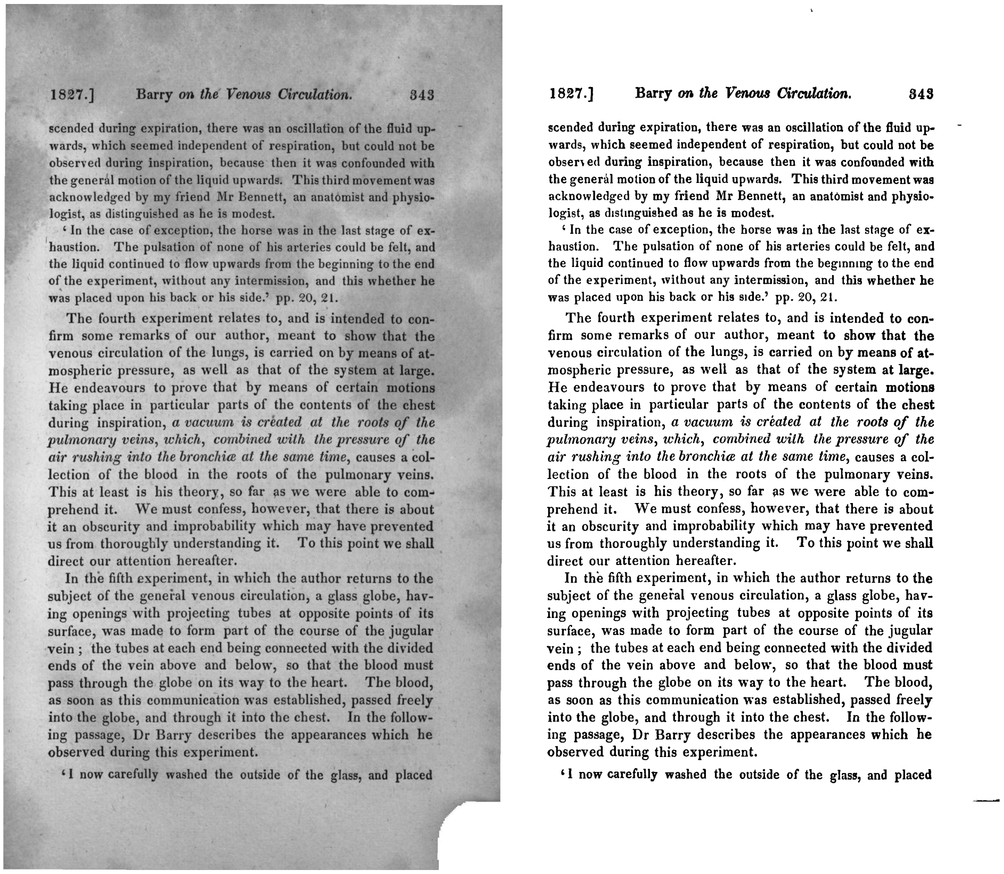
Finally, in order to give comparative results with a traditional Word-Spotting system, the system described in N. Doulgeri and E. Kavallieratou[9] was used. This system is very similar to the traditional ones[1-3], since it includes more of the stages described in fig.1. For our experiments, the parameters, as they were proved better in the paper, were used for that system. That is, synthesized words in 300 dpi, bold Times New Roman, interpolation of 175 points and smoothing of 5 points. Since the mentioned, in that paper, books were not available and the application of the traditional system to the ones mentioned here was impossible due to failure in segmentation, ten pages were used from the Google Book entitled *The Medico-chirurgical Review and Journal of Medical Science*[‡] that was published in London in 1826 (fig.9). Four samples were selected from the ones synthesized for the paper, and the pages were selected in order to have at least

[‡] http://books.google.gr/books?id=OhoUAAAAQAAJ&pg=PA489&lpg=PA489&dq=%22barry+on+the+venous+circulation%22&source=bl&ots=NZNHEOQ9ag&sig=X-9qRrYj0gUH1SlFdp15bzA7lgs&hl=el&sa=X&ei=9rygUJWaL7Cb1AXd_YHgAQ&ved=0CB8Q6AEwAA#v=onepage&q&f=false

two occurencies for each of them. Both systems were applied, as they are described at the corresponding papers. The results are presented in table 4.

Table 3. Experimental Results for Google book.

Queries	CPU time/per page (sec)	Similarity > 85%			Google OCR (occur.)	False positives
		Precision	Recall	F1		
	0.12	100	66.67	80	3	-
	5.56	88.89	100	94.11	8	
	9.81	100	66.67	80	24	-
	5.60	100	90	94.73	10	-
	6.17	100	100	100	2	-
	6.50	100	66.67	80	3	-
	7.90	100	66.67	80	3	-
	8.08	50	100	66.67	1	
	6.72	100	100	100	2	-



1827.] Barry *on the Venous Circulation.* 343

scended during expiration, there was an oscillation of the fluid upwards, which seemed independent of respiration, but could not be observed during inspiration, because then it was confounded with the general motion of the liquid upwards. This third movement was acknowledged by my friend Mr Bennett, an anatomist and physiologist, as distinguished as he is modest.

'In the case of exception, the horse was in the last stage of exhaustion. The pulsation of none of his arteries could be felt, and the liquid continued to flow upwards from the beginning to the end of the experiment, without any intermission, and this whether he was placed upon his back or his side.' pp. 20, 21.

The fourth experiment relates to, and is intended to confirm some remarks of our author, meant to show that the venous circulation of the lungs, is carried on by means of atmospheric pressure, as well as that of the system at large. He endeavours to prove that by means of certain motions taking place in particular parts of the contents of the chest during inspiration, *a vacuum is créated at the roots of the pulmonary veins, which, combined with the pressure of the air rushing into the bronchiæ at the same time*, causes a collection of the blood in the roots of the pulmonary veins. This at least is his theory, so far as we were able to comprehend it. We must confess, however, that there is about it an obscurity and improbability which may have prevented us from thoroughly understanding it. To this point we shall direct our attention hereafter.

In thė fifth experiment, in which the author returns to the subject of the general venous circulation, a glass globe, having openings with projecting tubes at opposite points of its surface, was made to form part of the course of the jugular vein; the tubes at each end being connected with the divided ends of the vein above and below, so that the blood must pass through the globe on its way to the heart. The blood, as soon as this communication was established, passed freely into the globe, and through it into the chest. In the following passage, Dr Barry describes the appearances which he observed during this experiment.

'I now carefully washed the outside of the glass, and placed

1827.] Barry *on the Venous Circulation.* 343

scended during expiration, there was an oscillation of the fluid upwards, which seemed independent of respiration, but could not be observed during inspiration, because then it was confounded with the general motion of the liquid upwards. This third movement was acknowledged by my friend Mr Bennett, an anatomist and physiologist, as distinguished as he is modest.

'In the case of exception, the horse was in the last stage of exhaustion. The pulsation of none of his arteries could be felt, and the liquid continued to flow upwards from the beginning to the end of the experiment, without any intermission, and this whether he was placed upon his back or his side.' pp. 20, 21.

The fourth experiment relates to, and is intended to confirm some remarks of our author, meant to show that the venous circulation of the lungs, is carried on by means of atmospheric pressure, as well as that of the system at large. He endeavours to prove that by means of certain motions taking place in particular parts of the contents of the chest during inspiration, *a vacuum is créated at the roots of the pulmonary veins, which, combined with the pressure of the air rushing into the bronchiæ at the same time*, causes a collection of the blood in the roots of the pulmonary veins. This at least is his theory, so far as we were able to comprehend it. We must confess, however, that there is about it an obscurity and improbability which may have prevented us from thoroughly understanding it. To this point we shall direct our attention hereafter.

In thė fifth experiment, in which the author returns to the subject of the general venous circulation, a glass globe, having openings with projecting tubes at opposite points of its surface, was made to form part of the course of the jugular vein; the tubes at each end being connected with the divided ends of the vein above and below, so that the blood must pass through the globe on its way to the heart. The blood, as soon as this communication was established, passed freely into the globe, and through it into the chest. In the following passage, Dr Barry describes the appearances which he observed during this experiment.

'I now carefully washed the outside of the glass, and placed

Figure 9. Sample page in original from Google books and binary from Doulgeri[9].

Table 4. Comparative Results on the Google book, for the proposed and the Doulgeri[9] systems.

Queries	Occurr.	Proposed System		Doulgeri[9]	
		true positive	false positive	true positive	false positive
close	3	2	1	3	1
difficult	2	2	2	2	0
English	3	3	0	2	2
habit	2	1	0	2	0

4. CONCLUSION

In this paper, a system of word spotting was presented and evaluated. The system proposes a simplified methodology that omits many tasks of the traditional word spotting approach. As preprocessing, it only requires a thresholding and it is a segmentation-free approach. Moreover, it does not include feature extraction and clustering or classification stages. The comparison and matching procedures are performed by image processing techniques.

The proposed system was applied to a collection of Greek document images of the Government Gazette of the Principality of Samos, that are kept at the General State Archives records (GSA) of Samos. Moreover, it was also applied to a Google book of the 17th century, in order to compare results based on the OCR text provided by Google. Finally, some examples are presented for the same queries from the proposed system and an older one, that includes segmentation and classification.

The results seem promising and we plan to research further with our technique, applying it even to handwritten historical document images. Moreover, we wish to improve the synthesis procedure. Another problem we wish to deal with is the fitting of the query for different size fonts.

REFERENCES

[1] T. Rath and R. Manmatha, "Word spotting for historical documents" *International Journal of Document Analysis and Recognition*, Vol. 9, No. 2-4, pp.139–152,(2007).

[2] T. Konidaris, B. Gatos, K. Ntzios, I. Pratikakis, S. Theodoridis, and J. Perantonis, "Keyword-guided Word Spotting in historical printed documents using synthetic data and user feedback" *International Journal of Document Analysis and Recognition*, Vol. 9, pp.167-177, (2007).

[3] H. Cao, A. Bhardwaj, V. Govindaraju, "A probabilistic method for keyword retrieval in handwritten document images", *Pattern Recognition*, Vol.42, No 12, pp.3374-3382, (2009).

[4] B. Gatos and I. Pratikakis, "Segmentation-free word spotting in historical printed documents" *Proc. of the Int. Conf. on Document Analysis and Recognition*, pp. 271–275, (2009).

[5] R. Farrahi Moghaddam. and M. Cheriet, " Application of multi-level classifiers and clustering for automatic word-spotting in historical document images", Proc. *of the Int. Conf. on Document Analysis and Recognition*, pp. 511–515, (2009).

[6] Y. Leydier, A. Ouji, F. LeBourgeois, and H. Emptoz, "Towards an omnilingual word retrieval system for ancient manuscripts", *Pattern Recognition*, Vol. 42, No. 9, pp. 2089–2105, (2009).

[7] M. Rusiňol, D. Aldavert, R. Toledo, J. Llados, "Browsing Heterogeneous Document Collections by a Segmentation-Free Word Spotting Method", *International Conference on Document Analysis and Recognition (ICDAR),* pp.63-67, (2011).

[8] E.Kavallieratou, N.Fakotakis, and G.Kokkinakis, "Un Off-line Unconstrained Handwritting Recognition System", *International Journal of Document Analysis and Recognition*, no 4, pp. 226-242, (2002).

[9] N. Doulgeri and E. Kavallieratou, "Retrieval of historical documents by word spotting" *Proceedings of SPIE, Volume 7247, Retrieval and Text Categorization*, pp.06, (2009).