A System for Processing and Recognition of Old Greek Manuscripts (The D-SCRIBE Project)

S.J. PERANTONIS[1], B. GATOS[1], K. NTZIOS[1], I. PRATIKAKIS[1], I. VRETTAROS[2],
A. DRIGAS[2], C. EMMANOUILIDIS[3], A. KESIDIS[4], D. KALOMIRAKIS[5]

[1]Computational Intelligence Laboratory, Institute of Informatics and Telecommunications
National Research Center "Demokritos", 153 10 Athens
GREECE
sper@iit.demokritos.gr   http://www.iit.demokritos.gr/cil/

[2] Net Media Laboratory
National Research Center "Demokritos", 153 10 Athens
GREECE
dr@imm.demokritos.gr   http://imm.demokritos.gr/

[3] ZENON S.A., Automation Technologies, R&D Division
Kanari 5, Glyka Nera Attikis 153 54, Athens
GREECE
christosem@zenon.gr   http://www.zenon.gr/

[4] BSI S.A., R&D Division
17 Noembriou 130, Xolargos 155 62, Athens
GREECE
akes@bsi.gr   http://www.bsi.gr/

[5] Mount Sinai Foundation
Doryleou 26, Athens 121 15
GREECE
kalomirakis@hotmail.com   http://www.sinaimonastery.com/

Abstract: - After many years of scholar study, manuscript collections continue to be an important source of novel information for scholars, concerning both the history of earlier times as well as the development of cultural documentation over the centuries. In this paper we present research efforts leading to the creation of a comprehensive software product, which can assist the content holders in turning an archive of manuscripts into a digital collection using automated methods. The system aims at supporting and facilitating current and future efforts in Early Christian Greek manuscript digitization and processing. A number of modules have been developed for the study and processing of the digital collections and are integrated under a common software environment. These modules include a document management system particularly adapted for manuscripts, a module for automatic processing and transliteration of manuscripts incorporating OCR techniques and a self study tool for the use of paleographers. In this paper, we focus on the description of the aforementioned modules that comprise the final system for processing and recognition of old Greek Manuscripts.

Key-Words: - Handwriting Recognition, Document Image Binarization, Segmentation-free, Feature Extraction, Historical Document Recognition, Old Manuscript Recognition, Self Study tool

1 Introduction

The advent of information technologies presents us with unique opportunities for the digital preservation and advancement of important treasures of our rich cultural heritage. A complete strategy for the exploitation of these opportunities should involve two key actions: Firstly, conversion of cultural heritage content to electronic form via digitization; and secondly, development of innovative information technology products that will assist scholars in the study of this content and illustrate its importance to all interested parties, including the general public.

In this paper we present research aiming at the creation of a comprehensive software product, which can assist the content holders in turning an archive of manuscripts into a digital collection using automated methods. This software is being developed in the framework of project D-SCRIBE, a Greek GSRT-funded R&D project which aims to support and facilitate current and future efforts in old Greek manuscript digitization and processing (http://iit.demokritos.gr/cil/dscribe/). Our final product will give various memory institutions the opportunity to:

- Digitize their manuscript collections according to quality metrics, leveraging existing material with state-of-the-art technical feasibility.
- Produce varying digital objects for varying purposes, e.g. access vs. preservation.
- Automate the transliteration of manuscripts, by employing manuscript-tuned OCR modules.
- Manage their content in the form of a digital library, by using a powerful document management system.
- Facilitate and expand the study of paleography, by providing self-study tools which will help students and researchers in coping with large volumes of data.

An immediate objective of the project is the digital preservation of a large number of important historical manuscripts of the early Christian and Byzantine era from St. Catherine's monastery, an outpost of the Hellenic world. Beyond this immediate goal, the product target includes an extensive number of organizations and companies related with the management of valuable manuscripts like monasteries, institutions, libraries, private collections etc, in Greece and other countries. Therefore, the D-SCRIBE software is expected to play a key role in the digital preservation, processing and study of old Greek manuscripts, thus contributing to the preservation and advancement of our cultural heritage.

Expected overall benefits include:

- The digitization and digital preservation of one of the largest and most valuable collections of early Christian manuscripts at Mt. Sinai monastery which will allow for easier and more comprehensive access by the scholars.
- Our effort is expected to advance the state of the art in the use of computer technologies in humanities, particularly in the sector of manuscript studies.
- The technology for integrated manuscript analysis produced in D-SCRIBE is of major significance, promising to assist in bridging the access gap between traditional experts and modern high-tech generations. Such a development can have a serious impact on economic issues, by introducing new models of work and cooperation as well as new tailored services that go beyond the "given" role of cultural institutions, therefore services that can be charged.
- Information technology companies that adopt the added value approach offered by new technology in the cultural informatics market – and more specifically in the market concerning access to valuable manuscripts - will clearly have a head start in this direction and will be able to capitalize their competitive advantage, by claiming a larger share of this emerging market.
- The emphasis given to the customizability of the tools and methods developed will help in showcasing technology that is readily deployable and transferable to similar applications.

The software comprises a number of modular tools. These tools include a document management system particularly adapted for manuscripts, a module for automatic processing and transliteration of manuscripts incorporating OCR techniques and a self study tool for the use of paleographers. In this paper, we focus on the description of the aforementioned modules that comprise the final system for processing and recognition of old Greek Manuscripts.

2 Document Digitization

In the D-SCRIBE project we aim to digitize a large part of the manuscript collection at the Mount Sinai Monastery. We will achieve this goal by selecting and digitizing samples from different eras. There are mainly two types of Greek manuscripts to be analyzed in the framework of this project.

- The first type concerns Greek manuscripts written in capitals which date back to the paleochristian period. The Sinaitic Codex, which contains the Bible, forms the basis of our digitization effort.
- The second type concerns Greek manuscripts written in lower case letters. Since Greek lower case writing has acquired several kinds of standardization as well as an extremely high variation through the ages, the basis for the digitization is formed by manuscripts coming from the group of the imperial laboratories of Constantinople which are linked to the Abbey of Studios. The Sinaitic Codex Number Three, which contains the Book of Job, as well as other relative manuscripts constitutes the main axis of this research.

It is expected that by the end of the project, 150.000 individual pages will have been digitized. For the digitization task, we use two 1200 dpi scanners and two digital photographic cameras: an RGB SINAR capable of 2000x3000 pixel resolution at 14bit and a grayscale SINAR capable of 3000x4000 pixel resolution at 12bit.

3 Document Management System

A document management subsystem has been developed based on a Client/Server architecture in order to handle digital input either digitized at a previous unknown time or through the relevant but independent components of the overall system (e.g. scanner drivers). For the purposes of this task, the development has been based on the document management technology already developed by D-SCRIBE partners ("eDoc" Document Management Software - www.edoplus.com). We provide the user with an intuitive interface allowing him to construct digital document collections and assign them meaningful metadata. We also provide the user with all the common document management facilities such as search and full text retrieval capabilities, fast document previews, printing, varying levels of access permission, backup facilities, annotations on document images etc. A demonstration of the document management system is given in Fig. 1.

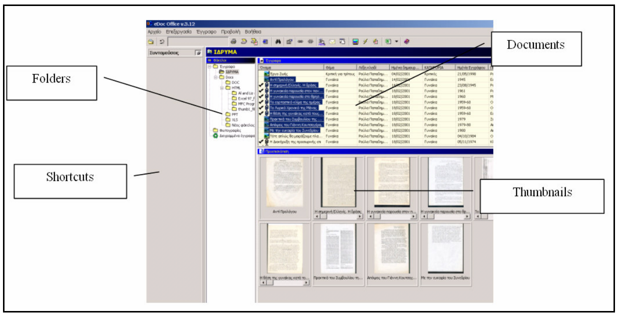


Figure 1. The Document Management workspace.

4 Processing and Recognition Module

In this paper, we focus on the problem of recognizing early Christian Greek manuscripts (see, e.g. Fig. 2). The processing and recognition module consists of several sub-modules that will be described in this section.

4.1 Binarization and Enhancement

Binarization is the starting step of most document image analysis systems and refers to the conversion of the gray-scale image to a binary image. Since historical document collections are most of the times of very low quality, an image enhancement stage is also essential.

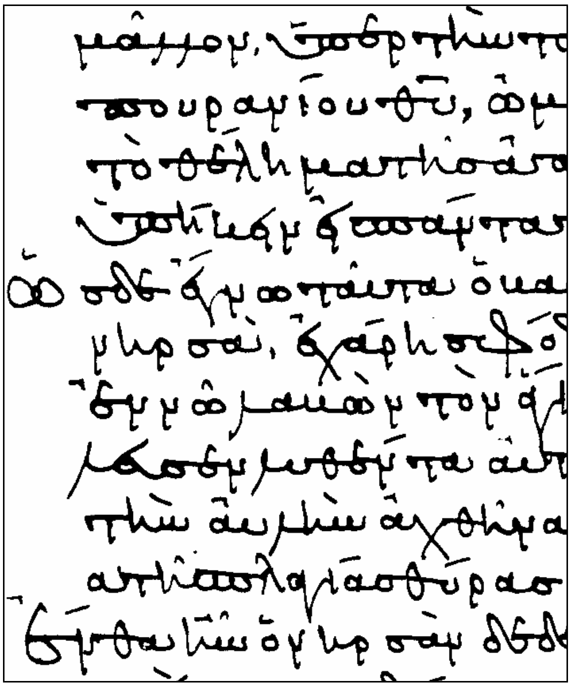


Figure 2. Early Christian Greek manuscript.

In the literature, binarization is usually reported to be performed either globally or locally. The global methods (global thresholding) use a single threshold value to classify image pixels into object or background classes [1], whereas the local schemes (adaptive thresholding) can use multiple values selected according to the local area information [2]. Most of the proposed algorithms for optimum image binarization rely on statistical methods, without taking into account the special nature of document images [3]. Global thresholding methods are not sufficient for document image binarization since document images usually have poor quality, shadows, nonuniform illumination, low contrast, large signal-dependent noise, smear and strains. Instead, techniques which are adaptive to local information have been developed for document binarization [4]. The proposed scheme for image binarization and enhancement is fully described in [5] and consists of five distinct steps: a pre-processing procedure using a low-pass Wiener filter, a rough estimation of foreground regions using Niblack's approach [3], a background surface calculation by interpolating neighboring background intensities (see Fig. 3), a thresholding by combining the calculated background surface with the original image and finally a post-processing step that improves the quality of text regions and preserves stroke connectivity. An example of the image binarization and enhancement result is demonstrated in Figs. 4 and 5.

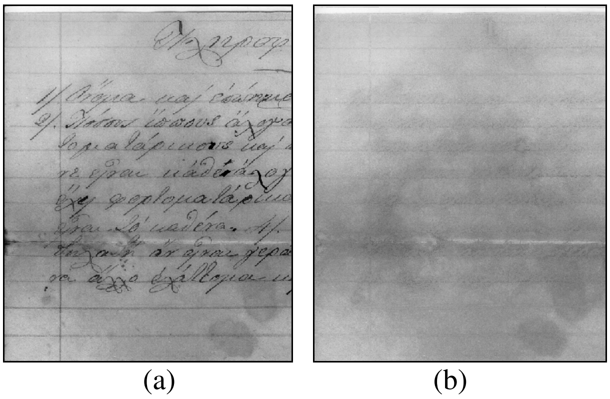


Figure 3. Background surface estimation:(a) Original image I; (b) Background surface B.

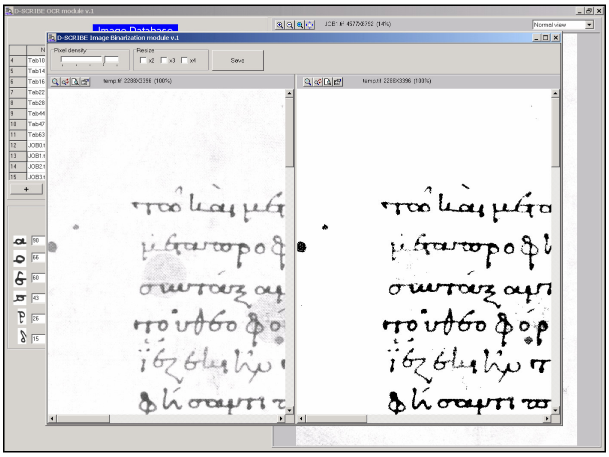
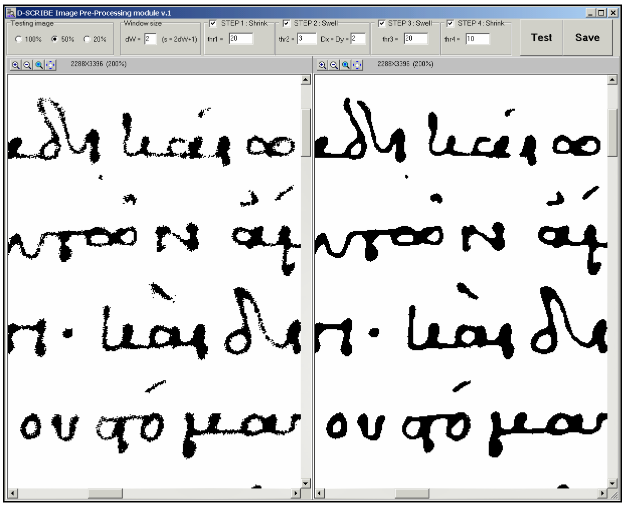


Figure 4. Image binarization step.



Testing image: ○ 100% ● 50% ○ 20%

Window size: dW = 2 (s = 2dW+1)

☑ STEP 1 : Shrink: thr1 = 20

☑ STEP 2 : Swell: thr2 = 3 Dx = Dy = 2

☑ STEP 3 : Swell: thr3 = 20

☑ STEP 4 : Shrink: thr4 = 10

Figure 5. Image enhancement step.

4.2 Recognition

In the field of handwritten character recognition a great progress has occurred during the past years [6]. Many methods were developed for a variety of applications like automatic reading of postal addresses [7], fax forms [8] and bank checks [9], form processing, etc. In the literature, two main approaches can be identified: the global approach [10] and the segmentation approach [11]. The global approach entails the recognition of the whole word while the segmentation approach requires that each word has to be segmented into letters. Some approaches that do not involve any segmentation task are based on concepts and techniques that have been used in object recognition with occlusions [12]. According to these approaches, significant geometric features such as short line segments, enclosed regions and corners are extracted from a fully unsegmented raw document bitmap by methods like template matching [13], peephole method [14], n-tuple feature [15] and hit-or-miss operator [16].

In this paper we focus on the specific example of the Sinaitic Codex Number Three, which contains the Book of Job, one of the best Greek manuscripts and one of the major masterpieces of world literature (see Fig. 2). Written in Hebrew initially, the Book was translated into Greek approximately the 3rd century B.C. for the sake of the Hellenized Hebrews of Alexandria.

Traditional techniques for handwriting recognition cannot be applied to such early Christian Greek manuscripts written in lower case letters, since continuity between characters of the same or consecutive words does not permit character or word segmentation. Furthermore, the aforementioned manuscripts entail several unique characteristics as in the following:

- High script standardization. Although, we refer to handwritten manuscripts, the corresponding characters are highly standardized since the manuscripts are immediate predecessors of early printed books.
- Frequent appearance of character ligatures
- Frequent appearance of closed cavities in the majority of character and character ligatures. As shown in Fig. 6, closed cavities appear in letters “α”, “ο”, “σ”, “ε”, “δ”, “ω”, “π”, “θ”, “φ” as well as in letter ligatures “σπ”, “εσ” etc. These constitute 60% of complete character set used in a typical old Greek manuscript.

The continuity between characters of the same or consecutive words guided us to develop a segmentation-free recognition technique as a fundamental assistance to Old Greek handwritten Manuscript OCR. Based on the existence of closed cavities in the majority of characters and character ligatures, we propose a technique for the detection and recognition of characters that contain closed cavities. It is a novel method whose originality is based on two aspects: First, a novel segmentation-free approach based on the detection of the closed cavities. This aids toward the proposed character representation since the hole regions exist in the majority of characters and character ligatures. Second, novel features are used that are based on the protrusions in the outer contour of the connected components that contain closed cavities.

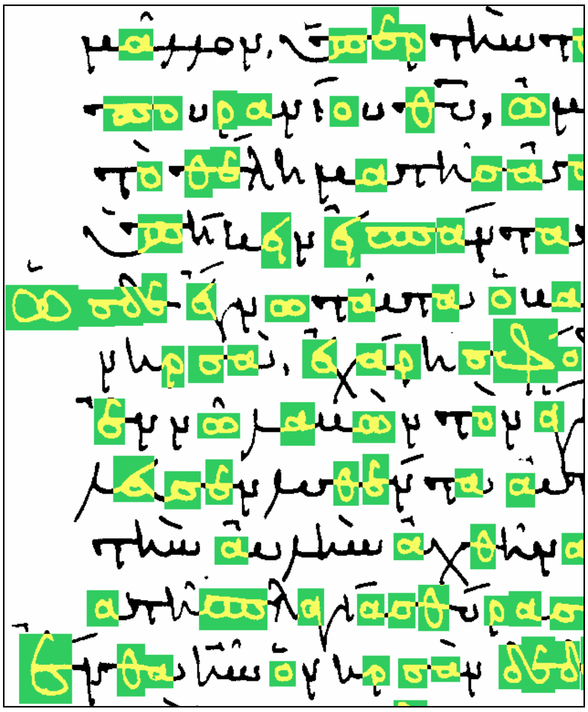
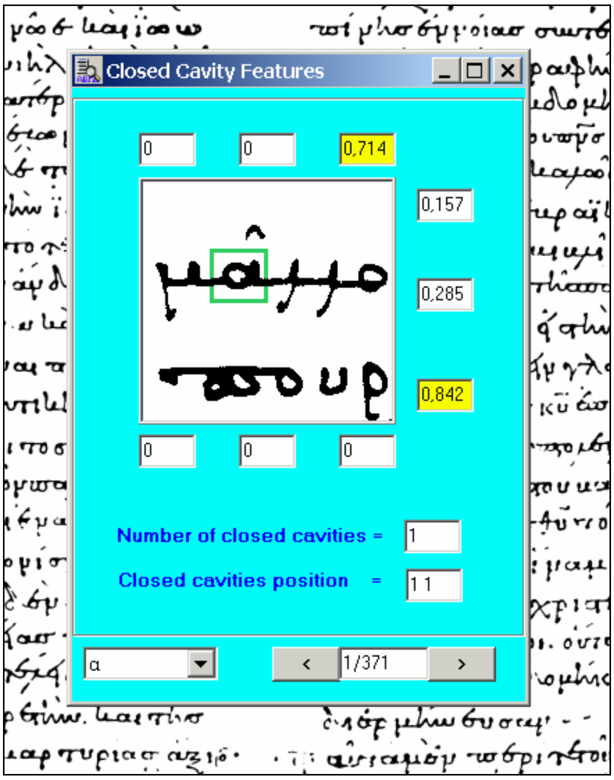


Figure 6. Early Christian Greek manuscript. Identified characters or character ligatures that contain closed cavities.

The proposed methodology consists of several distinct stages [17]. First, we apply a binarization and image enhancement technique to get an improved quality black and white (b/w) image. Second, we trace closed cavities that exist in character bodies. We suggest a novel fast algorithm based on processing the white runs of the initial b/w image. This algorithm permits the extraction of the character closed cavities but rejects closed cavities of larger dimension, such as closed cavities inside frames, diagrams, etc. In the next step, all closed cavities in characters are initially grouped into several categories based on their spatial proximity and topology. In this way, character closed cavities are classified as: a single closed cavity, two horizontal neighboring closed cavities, three horizontal neighboring closed cavities, four horizontal neighboring closed cavities, two vertical closed cavities and two vertical neighboring patterns that consist of a single closed cavity and two neighboring closed cavities (see Table 1). The final stage of our approach concerns classification of the aforementioned closed cavity patterns into a character or a ligature. It is based on the protrusions that appear in the outer contour outline of the connected components which contain the character closed cavities (see Fig. 7).

Table1. The proposed dictionary for closed cavity patterns.

Pattern ID	1	2	3	4	5	6
Pattern	o	o o	o o o	o o o o	o o	o o o
Characters or character ligatures	(α), (ο) (ε), (σ) (ρ), (δ)	(π), (ω), (εσ)	(σπ), (επ)	(απο)	(θ)	(φ)



Number of closed cavities = 1

Closed cavities position = 1 1

Figure 7. Feature extraction window.

The performance of the proposed recognition algorithm was tested using several image samples originating from different writers and manually labeled with the correct answers [17]. For this reason we also developed a ground-truth tool that helped up label all characters and character ligatures having closed cavities (see Fig. 8). Experimental results show that the proposed method gives highly accurate results that offer a great assistance to old Greek handwritten manuscript interpretation. The recognition accuracy for characters with closed cavities is approximately 95% on the sample images taken from the Book of Job.

In Fig. 9, the final recognition result for an old manuscript is demonstrated. The recognized characters and character ligatures have been placed on the original image.

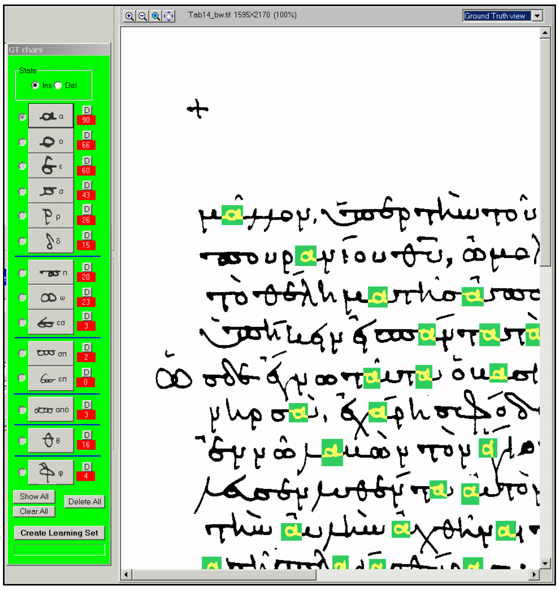


Figure 8. Ground-truth tool.

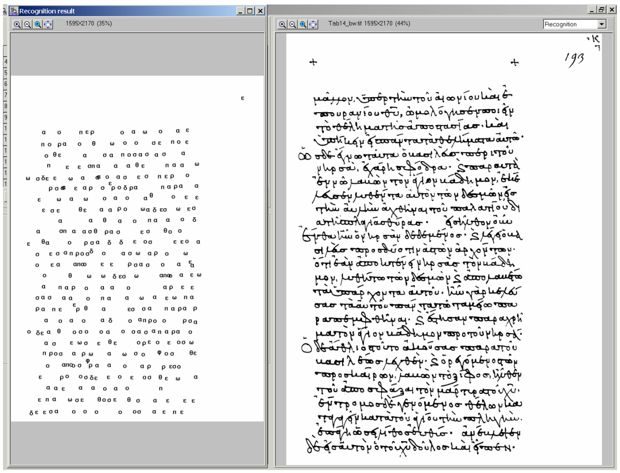


Figure 9. OCR result.

4.3 Slant estimation

By the term “character slant” we imply the angle in degrees clockwise from vertical at which the characters are drawn. It is known that each manuscript of a particular writer has some characteristics that are highly possible to be met in other manuscripts of the same writer. One of these characteristics is the character slant. Thus, the character slant is thought to be very important information which can help to point out the writer of a text. Regarding early Christian manuscripts, the slant is often used to distinguish between different scriptoria. In addition to that, character slant estimation can be very helpful in text processing. Knowing the slant’s value we can correct it in order to facilitate processing and recognition.

In the bibliography, there are various approaches concerning character slant estimation [18, 19, 20]. We used an approach similar to the one proposed by Bozinovich and Srihari [20]. According to this approach, for a given word all horizontal lines which contain at least one run of length greater than a parameter M (depending on the width of the letters) are removed. Additionally, all horizontal strips of small height are removed. By deleting these horizontal lines, only orthogonal window parts of the picture remain in the text. For each letter the parts that remain (after the deletion) are those that have relatively small horizontal slant. For each of these parts, we estimate the angle between the line indicated by the centers of gravity for its upper and lower halves and the vertical side of the page. The mean value of these slants is the overall text’s character slant.

In order to estimate the character slant in the manuscripts, we used a more efficient version of the above described approach. Because of the fact that the width of the letters differs from text to text, we used the outer contour of the character for the estimation of the slant. In Figure 10 we can see the respective results. In Figure 11 we can see an example concerning the estimation of the characters slant in manuscripts. The slant value can be either a positive or a negative number depending on whether we have a right or left slant of the letters of the manuscripts.

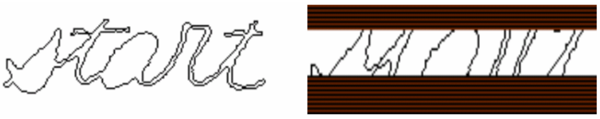


Figure 10. Estimating character slant using the outer contour of the characters.

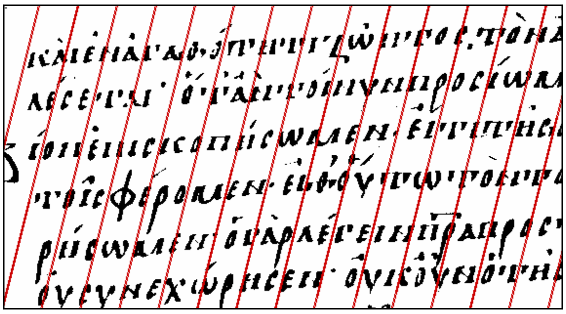


Figure 11. Examples of character slant estimation in manuscripts. The estimated slant is +14°.

4.4 Querying by example

Querying by example involves searching for words similar to a word image marked in a particular location of a manuscript. The search procedure is applied to a set of images. The steps we follow in order to trace the words can be divided into two groups. The first group consists of “off-line” steps that are followed only once for each image of a manuscript when images are initially imported. The second group consists of "on-line" steps that are followed during searching for the marked word.

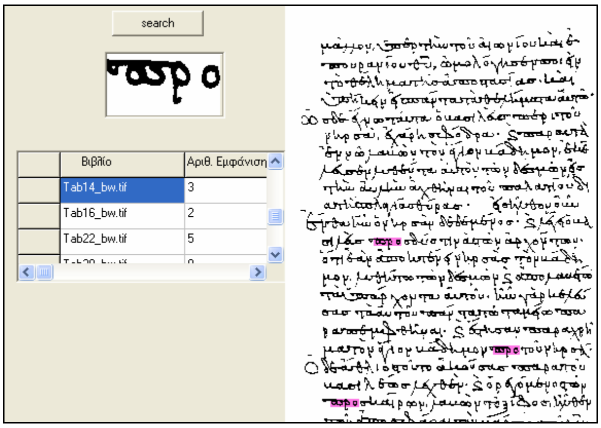
The first group of "off-line" steps is as follows:

- All closed cavities of the image are traced and features extracted in their vicinity are stored. Features are extracted as in section 4.2.
- The text lines of the image of the manuscript are extracted and the line to which each closed cavity belongs is identified [21, 22].
- The closed cavities of each line are sorted in a list according to their distance from the beginning of the line. The hole which is nearest to the beginning of the line is the first one in the list.

The second group of steps is described as follows:

- The user marks the word he/she wants.
- All closed cavities of the selected part of the image are traced and the corresponding features are calculated
- For each text line in the stored images, a search is performed for combinations of features similar to the ones extracted in the previous step.
- Best matching words are marked on the manuscript image so as to be easily spotted by the user.

In Figure 12 we give an example of the search procedure. Te user has marked the world "προ" in one of the available manuscript images. In all available manuscript images, words similar to the original word "προ" selected by the user are found by the system and automatically marked.



	Βιβλίο	Αριθ. Εμφάνιση
	Tab14_bw.tif	3
	Tab16_bw.tif	2
	Tab22_bw.tif	5

Figure 12. Searching by using examples

5 Self-Study Tool

Apart from the tools for processing the manuscripts, an innovative educational software tool has been developed to support the process of studying and transliterating the manuscripts by experts and students. This is envisaged as a self-study tool that will be able to guide manuscript readers for using the technology developed in the D-SCRIBE project and teach them to understand the main different kinds of handwriting styles used by early Greek manuscript transcribers.

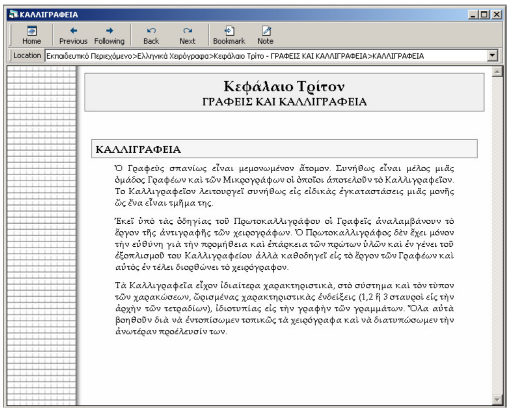
The self-study tool is called STUD-IOS and consists of two subsystems.

The 1st subsystem aims at covering the need for self-instruction of users for the digitalization and processing of old Greek manuscripts. It serves as an extensive manual of the technologies developed in the D-SCRIBE project. It helps the user to formulate a clear picture of all steps that need to be carried out for every operation.

The 2nd subsystem covers the subject of paleography, explaining the different types of writing, faculties of paleography, materials used for paleography, techniques of paleography etc. It also serves to teach methods of manuscript transcription in Modern Greek script. The main characteristics of the 2nd subsystem are the following:

- Presentation of material with use of lists of contents in tree form, beginning from units, sub-sections and leading to the pages of educational material.
- Extensive use of index of terms with direct link in the pages in which terms are contained.
- Use of a tool for keeping of notes which can be interlinked with each page of the educational material.
- Provision of tools for conducting tests for the self-assessment in each cognitive unit. This operation supports a more complete form of learning.
- Use of friendly interactive techniques for the automatic evaluation of progress in tests and exercises.

Sample screenshots from the self-study tool are shown in Figs. 13 and 14. Fig. 13 depicts a typical lesson in paleography while Fig. 14 shows a typical test for the student.



Εκπαιδευτικό Περιεχόμενο>Ελληνικά Χειρόγραφα>Κεφάλαιο Τρίτο - ΓΡΑΦΕΙΣ ΚΑΙ ΚΑΛΛΙΓΡΑΦΕΙΑ>ΚΑΛΛΙΓΡΑΦΕΙΑ

Κεφάλαιο Τρίτον
ΓΡΑΦΕΙΣ ΚΑΙ ΚΑΛΛΙΓΡΑΦΕΙΑ

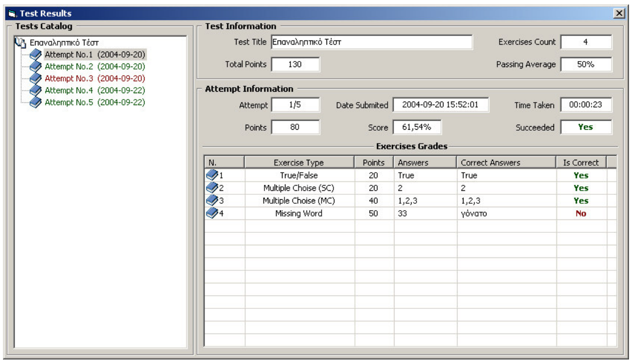
ΚΑΛΛΙΓΡΑΦΕΙΑ

Ὁ Γραφεὺς σπανίως εἶναι μεμονωμένον ἄτομον. Συνήθως εἶναι μέλος μιᾶς ὁμάδος Γραφέων καὶ τῶν Μικρογράφων οἱ ὁποῖοι ἀποτελοῦν τὸ Καλλιγραφεῖον. Το Καλλιγραφεῖον λειτουργεῖ συνήθως εἰς εἰδικὰς ἐγκαταστάσεις μιᾶς μονῆς ὥς ἕνα εἶναι τμῆμα της.

Ἐκεῖ ὑπὸ τὰς ὁδηγίας τοῦ Πρωτοκαλλιγράφου οἱ Γραφεῖς ἀναλαμβάνουν τὸ ἔργον τῆς ἀντιγραφῆς τῶν χειρογράφων. Ὁ Πρωτοκαλλιγράφος δὲν ἔχει μόνον τὴν εὐθύνη γιὰ τὴν προμήθεια καὶ ἐπάρκεια τῶν πρώτων ὑλῶν καὶ ἐν γένει τοῦ ἐξοπλισμοῦ του Καλλιγραφείου ἀλλὰ καθοδηγεῖ εἰς τὸ ἔργον τῶν Γραφέων καὶ αὐτὸς ἐν τέλει διορθώνει τὸ χειρόγραφον.

Τὰ Καλλιγραφεῖα εἶχον ἰδιαίτερα χαρακτηριστικὰ, στὸ σύστημα καὶ τὸν τὺπον τῶν χαρακώσεων, ὥρισμένας χαρακτηριστικὰς ἐνδείξεις (1,2 ἢ 3 σταυροὶ εἰς τὴν ἀρχὴν τῶν τετραδίων), ἰδιοτυπίας εἰς τὴν γραφὴν τῶν γραμμάτων. Ὅλα αὐτὰ βοηθοῦν διὰ νὰ ἐντοπίσωμεν τοπικῶς τὰ χειρόγραφα καὶ νὰ διατυπώσωμεν τὴν ἀνωτέραν προέλευσίν των.

Figure 13. Typical paleography lesson in the STUD-IOS self study tool.



Επαναληπτικό Τέστ
- Attempt No.1 (2004-09-20)
- Attempt No.2 (2004-09-20)
- Attempt No.3 (2004-09-20)
- Attempt No.4 (2004-09-22)
- Attempt No.5 (2004-09-22)

Test Title: Επαναληπτικό Τέστ
Exercises Count: 4
Total Points: 130
Passing Average: 50%

Attempt: 1/5
Date Submited: 2004-09-20 15:52:01
Time Taken: 00:00:23
Points: 80
Score: 61,54%
Succeeded: Yes

N.	Exercise Type	Points	Answers	Correct Answers	Is Correct
1	True/False	20	True	True	Yes
2	Multiple Choise (SC)	20	2	2	Yes
3	Multiple Choise (MC)	40	1,2,3	1,2,3	Yes
4	Missing Word	50	33	γόνατο	No

Figure 14. Typical test in the STUD-IOS self study tool.

6 Conclusions and Further Work

A number of modular software tools have been developed for supporting and facilitating old Greek manuscript digitization and processing. These tools include a document management system particularly adapted for manuscripts, a module for automatic processing and transliteration of manuscripts incorporating OCR techniques and a self study tool for the use of paleographers.

In this paper, we focus on the description of the aforementioned modules that comprise the final system for processing and recognition of old Greek Manuscripts. We propose a novel digital image binarization scheme for low quality historical documents allowing further content exploitation in an efficient way. Additionally, we present a novel methodology that assists recognition of early Christian Greek manuscripts. We strive toward an assessment of the recognition procedure by tracing and recognizing the most frequently appearing characters or character ligatures, using a segmentation-free, quick and efficient approach. Based on the observation that closed cavities appear in the majority of characters and character ligatures, we propose a recognition technique that consists of several distinct stages. Additional tools for character slant estimation and querying by example have been developed. Apart from the tools for processing the manuscripts, an innovative educational software tool has been developed to support the process of studying and transliterating the manuscripts by experts and students.

Future work mainly involves the detection and recognition of the remaining old Greek handwritten character and character ligatures that do not include closed cavities, as well as the testing of the performance of the proposed technique for other types of old handwritten historical manuscripts. Additionally, an efficient post-processing lexicon matching technique will be employed in order to further increase the accuracy of the results. To achieve this, a higher word level comparison strategy will be developed. In particular, the OCR results will be grouped into words and then searched within a lexicon. This lexicon will be given by the expert users and shall constitute of the majority of the words found on the manuscript images. For this purpose, spelling correction software based on efficient associative memory techniques will be extended for handling the task at hand. An additional evaluation procedure will then take place in order to determine the overall recognition accuracy as well as the improvement of the recognition accuracy after post-processing. Finally, it is planned to enrich the self study tool with more training material and tests for the interested students of paleography.

References:

[1] N. Otsu, A threshold selection method from gray-level histograms, *IEEE Trans. Systems Man Cybernet*, 9 (1), 1979, pp. 62-66.

[2] I. K. Kim, R. H. Park, Local adaptive thresholding based on a water flow model, *Second Japan-Korea Joint Workshop on Computer Vision*, Japan, 1996, pp. 21-27.

[3] W. Niblack, *An Introduction to Digital Image Processing*. Englewood Cliffs, N. J., Prentice Hall, 1986, pp. 115-116.

[4] J. Sauvola, M. Pietikainen, Adaptive Document Image Binarization, *Pattern Recognition* 33, 2000, pp. 225-236.

[5] B. Gatos, I. Pratikakis, S. J. Perantonis, Locating Text in Historical Collection Manuscripts, *Lecture Notes on AI, SETN 2004*, pp. 476-485.

[6] A. Vinciarelli, A survey on off-line Cursive Word Recognition, *Pattern Recognition* 35, 2002, pp. 1433-1446.

[7] Y. Lu, C. L. Tan, Combination of multiple classifiers using probabilistic dictionary and its application to postcode recognition, *Pattern Recognition* 35, 2002, pp. 2823-2832.

[8] T. Hirano, Y. Okada, F. Yoda, Field Extraction Method from Existing Forms Transmitted by Facsimile, *Sixth International Conference on Document Analysis and Recognition, ICDAR2001*, 2001, pp. 738-742.

[9] Q. Xu, L. Lam, C. Y. Suen, A Knowledge-based Segmentation System for Handwritten Dates on Bank Cheques, *Sixth International Conference on Document Analysis and Recognition, ICDAR2001*, 2001, pp. 384-388.

[10] C. Y. Suen et al., Building a New Generation of Handwriting Recognition Systems, *Pattern Recognition Letters* 14, 1993, pp. 303-315.

[11] E. Kavallieratou, N. Fakotakis, G. Kokkinakis, Handwritten character recognition based on structural characteristics, *16th International Conference on Pattern Recognition*, 2002, pp. 139-142.

[12] C. H. Chen, J. de Curtins. Word Recognition in a Segmentation-Free Approach to OCR, *Second International Conference on Document Analysis and Recognition (ICDAR'93)*, 2003, pp. 573-576.

[13] R. Duda, E. Hart, *Pattern Classification and Scene Analysis*, Wiley 1973.

[14] S. Mori, C. Y. Suen, K. Yamamoto, Historical review of OCR research and development, *Proc. IEEE*, vol. 80, 1992, pp. 1029-1058.

[15] D. M. Jung, M. S. Krishnamoorty, G. Nagy, A. Shapira, N-tuple features for OCR revisited, *IEEE Trans. PAMI* vol. 18, no. 7, 1996, pp. 734-745.

[16] R. C. Gonzalez, R. E. Woods, *Digital Image Processing*, Addison-Wesley, 1992.

[17] B. Gatos, K. Ntzios, I. Pratikakis, S. Petridis, T. Konidaris and S. J. Perantonis, A segmentation-free recognition technique to assist old Greek handwritten manuscript OCR, *IAPR Workshop on Document Analysis systems (DAS'2004)*, *Lecture Notes in Computer Science (3163)*, Florence, Italy, September 2004, pp. 63-74.

[18] V.K. Sagar, S.W. Chong, Slant Manipulation And Character Segmentation For Forensic document Examination, *IEEE TENCON-Digital Signal Processing Applications*, 1996, pp.933-938.

[19] E. Kavallieratou N.Fakotakis G Kokkinakis, New algorithms for skewing correction and slant removal on word-level, *Proc. IEEE,* 1999, pp 1159-1162.

[20] A. Bozinovich, A.Srihari, Off-Line Cursive Script Word Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol II, No 1,January 1989, pp. 69-82.

[21] A. Zahour, B. Taconet, Arabic hand-written text-line extraction, *Proc IEEE,* 2001, pp.281-285.

[22] M. Sawaki, N. Hagita, Text-line Extraction and Character Recognition of Japanese Newspaper Hea dlines with, IEEE, *Proceedings of ICPR 1996*, pp. 73-78.