

An old greek handwritten OCR system based on an efficient segmentation-free approach

K. Ntzios · B. Gatos · I. Pratikakis · T. Konidaris ·
S. J. Perantonis

Received: 21 February 2005 / Revised: 24 March 2006 / Accepted: 25 July 2006 / Published online: 24 January 2007
© Springer-Verlag 2007

Abstract Recognition of Old Greek Early Christian manuscripts is essential for efficient content exploitation of the valuable Old Greek Early Christian historical collections. In this paper, we focus on the problem of recognizing Old Greek manuscripts and propose a novel recognition technique that has been tested in a large number of important historical manuscript collections which are written in lowercase letters and originate from St. Catherine's Mount Sinai Monastery. Based on an open and closed cavity character representation, we propose a novel, segmentation-free, fast and efficient technique for the detection and recognition of characters and character ligatures. First, we detect open and closed cavities that exist in the skeletonized character body. Then, the classification of a specific character or

character ligature is based on the protrusible segments that appear in the topological description of the character skeletons. Experimental results prove the efficiency of the proposed approach.

Keywords Historical document recognition · Handwriting character recognition · Segmentation-free OCR

1 Introduction

Recognition of old Greek manuscripts is essential for quick and efficient content exploitation of the valuable old Greek historical collections. In this paper, we focus on the recognition of Early Christian Greek manuscripts written in lower case letters (see Fig. 1a). Old Greek manuscripts can be found at writings from the Jewish Bible that became part of the Christian Old Testament, at copies of early extra-canonical writings such as the Gospel of Thomas or the Shepherd of Hermas, and at fragments of other, unknown writings, as well as liturgical and theological texts. Any manuscript of Christian provenance can provide valuable historical information about early Christianity. Particularly, the Sinaitic Codex Number Three, which contains the Book of Job, is one of the best Greek manuscripts and one of the major masterpieces of world literature. Written in Hebrew initially, the Book was translated into Greek approximately in the third century BC for the sake of the Hellenized Hebrews of Alexandria.

The work described in this paper has been developed within the framework of the Greek Ministry of Research funded R&D project, D-SCRIBE, which aims

K. Ntzios (✉) · B. Gatos · I. Pratikakis ·
T. Konidaris · S. J. Perantonis
Computational Intelligence Laboratory,
Institute of Informatics and Telecommunications,
National Research Center "Demokritos",
153 10 Athens, Greece
URL: <http://www.iit.demokritos.gr/cil>
e-mail: ntzios@iit.demokritos.gr

K. Ntzios
Department of Informatics and Telecommunications,
National and Kapodistrian University of Athens,
Athens, Greece
e-mail: ntzios@di.uoa.gr

B. Gatos
e-mail: bgat@iit.demokritos.gr

I. Pratikakis
e-mail: ipratika@iit.demokritos.gr

T. Konidaris
e-mail: tkonid@iit.demokritos.gr

S. J. Perantonis
e-mail: sper@iit.demokritos.gr

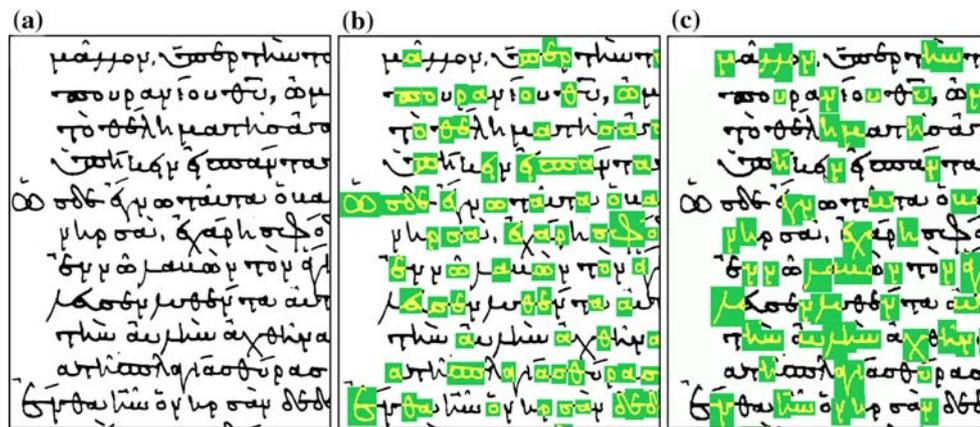


Fig. 1 **a** Early Christian Greek manuscript; **b** Identified characters or character ligatures that contain closed cavities **c** Identified characters or characters ligatures that contain open cavities

to develop an integrated system for digitization and processing of Old Greek manuscripts. It is expected that by the end of the project, 150,000 individual pages will be processed. D-SCRIBE strives toward 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. 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 cultural heritage.

In the field of handwriting recognition a great progress has occurred during the past years [35]. Many methods were developed for a variety of applications like automatic reading of postal addresses [2,19], fax forms [14] and bank checks [11,39], form processing, etc. In methodology, two general approaches can be identified: the segmentation approach [7,16] and the global or segmentation-free approach [12,13,34]. The segmentation approach requires that each word has to be segmented into characters while the global approach entails the recognition of the whole word.

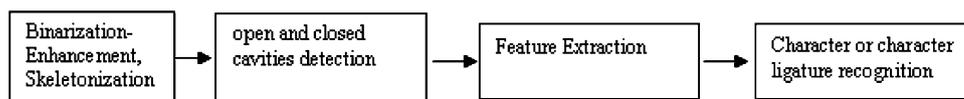
In the segmentation approach, the crucial step is to split a scanned bitmap image of a document into individual characters. Many segmentation algorithms have been proposed for handwritten words and digits. Lu and Shridhar gave an overview of the various techniques

for the segmentation of handwritten characters [20]. Xiao and Leedman proposed a segmentation method based on certain knowledge of the handwriting [37], while Plamondon and Privitera introduced a segmentation method that partly simulates the cognitive-behavioral process used by human beings in order to recover the temporal sequence of the strokes that composed the original pen movement [30]. Chi et al. proposed a contour curvature-based algorithm to segment single and double-touching handwritten digit strings [3]. Shuyan et al. proposed a two-stage approach to segment unconstrained handwritten Chinese characters [32]. In their algorithm a character string is first coarsely segmented on the basis of the background skeleton, a vertical projection and a set of geometric features. All possible segmentations paths are evaluated by using the fuzzy decision rules learned from examples discarding unsuitable segmentation paths.

Global approaches avoid character segmentation, looking at words as entities using statistical methods to classify word samples [8]. Holistic strategies employ top-down approaches for recognizing the whole word, thus eliminating the segmentation problem [21,22,33]. In these strategies, global features extracted from the entire word image are used for the recognition of limited-size lexicon. As the size of the lexicon becomes larger, the complexity of algorithms increases linearly due to the need for a larger search space and a more complex pattern representation. Although the global approaches are referred in the literature as "segmentation-free" approaches, they involve a word detection task.

Some approaches that do not involve any segmentation task are based on concepts and techniques that have been used in object recognition with occlusions [4,5]. According to these approaches, significant geometric features, such as short line segments, enclosed

Fig. 2 Overview of handwritten recognition system



regions and corners, are extracted from a fully unsegmented raw document bitmap by methods like template matching [1,6], peephole method [24], n -tuple feature [15,35] and hit-or-miss operator [12].

In the case of historical documents, Manmatha and Croft [23] presented a method for word spotting wherein matching was based on the comparison of entire words rather than individual characters. In this method, an off-line grouping of words in a historical document and the manual characterization of each group by the ASCII equivalence of the corresponding words are required. The volume of the processed material was limited to a few pages. This process can become very tedious for large collections of documents. Furthermore in [10] is presented a novel segmentation-free approach for keyword search in historical typewritten documents combining image preprocessing, synthetic data creation, word spotting and user's feedback technologies. It aims to search for keywords typed by the user in a large collection of digitized typewritten historical documents.

Traditional techniques for handwriting recognition cannot be applied to Old Greek manuscripts written in lower case letters, since continuity in writing of the same or consecutive words does not permit character or word segmentation. Furthermore, the discussed manuscripts entail several unique characteristics that are described in the following:

- Consistent script writing. Although we refer to handwritten manuscripts, the corresponding characters are highly standardized since the manuscripts are precursors of early printed books.
- Frequent appearance of character ligatures.
- Frequent appearance of open and closed cavities in the majority of character and character ligatures. As shown in Fig. 1b,c, open and closed cavities appear in letters “ α ”, “ \omicron ”, “ σ ”, “ ϵ ”, “ π ”, “ φ ”, “ β ”, “ μ ”, “ ν ”, “ λ ”, “ υ ”, etc. as well as in letter ligatures “ $\sigma\pi$ ”, “ $\epsilon\sigma$ ”, “ $\upsilon\nu$ ”, “ $\pi\tau$ ”, “ $\upsilon\pi$ ” etc. These constitute 95% of complete character set used in a typical old Greek manuscript. (see Fig. 1b, c).

The continuity in writing for characters of the same or consecutive words as well as the unique characteristics of the lower case script in Early Greek Manuscripts guided us to develop a segmentation-free recognition technique as a fundamental assistance to Old Greek

handwritten Manuscript OCR. Based on the existence of open and closed cavities in the majority of characters and character ligatures, we propose a technique for the detection and recognition of characters that contain open and closed cavities. The originality of the proposed method relies on two aspects. First, a set of discriminant features are used which are based on the protrusions that appear in the topological description of character skeletons. Second, we strive toward the detection of open and closed cavities that sets the base for a robust classifier in combination with the aforementioned discriminant features.

In the proposed method, the document image is binarized, enhanced and skeletonized. Next, we detect the open and closed cavities of the skeletonized characters where we apply a feature extraction that sets the base for the recognition process. Finally, the individual cavities are recognized on the basis of their features. In Fig. 2, an overview of this handwritten recognition system is shown.

2 Preprocessing

2.1 Image 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. 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 objects or background classes [26], whereas the local schemes (adaptive thresholding) can use multiple values selected according to the local area information [17]. Most of the proposed algorithms for optimum image binarization rely on statistical methods, without taking into account the special nature of document images [25]. Global thresholding methods are not sufficient for document image binarization since document images usually have poor quality, shadows, no uniform illumination, low contrast, large signal-dependent noise, smear and strains. Instead, techniques which are adaptive to local information have been developed for document binarization [31]. The proposed scheme for

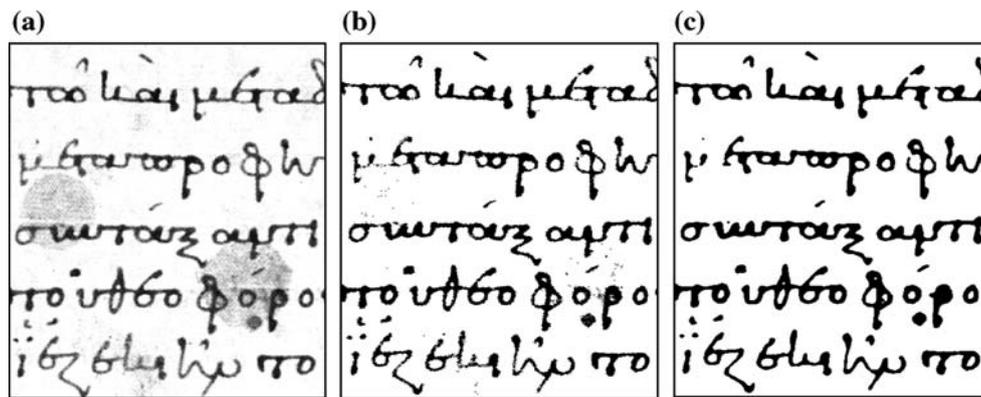


Fig. 3 Image binarization and enhancement example. **a** Original gray scale image; **b** Resulting image after binarization; **c** Resulting image after image enhancement

image binarization and enhancement is fully described in [9] and consists of five distinct steps: a preprocessing procedure using a low-pass Wiener filter, a rough estimation of foreground regions using Niblack's approach [25], a background surface calculation by interpolating neighboring background intensities, a thresholding by combining the calculated background surface with the original image and finally a postprocessing step that improves the quality of text regions and preserve stroke connectivity. An example of the image binarization and enhancement result is demonstrated in Fig. 3.

2.2 Skeletonization

For the skeletonization process, we use an iterative method presented in [18]. This method is simply an extension of the method of Zhang and Suen [40]. The skeleton obtained is not truly 8-connected, since some non-junction pixels have more than two neighbors, making the skeleton useless for algorithms that require this constraint. Therefore, some pixels have to be removed. The skeleton is inspected, and each pixel is tested using a lookup table. The result is a true 8-connected skeleton where only junction pixels have more than two 8-neighbors (see Fig. 4).

3 Open and closed cavities detection

In this step, open and closed cavities are detected in the skeletonized image. For the closed cavity, several detection algorithms exist that are mainly based on contour following techniques that distinguish the external from internal contours [29,38]. We suggest a novel fast algorithm for closed cavity detection based on processing the white runs of the b/w image. In the following,

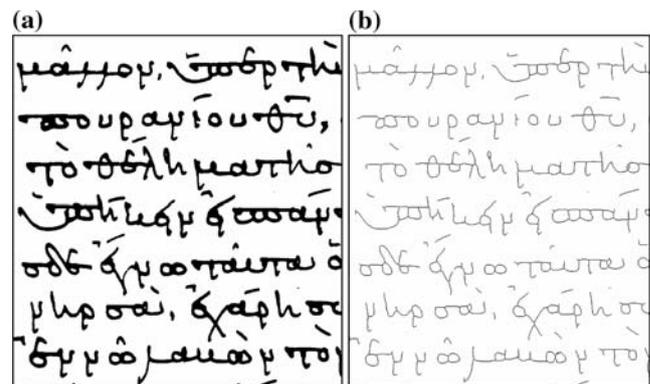


Fig. 4 **a** Binarized image. **b** Skeletonized image

a step-by-step description of the proposed algorithm is given.

- step 1 All horizontal and vertical image white runs that neighbor with image borders or have a length greater than L , get flagged, where L denotes length which reflects character size. The proposed algorithm for closed cavity detection extracts only the character closed cavities and not other closed cavities of larger dimension, with white run length greater than L , such as closed cavities inside frames, diagrams etc.;
 - step 2 All horizontal and vertical white runs of unflagged pixels that neighbor with the flagged pixels of Step 1, get flagged as well;
 - step 3 Repeat Step 2 until no pixel remains to be flagged;
 - step 4 All remaining white runs of unflagged pixels belong to image closed cavities.
- An example of the proposed closed cavity detection algorithm is demonstrated in Fig. 5.

Fig. 5 Demonstration of closed cavity detection algorithm: a–d Resulting image after 1, 2, 3 and 4 iterations, respectively

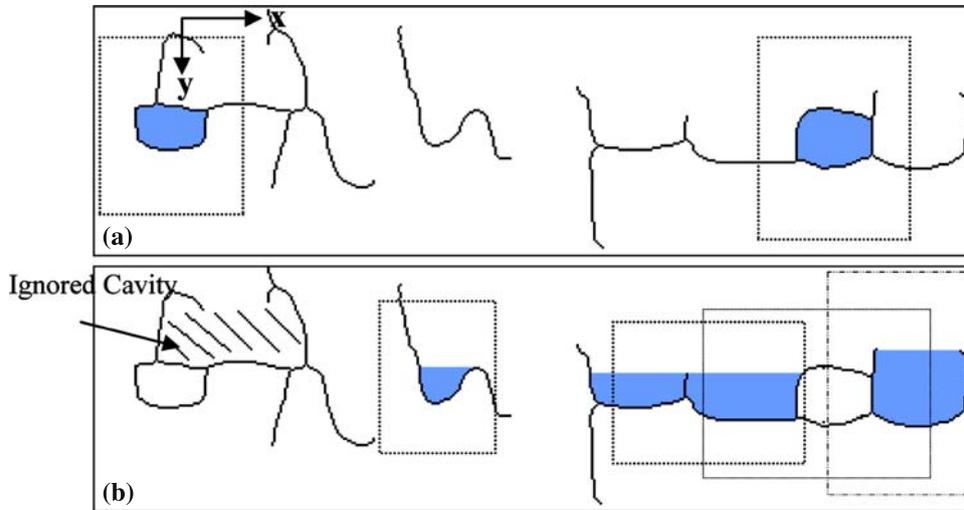
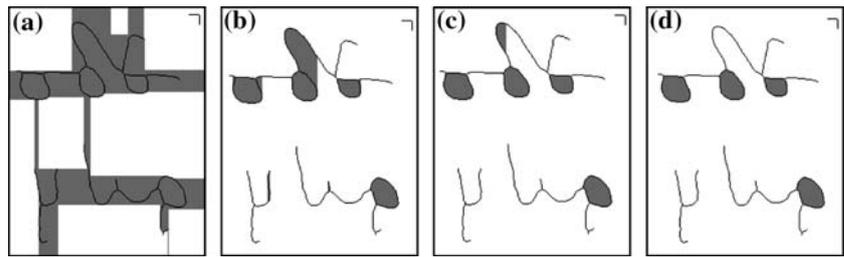


Fig. 6 **a** The skeletonized components with the respective bounding box around each closed cavity. **b** The skeletonized components with the respective bounding box around each open cavity

For the open cavities detection the water reservoir principle is used. If water is poured from top of the connected component, the cavity regions of the component where water will be stored are considered as reservoirs, [27,28] (see Fig. 6 b).

Some of the detected characters are ignored and are not considered for future processing. We consider only the cavities having width greater than a threshold T which is chosen to be the one-third of the mean width of all cavities. Additionally, an open cavity is ignored when it shares a common boundary with a closed cavity and the following condition holds:

$$\text{mean}(y_i^u) < \frac{3 * \text{mean}(y_j^o)}{5} \tag{1}$$

where $\text{mean}(y_i^u)$ is the mean value of all y -coordinates of the pixels that compose the open cavity and $\text{mean}(y_j^o)$ is the mean value of all y -coordinates of the pixels that compose the neighbor closed cavity. In Fig. 6b an ignored open cavity is shown as a shaded area.

4 Feature estimation

4.1 Character detection

Feature extraction is applied to characters that contain one or more open or closed cavities. The proposed method creates a bounding box W with the following top-left (x_{TL}, y_{TL}) and bottom right corner coordinates (x_{BR}, y_{BR}) , around the segment that has been characterized as open or closed cavity. Let $x_i \in X$, where X denotes the set of pixel coordinates of the cavity in the x direction and $y_i \in Y$, where Y denotes the set of pixel coordinates of the cavity in the y direction. The bounding box is computed as follows:

$$\begin{aligned} (x_{TL}, y_{TL}) &= \left(\min(x_i) - \frac{\text{mean}(x_i)}{2}, \min(y_i) - \frac{\text{mean}(y_i)}{2} \right) \\ (x_{BR}, y_{BR}) &= \left(\max(x_i) + \frac{\text{mean}(x_i)}{2}, \max(y_i) + \frac{\text{mean}(y_i)}{2} \right) \end{aligned} \tag{2}$$

where, $\min(\cdot)$, $\max(\cdot)$, $\text{mean}(\cdot)$ denote the minimum value, the maximum value and the average value, of the set X or Y , respectively.

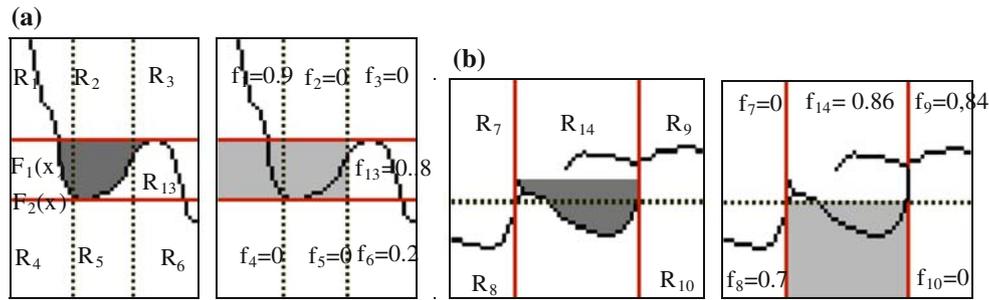


Fig. 7 a Vertical mode for the Greek letter “ η ”. *Left*: Blocks R_1, \dots, R_6 and R_{13} defined by $F_1(x)$ and $F_2(x)$. *Right*: The values of features f_1, \dots, f_6 and f_{13} . For this character $f_{11} = -0.13, f_{12} = 0$

and $f_{15} = 126^\circ$. **b** Horizontal mode for the Greek letter “ σ ”. *Left*: Blocks R_7, \dots, R_9 and R_{14} defined by $F_1(y)$ and $F_2(y)$. *Right*: The values of features f_7, \dots, f_{10} and f_{14} . For this character $f_{15} = 153^\circ$

Fig. 8 Feature f_{15} . **a** Points determination. **b** The open angle $\hat{A}P\hat{B}$

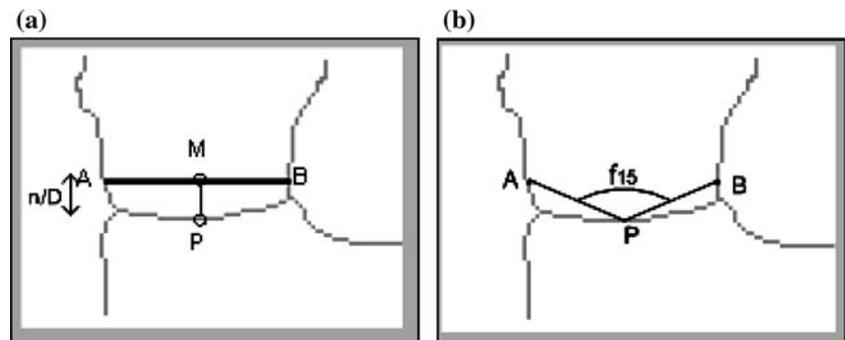


Figure 6a, b shows the skeletonized components with the corresponding bounding box W around each open and closed cavity.

4.2 Feature extraction

The feature extraction stage identifies all segments that belong to a protrusion of an isolated character’s cavity. It is applied in two consecutive modes: a *vertical* and a *horizontal* mode. The vertical mode is used to describe the protrusible segments that exist either at the top or at the bottom of the character’s cavity while the horizontal mode is used to describe the protrusible segments that exist either at the right or at the left side of the character. The feature set is composed of 15 features $\mathcal{F} = \{f_1, f_2, \dots, f_{15}\}$. More specifically, $\{f_1, f_2, f_3\}$ denotes the length of protrusible segments that appear on the top of the character’s cavity; $\{f_4, f_5, f_6\}$ denotes the length protrusible segments that appear on the bottom of the character, $\{f_7, f_8\}$ and $\{f_9, f_{10}\}$ denote the protrusible segments that appear on the left and the right side of the character and f_{11} denotes the upper slope of the protrusible segments. The remaining features ($f_{12} - f_{15}$) are used only for the characters with open cavities. Feature f_{12} denotes the lower slope of the protrusible segment, features f_{13} and f_{14} denote the length of segments that

appear in the block R_{13} and R_{14} (see Fig. 7a, b), while f_{15} denotes the opening angle of the open cavity. This angle is constructed as in the following: we first determine points A, B which denote the intersection points at the cavity and the horizontal line at a height $n \frac{1}{D}$ from the lower part of the cavity where D denotes the total height of the cavity, while n is chosen equal to 2. Then, we determine point P , which is the projection of the middle point in line AB to the lower part of the cavity. Finally the opening angle is the $\hat{A}P\hat{B}$ angle (see Fig. 8).

Feature estimation is employed in the following two steps.

- step 1: *Bounding box division into blocks*

Let $\mathcal{H} = \{(x_i^T, y_i^T), i \in [1, n]\}$, be the set of the pixel coordinates, that composes the closed or open cavity. In vertical mode we divide W into three vertical areas of equal width and assign two divide lines $F_1(x) = \min(y_i)$ and $F_2(x) = \max(y_i)$ as it shown in Fig. 9a, resulting in six blocks $\{R_1, \dots, R_6\}$. Furthermore, for the horizontal mode we divide W into two horizontal areas of equal width and assign two divide lines $F_1(y) = \min(x_i)$ and $F_2(y) = \max(x_i)$ as it is shown

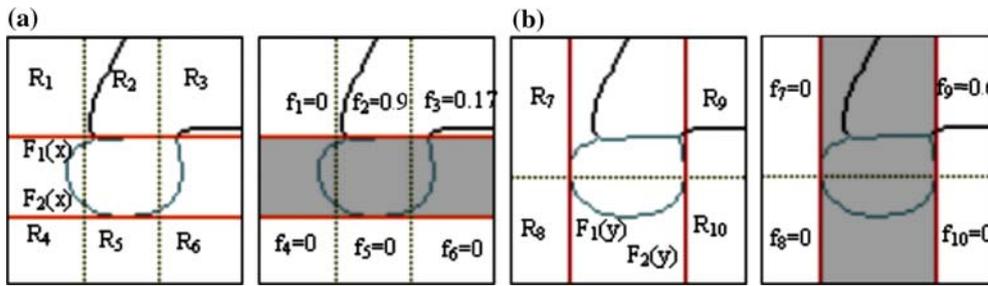
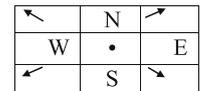


Fig. 9 **a** Vertical mode. Left: Blocks R_1, \dots, R_6 defined by $F_1(x)$ and $F_2(x)$. Right: The values of features f_1, \dots, f_6 . **b** Horizontal mode. Left: Blocks R_7, \dots, R_{10} defined by $F_1(y)$ and $F_2(y)$. Right: The values of features f_7, \dots, f_{10} . For this character $f_{11} = +0.89$

Table 1 $g_i(\cdot)$ in 8-connectivity skeleton

	E	NE	N	NW	W	SW	S	SE
g_{1-3}	0	1	1	1	0	0	0	0
g_{4-6}	0	0	0	0	0	1	1	1
$g_{7,8}$	1	1	0	0	0	0	0	1
$g_{9,10}$	0	0	0	1	1	1	0	0
g_{11}	0	1	0	-1	0	0	0	0
g_{12}	0	0	0	0	0	1	0	-1
g_{13}	0	0	0	0	0	1	1	1
g_{14}	0	0	0	1	1	1	0	0



in Fig. 9b, resulting in extra four blocks $\{R_7, \dots, R_{10}\}$.

We further consider two other blocks denoted as $R_{11} = R_1 \cup R_2 \cup R_3$ and $R_{12} = R_4 \cup R_5 \cup R_6$.

• step 2: *Block-based feature computation*

In this step, we estimate the length of a protrusion by taking into account pairs of adjacent pixels during a tracing, starting from each skeleton pixel of \mathcal{H} being in the vicinity of pixel that does not belong to \mathcal{H} .

Let, $\mathcal{H}_{R_i} = \{(x_j^{\mathcal{H}_{R_i}}, y_j^{\mathcal{H}_{R_i}}), i \in [1, 14], (x_j^{\mathcal{H}_{R_i}}, y_j^{\mathcal{H}_{R_i}}) \notin \mathcal{H}\}$ be the set of pixel coordinates depicted in block R_i and meanwhile they do not comprise pixel of the cavity. For each pixel j of \mathcal{H}_{R_i} we determine its local orientation s_j taking nominal values from the set $\{W, SW, S, SE, E, NE, N, NW\}$ in terms of the previous pixel during the tracing. Once the directions are evaluated the proposed feature f_i for closed cavities and f_j for open cavities are defined as follows:

$$f_i = \frac{1}{D} \sum_{k=1}^{m_i} g_i(s_k^i), \quad i \in [1, 11] \tag{3}$$

$$f_j = \begin{cases} \frac{1}{D} \sum_{k=1}^{m_i} g_j(s_k^j), & j \in [1, 14] \\ A\hat{P}B, & j = 15 \end{cases} \tag{4}$$

where $g_i(\cdot)$ is a function depending on the orientation of the pixel and the block considered and m_i is the total

number of pixels of the skeleton in block R_i . The term D denotes the mean of the character’s cavity height and it is used as a normalization factor allowing the feature to be invariant with respect to character scaling. The $g_i(\cdot)$ ’s are explicitly defined in Table 1, denoting the contribution of certain orientation to the corresponding protrusible segment at the region R_i .

4.3 Protrusible artifacts

For the feature estimation of the characters an upper or lower protrusible segment cannot be considered as a protrusion of more than one character, although a protrusible segment can be found for more than one character’s bounding box. Therefore, a methodology is required to assign a protrusible segment to only one character. To accomplish this, we consider a methodology strictly following the next steps:

step 1: During the closed cavity feature estimation step, we mark the pixels of their upper and lower protrusions in order to be not considered in future processing.

step 2: The bounding boxes W_i of the corresponding open cavities are sorted by starting from the left most bounding box and ending to the right most. Then, we apply a two pass process. At the first pass, for each of the sorting boxes we estimate all the features apart from f_3 and f_6 because the corresponding protrusible segments may belong to the right neighbor cavity.

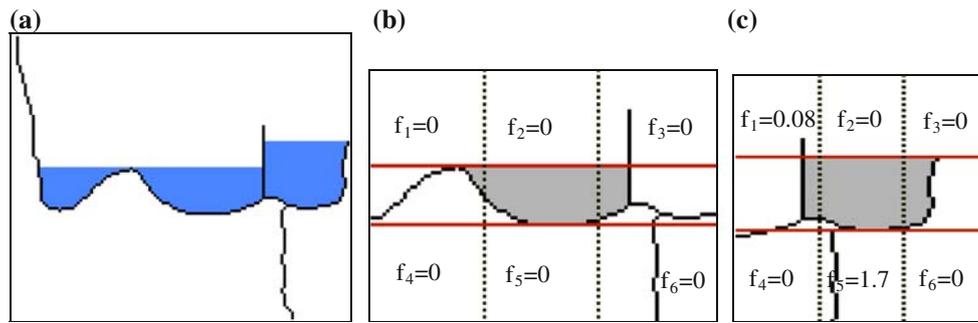


Fig. 10 **a** A connected component with three open cavities. **b** The features of the second cavity. Although at the blocks R_3 and R_6 there are protrusible segments the respective features are 0 because these segments belong to the next cavity. **c** The features of the last cavity

Table 2 The proposed dictionary for closed cavity patterns

Pattern ID	1	2	3	4	5	6
Pattern	o	oo	ooo	oooo	o	o
Characters or character ligatures with closed cavities	α (α), λ (α), ϵ (ϵ), σ (σ) ρ (ρ), $\epsilon\tau$ ($\epsilon\tau$), $\sigma\tau$ ($\sigma\tau$) δ (δ), ϵ (ϵ), o (o)	π (π), ω (ω), $\epsilon\sigma$ ($\epsilon\sigma$) $\epsilon\sigma\tau$ ($\epsilon\sigma\tau$) $\nu\pi$ ($\nu\pi$)	$\sigma\pi$ ($\sigma\pi$), $\epsilon\pi$ ($\epsilon\pi$)	$\alpha\pi$ ($\alpha\pi$)	θ (θ)	ϕ (ϕ)

All the skeleton pixels that are involved in this feature extraction phase are marked in order to be not considered in future processing. Finally, we apply a second pass and we estimate features f_3 and f_6 for each cavity taking into account only the not marked pixels (see Fig. 10)

$\min(y_j)$ is the maximum and minimum y-coordinate of j closed cavity.

Moreover, in this stage two or more open cavities that have a common boundary and they do not have upper and lower protrusible segments are merged, and the resulting cavity is characterized as a cavity with two or three open cavities (see Table 2). After merging the features of the resulting cavity are estimated exactly as being single.

4.4 Cavity merging

In this stage two or more cavities are merged when they share a common boundary. The merged closed characters can be characterized as (i) a character with two, three or four horizontal closed cavities, (ii) as a character with two vertical closed cavities and (iii) as a character with horizontal and vertical closed cavities (see Table 1). Therefore, when two closed cavities i and j , have common pixels, the merged character is characterized as horizontal if Eq. 5 is true, otherwise it is characterized as vertical.

$$\min(|\max(y_i) - \max(y_j)|, |\min(y_i) - \min(y_j)|) < \min(|\max(y_i) - \min(y_j)|, |\min(y_i) - \max(y_j)|) \quad (5)$$

where $\max(y_i)$ and $\min(y_i)$ is the maximum and minimum y-coordinate of i closed cavity and $\max(y_j)$ and

5 Character recognition

The character recognition process consists of two basic stages. In the first stage each character is classified into a pattern by their spatial configuration as shown in Tables 2 and 3. For example, the characters that have one closed cavity are classified to the pattern with ID 1 and the characters with one open cavity are classified to the pattern with ID 7. In the second stage for each pattern except the patterns with ID 4–6,8 that correspond to a unique character, there is a classification binary decision tree. Decision is taken at each node after the examination of specific feature valuation. All conditions, upon which a tree traversal is progressing, can be shown in Figs. 11,12,13,14 and 15. The corresponding

Fig. 11 Classification tree for characters of pattern with ID1

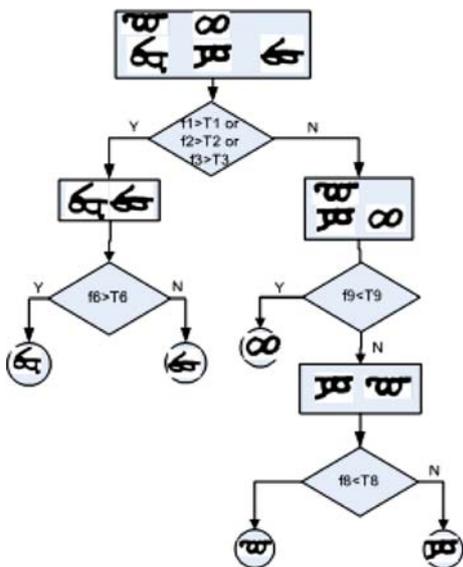
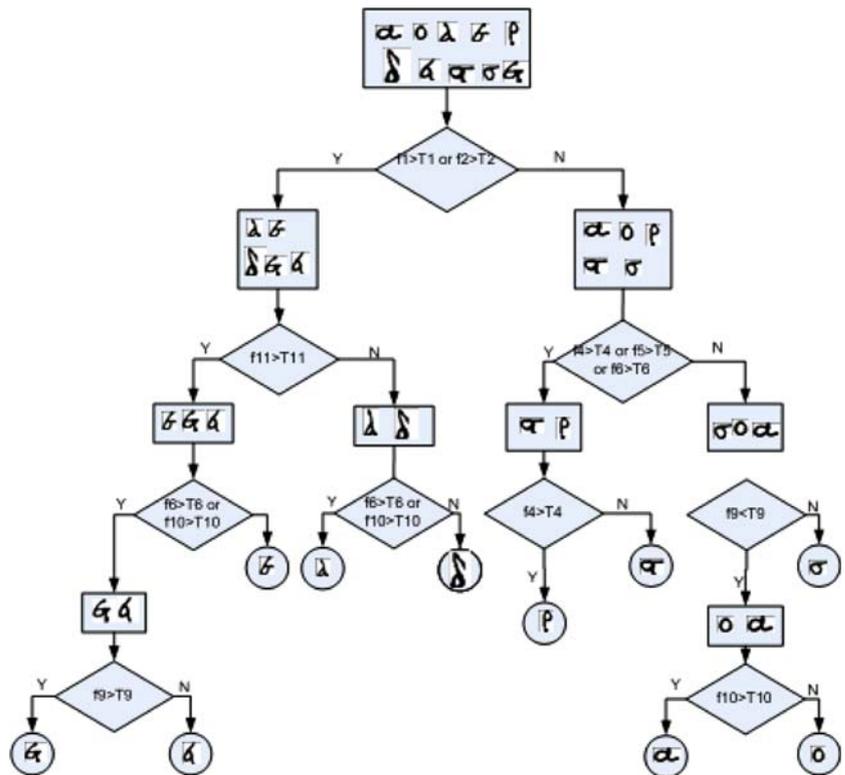


Fig. 12 Classification tree for characters of pattern with ID 2

threshold values T_i , that support the required conditioning is computed in the following:

$$T_i = \text{mean}(f_i^j) \quad i \in [1, 15], \quad j \in C \quad (6)$$

where C is the set of the training set cavities (2,497 characters).

Table 3 The proposed dictionary for open cavity patterns

Pattern ID	7	8	9
Pattern	u	uu	uuu
Characters or character ligatures with open cavities	μ (λ) ϵ (ν) ξ (κ) σ (σ) ν (ν)	η (η) μ (μ) χ (χ) ϵ (ϵ) γ (γ)	β (β) ν (ν) $\nu\nu$ ($\nu\nu$)

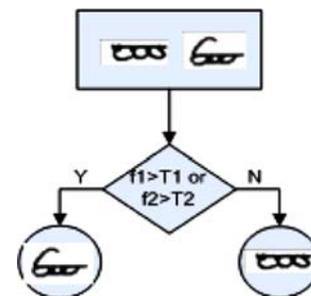


Fig. 13 Classification tree for characters of pattern with ID 3

6 Experimental results

The purpose of the experiments was to test the classification performance of the proposed handwritten manuscripts recognition procedure. The overall experimental samples originate from different manuscripts for training and testing of the Book of Job collection, manually labeled with the ground truth. We have built a dictionary of open and closed cavity patterns that contains a total of 12,332 characters and character ligatures where 2,497 characters are used for the training set and 9,835 for the testing set. The annotation was done manually in the character or character ligature level.

For the training and testing set detailed distribution of the underlying patterns along with their spatial configuration is shown in Tables 2 and 3. Table 4 shows the results obtained by applying the algorithm, indicating the recall and the precision rates for each one of the

characters. Recall (R) is the correct number of open and closed cavities classified divided by the total number open and closed cavities. Precision (P) is the number of correct open and closed cavities classified divided by the total number of open and closed cavities classified. We further compute an overall Figure of Merit (FOM) which takes into account the average precision and recall, denoted as

$$F = 2 \frac{P * R}{P + R} \tag{7}$$

Our system recognizes basic characters with an average recall of 89, 49% and overall FOM of 89, 27%. Some errors come from the bad quality of the document which cannot be overcome from the preprocessing step. In bad quality documents broken characters are recognized as open cavities whereas they belong to the closed cavity patterns. Table 5 shows the confusion matrix for the recognized characters. It can be noticed that we get certain cases of misclassifications. In particular it can be noticed that characters “λ” and “μ” are mutually misclassified 69 and 26 times, respectively. Character “δ” is misclassified as “o” 11 times, while character “ν” misclassified as “ν” 32 times.

In Tables 6 and 7 we can see for each closed (Table 6) and open (Table 7) cavity the mean value of their features and in the brackets the standard deviation of them. The features that are significant for the character classification according to the respective classification tree are marked in bold.

In Table 8, we can see some instances of “α” and “ν” characters, along with their corresponding feature set. In this table the features that are marked as black are the nonsignificant features whose values do not play any discriminant role to the character classification. The green marked features are the features whose values are very small and the red marked features are the features that describe a protrusible segment.

Table 4 Precision and recall for each character

		Number of characters	Recall	Precision
	α1	1725	96,99	98,99
	α2	79	74,68	53,15
	απ	4	100,00	100,00
	β	173	90,17	70,59
	γ	235	72,77	82,61
	δ	366	79,51	96,04
	ε1	792	91,41	95,14
	ε2	15	93,33	38,89
	ει	301	81,40	77,53
	επ	55	81,82	97,83
	ετ	57	80,70	100,00
	εστ	1	100,00	20,00
	εσ	8	100,00	61,54
	η	516	87,40	96,57
	θ	234	91,88	88,84
	κ	418	95,93	87,94
	λ	425	69,65	93,97
	μ	334	85,63	64,71
	ν	1212	92,57	93,73
	ν1	80	72,50	95,08
	ο	1812	98,51	99,28
	π	399	85,46	96,88
	ρ	463	93,30	96,43
	σ	751	96,67	87,58
	σ1	65	81,54	72,60
	στ	15	100,00	75,00
	στ	50	84,00	91,30
	υ	683	98,83	68,04
	υπ	6	83,33	83,33
	υν	66	87,88	51,79
	ψ	154	79,22	90,37
	χ	122	54,10	97,06
	ω	672	95,98	96,99
Total average		12332	89,49	89,06

7 Conclusions

In this paper, we present a novel methodology for recognition of Early Christian Greek manuscripts written in lower case letters. Using a robust character representation based on open and closed cavities, we propose a segmentation-free, quick and efficient recognition technique for the detection and recognition of characters and character ligatures. Experimental results show that the proposed method gives highly accurate results and offers a great assistance to Old Greek handwritten interpretation. We strongly believe that this system in combination with an efficient postprocessing lexicon technique on

Fig. 14 Classification tree for characters of pattern with ID 7

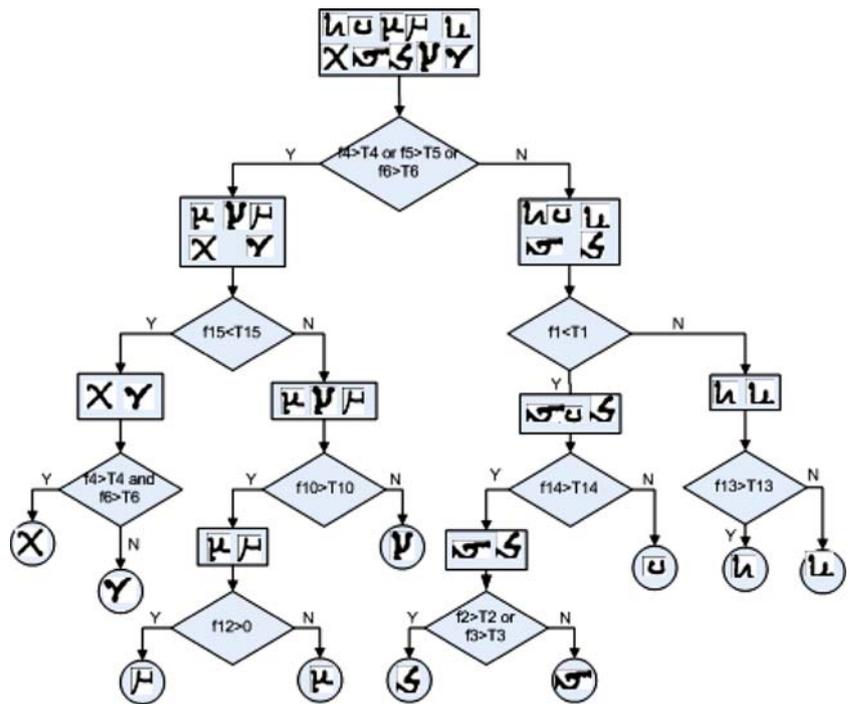


Table 5 Confusion matrix for all characters. The symbol “+” in the upper right corner describes the characters that do not belong to the character list that our system recognizes

	α1	α2	απ	β	γ	δ	ε1	ε2	ετ	επ	ετ	εστ	εσ	η	θ	κ	λ	μ	ν	ν1	ο	π	ρ	σ	σ1	σπ	στ	υ	υπ	υν	φ	χ	ω	+
α1	1673	10	0	5	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	3	0	0	7	0	0	0	5	0	0	0	1	1	17
α2	0	59	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	4	1	0	0	0	0	0	9	
απ	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
β	0	0	0	156	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	6	0	0	0	0	0	5	
γ	1	0	0	0	171	0	0	0	0	0	0	0	0	1	0	3	0	5	3	0	0	0	0	0	0	0	2	0	0	0	0	0	5	
δ	0	14	0	0	0	291	5	0	0	0	0	0	0	0	1	0	0	0	0	0	11	0	0	13	1	0	0	0	0	3	0	0	11	
ε1	0	0	0	0	0	0	724	0	10	0	0	0	1	0	0	2	0	2	0	0	0	0	0	17	0	0	0	0	0	1	0	0	35	
ε2	0	0	0	0	0	0	1	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ετ	0	12	0	0	0	0	8	245	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	24	
επ	0	0	0	0	0	0	0	0	45	0	1	2	0	0	0	0	0	0	0	0	0	0	2	0	0	0	2	0	0	0	0	0	5	
ετ	0	0	0	0	0	0	0	7	0	46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	
εστ	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
εσ	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
η	0	1	0	5	0	0	0	0	0	0	0	0	0	451	0	38	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	18	
θ	0	0	0	0	0	1	0	0	0	0	0	0	0	0	215	0	1	0	0	1	0	0	0	0	0	0	0	0	0	7	0	0	9	
κ	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	401	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	1	
λ	0	0	0	1	3	0	0	0	0	0	0	0	1	0	0	280	69	2	0	0	0	0	0	0	0	0	41	0	0	0	0	0	22	
μ	0	0	0	1	0	0	0	0	0	0	0	0	0	0	6	26	206	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ν	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	1	44	122	0	0	0	0	0	0	0	32	0	0	0	0	0	0	11	
ν1	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	58	0	0	0	0	0	0	0	2	0	13	0	0	5		
ο	0	12	0	0	0	6	1	0	0	0	0	0	0	0	0	0	0	0	0	1779	0	0	9	0	0	0	0	0	0	0	0	0	5	
π	9	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	341	1	28	0	0	0	1	0	0	0	0	17	0	
ρ	0	0	0	0	0	4	4	0	0	0	0	0	0	0	0	0	0	1	0	1	0	432	13	2	0	0	0	0	0	0	0	0	6	
σ	2	0	0	0	1	9	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	726	6	0	0	0	0	0	0	0	0	6	
σ1	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	5	53	0	0	1	0	0	0	0	2		
σπ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0	0	0		
στ	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	2	0	0	42	0	0	0	0	2	0		
υ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	1	0	0	1	0	0	675	0	0	0	0	0	2		
υπ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	5	0	0	0	0		
υν	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	1	0	58	0	0	0		
φ	0	3	0	0	11	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	122	0	6		
χ	0	0	0	13	26	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	8	0	0	66	0	8		
ω	2	0	0	0	0	0	0	13	0	3	0	0	0	0	0	0	0	0	0	2	0	0	2	0	5	0	0	0	0	0	645	0		

Circles show notable misclassification errors

Table 6 Mean value (standard deviation) of each feature for all characters with closed cavities

	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11
α	0(0)	0(0)	0,26(0,20)	0,006(0,00)	0(0)	0,04(0,08)	0,36(0,34)	0,23(0,26)	0,13(0,012)	0,98(0,10)	0,01(0,01)
β	0,92(0,11)	0,97(0,12)	0(0)	0(0)	0(0)	0,78(0,23)	0,43(0,41)	0,1(0,21)	0,2(0,16)	0,89(0,16)	-0,44(0,23)
γ	0,11(0,14)	0,01(0,01)	0(0)	0,03(0,01)	0(0)	0(0)	0(0)	0,13(0,24)	0,93(0,16)	0,11(0,06)	0(0)
δ	0,93(0,14)	0,99(0,15)	0,16(0,14)	0(0)	0(0)	0(0)	0,45(0,36)	0,44(0,3)	0,41(0,52)	0(0)	-0,66(0,44)
ε	0,95(0,13)	1,14(0,17)	0,14(0,19)	0,06(0,03)	0,014(0,08)	0,011(0,03)	0,42(0,40)	0,26(0,32)	1,06(0,21)	0,08(0,11)	0,57(0,26)
ζ	0(0)	1,07(0,11)	0,7(0,15)	0(0)	0(0)	1,17(0,21)	0,20(0,38)	0,31(0,41)	0,25(0,22)	0,65(0,34)	1,14(0,13)
η	0,99(0,13)	0,94(0,16)	0(0)	0(0)	0(0)	0(0)	0,34(0,24)	0(0)	0,99(0,13)	0(0)	0,99(0,17)
θ	0(0)	1,01(0,14)	0,97(0,33)	0(0)	0(0)	0,95(0,22)	0,44(0,45)	0,37(0,29)	0,23(0,20)	0,04(0,02)	0,95(0,17)
ι	0,88(0,17)	1,04(0,19)	0,86(0,22)	0(0)	0(0)	0,79(0,18)	0,42(0,40)	0,29(0,26)	0,97(0,12)	0,44(0,23)	0,88(0,19)
κ	0(0)	1,06(0,18)	0,79(0,23)	0(0)	0(0)	0(0)	0,44(0,4)	0,33(0,29)	1,1(0,09)	0(0)	0,92(0,2)
λ	0(0)	0,16(0,19)	0,02(0,01)	0(0)	0(0)	0(0)	0,44(0,36)	0,40(0,29)	0,45(0,33)	0,27(0,22)	0(0)
μ	0,02(0,01)	0,11(0,03)	0(0)	0(0)	0(0)	0(0)	0,46(0,4)	0,36(0,41)	0,02(0,09)	0,02(0,11)	0(0)
ν	0(0)	0(0)	0,13(0,11)	0(0)	0(0)	0(0)	1,02(0,11)	0(0)	0,98(0,14)	0,10(0,10)	0(0)
ξ	0,11(0,06)	0(0)	0(0)	0,99(0,13)	0,98(0,12)	0(0)	0,46(0,41)	0,41(0,40)	0(0)	0(0)	0(0)
ο	0,12(0,14)	0(0)	0,22(0,18)	0,03(0,04)	0(0)	0(0)	0,38(0,41)	0,42(0,28)	0,99(0,06)	0(0)	0(0)
π	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0,52(0,42)	0(0)	0,36(0,41)	0(0)	0(0)
ρ	0(0)	0(0)	0(0)	0(0)	0(0)	0,93(0,16)	0,36(0,33)	0,29(0,31)	1,01(0,13)	0(0)	0(0)
σ	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0,96(0,05)	0,79(0,28)	0,99(0,13)	0(0)	0(0)
τ	0(0)	0(0)	0(0)	0(0)	0,99(0,06)	0(0)	0(0)	0,55(0,47)	0(0)	0(0)	0(0)
υ	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)

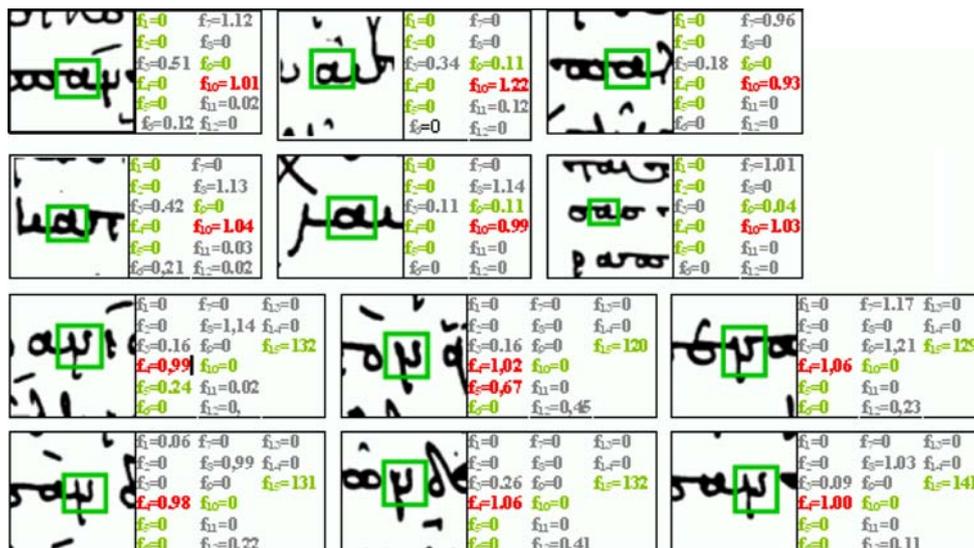
The features that are playing a discriminant role to the character classification are marked in bold

Table 7 Mean value (standard deviation) of each feature for all characters with open characters

	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14	f15
α	0,02(0,05)	0,02(0,04)	0,09(0,07)	0(0)	0,73(0,45)	0,16(0,15)	1(0,34)	0,12(0,37)	0,83(0,52)	0,5(0,44)	0,06(0,14)	0,19(0,58)	0,78(0,28)	0(0)	74,6(7,1)
β	0(0)	0,89(0,14)	0,92(0,13)	0(0)	0(0)	0(0)	0(0)	0,54(0,36)	0,89(0,18)	0(0)	0,68(0,16)	0(0)	0(0)	0,89(0,14)	143(24)
γ	1,03(0,16)	0(0)	0(0)	0(0)	0(0)	0(0)	0,44(0,29)	0,49(0,43)	0(0)	0,53(0,36)	0,13(0,17)	0(0)	0,24(0,18)	0(0)	154(25)
δ	1,02(0,15)	0(0)	0(0)	0(0)	0(0)	0(0)	0,45(0,33)	0,55(0,46)	0(0)	0,52(0,16)	0,22(0,27)	0(0)	0,64(0,11)	0(0)	128(17)
ε	0,23(0,20)	0(0)	0,18(0,12)	1,05(0,16)	0(0)	0,54(0,38)	0,45(0,23)	0,22(0,25)	0,22(0,16)	0,89(0,15)	0(0)	0,67(0,15)	0(0)	0(0)	155(22)
ζ	0,31(0,17)	0(0)	0,18(0,19)	0,99(0,12)	0(0)	0(0)	0,32(0,38)	0,64(0,24)	0(0)	0,99(0,17)	0(0)	0,06(0,14)	0(0)	0(0)	134(23)
η	0,42(0,25)	0(0)	0,41(0,46)	0,94(0,16)	0,93(0,17)	0(0)	0,44(0,36)	0,52(0,47)	0,33(0,34)	0(0)	0(0)	0,08(0,13)	0(0)	0(0)	129(24)
θ	0(0)	0(0)	0,18(0,17)	0(0)	0(0)	0(0)	0,55(0,43)	0,32(0,13)	1,02(0,16)	0(0)	0(0)	0(0)	0(0)	0,66(0,13)	144(25)
ι	0,14(0,08)	0(0)	0,24(0,23)	0(0)	0(0)	0(0)	0,54(0,5)	0(0)	0,34(0,41)	0(0)	0(0)	0(0)	0(0)	0(0)	140(24)
κ	0,31(0,22)	0(0)	0,32(0,36)	0,99(0,14)	0(0)	0,97(0,16)	0,45(0,33)	0(0)	0,45(0,52)	0(0)	0,02(0,01)	0,14(0,33)	0(0)	0(0)	79(13)

The features that are playing a discriminant role to the character classification are marked in bold

Table 8 Some instances of characters ‘α’ and ‘ν’ with their corresponding feature values



The features that are playing a discriminant role to the character classification are marked with green or red color. With red color are marked the features that describe a remarkable protrusible segment

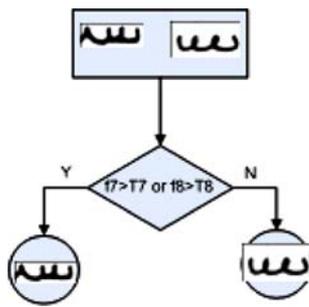


Fig. 15 Classification tree for characters of pattern with ID 9

Early Christian Greek manuscripts will further increase the accuracy of the results.

Acknowledgments This research is carried out within the framework of the Greek Ministry of Research funded R&D project, D-SCRIBE, which aims to develop an integrated system for digitization and processing of Old Greek manuscripts.

References

- Amin, A., Masini, G.: Machine recognition of cursive Arabic words, application of digital image processing IV. San Diego, CA, vol. SPIE-359, pp. 286–292 (1982)
- Brakensiek, A., Rottland, J., Rigoll, G.: Confidence measures for an address reading system. In: 7th International Conference on Document Analysis and Recognition, ICDAR 2003, pp. 294–298 (2003)
- Chi, Z., Suters, M., Yan, H.: Separation of single- and double-touching handwritten numeral strings. *Opt. Eng.* **34**, 1159–1165 (1995)
- Chen, C.H., Curtins, J.: Word recognition in a segmentation-free approach to OCR. In: 2nd International Conference on Document Analysis and Recognition (ICDAR'93), pp. 573–576 (2003)
- Chen, C.H., Curtins, J.: A Segmentation-free approach to OCR. *IEEE Workshop on Applications of Computer Vision*, pp. 190–196 (1992)
- Duda, R., Hart, E.: *Pattern Classification and Scene Analysis*. Wiley, New York (1973)
- Eastwood, B., Jennings, A., Harvey, A.: A feature based neural network segmenter for handwritten words. In: International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'97), pp. 286–290. Australia (1997)
- Farag, R.: Word-level recognition of cursive script. *IEEE Trans. Comput.* **C-28**, pp. 172–175 (1979)
- Gatos, B., Pratikakis, I., Perantonis, S.J.: Adaptive degraded document image binarization. *Pattern Recogn.* **39**, 317–327 (2006)
- Gatos, B., Konidaris, T., Ntzios, K., Pratikakis, I., Perantonis, S.: A segmentation-free approach for keyword search in historical typewritten documents. In: 8th International Conference on Document Analysis and Recognition (ICDAR'05), Seoul, Korea, (2005)
- Gorski, N., Anisimov, V., Augustin, E., Baret, O., Price, D., Simon, J.C.: A2iA check reader: a family of bank check recognition systems. In: Proceedings of 5th International Conference on Document Analysis and Recognition, pp. 523–526 (1999)
- Gonzalez, R.C., Woods, R.E.: *Digital Image Processing*. Addison-Wesley, Reading (2003)
- Guillevic, D., Suen, C.Y.: HMM word recognition engine. In: 4th International Conference on Document Analysis and Recognition ICDAR97, pp. 544 (1997)
- Hirano, T., Okada, Y., Yoda, F.: Field extraction method from existing forms transmitted by facsimile. In: 6th International Conference on Document Analysis and Recognition, ICDAR2001, pp. 738–742 (2001)
- Jung, D.M., Krishnamoorthy, M.S., Nagy, G., Shapira, A.: N-tuple features for OCR revisited. *IEEE Trans. PAMI* **18**(7), 734–745 (1996)
- Kavallieratou, E., Fakotakis, N., Kokkinakis, G.: Handwritten character recognition based on structural characteristics. In: 16th International Conference on Pattern Recognition, pp. 139–142 (2002)
- Kim, I.K., Park, R.H.: Local adaptive thresholding based on a water flow model. In: 2nd Japan–Korea Joint Workshop on Computer Vision, pp. 21–27. Japan (1996)
- Lee, H.J., Chen, B.: Recognition of handwritten Chinese characters via short line segments. *Pattern Recogn.* **25**(5), 543–552 (1992)
- Lu, Y., Tan, C.L.: Combination of multiple classifiers using probabilistic dictionary and its application to postcode recognition. *Pattern Recogn.* **35**, 2823–2832 (2002)
- Lu, Y., Shridhar, M.: Character segmentation in handwritten words—an overview. *Pattern Recogn.* **29**(1), 77–96 (1996)
- Madhvanath, S., Kleinger, E., Govindaraju, V.: Holistic verifications of handwritten phrases. *IEEE Trans. PAMI* **21**: 1344–1356 (1999)
- Madhvanath, S., Govindaraju, V.: Holistic lexicon reduction. In: Proceedings of the 3rd International Workshop on Frontiers in Handwriting Recognition, pp. 71–82 Buffalo, NY (1993)
- Manmatha, R., Croft, W.B.: A draft of word spotting: indexing handwritten manuscripts. In: *Intelligent Multimedia Information Retrieval*, pp. 43–64. MIT Press, Cambridge, MA (1997)
- Mori, S., Suen, C.Y., Yamamoto, K.: Historical review of OCR research and development. *Proc. IEEE*, **80**, 1029–1058 (1992)
- Niblack, W.: *An Introduction to Digital Image Processing*. pp. 115–116. Prentice Hall, Englewood Cliffs, NJ, (1986)
- Otsu, N.: A threshold selection method from gray-level histograms, *IEEE trans. Syst. Man Cybern.* **9**(1), 62–66 (1979)
- Pal, U., Sarkar, A.: Recognition of printed urdu script. In: Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR 2003)
- Pal, U., Belaid, A., Choisy, Ch.: Touching numeral segmentation using water reservoir concept. *Pattern Recogn. Lett.* **24**, 261–272 (2003)
- Pavlidis, T.: *Algorithms for Graphics and Image Processing*. Computer Science Press, Rockville, MD (1992)
- Plamondon, P., Privitera, C.M.: The segmentation of cursive handwritten: an approach based on off-line recovery of the motor-temporal information. *IEEE Trans. Image Process.* **8**, 80–91 (1999)
- Sauvola, J., Pietikainen, M.: Adaptive document image binarization, *Pattern Recogn.* **33**, pp. 225–236 (2000)
- Shuyan, Z., Zheru, C., Penfei, S., Hong, Y.: Two-stage segmentation of unconstrained handwritten Chinese characters. *Pattern Recogn.* **36** 145–156 (2003)
- Simon, J.: Off-line cursive word recognition. In: *Proc. IEEE* **80**, 1150–1161 (1992)
- Suen, C.Y.: Building a new generation of handwriting recognition systems. *Pattern Recogn. Lett.* **14**, 303–315 (1993)

35. Ulmann, J.R.: Experiments with the n -tuple method of pattern recognition. *IEEE Trans. Comput.* **18**(12), 1135–1137 (1969)
36. Vinciarelli, A.: A survey on off-line cursive word recognition. *Pattern Recogn.* **35**, 1433–1446 (2002)
37. Xiao, X., Leedham, G.: Cursive script segmentation incorporating Knowledge of writing. In: *Proceedings of the 5th International Conference on Document Analysis and Recognition*, pp. 535–538 (1999)
38. Xia, F.: Normal vector and winding number in 2D digital images with their application for hole detection. *Pattern Recogn.* **36**, 1383–1395 (2003)
39. Xu, Q., Lam, L., Suen, C.Y.: A knowledge-based segmentation system for handwritten dates on bank cheques. In: *Sixth International Conference on Document Analysis and Recognition, ICDAR2001*, pp. 384–388 (2001)
40. Zhang, M., Suen, C.: *Digital Image Processing*, 2nd edn, pp. 398–402 (1987)