An Old Greek Handwritten OCR System

K. Ntzios, B. Gatos, I. Pratikakis,T. Konidaris and S. J. Perantonis
Computational Intelligence Laboratory, Institute of Informatics and Telecommunications, National Research Center "Demokritos",
153 10 Athens, Greece
{ntzios,bgat,ipratika,tkonid,sper}@iit.demokritos.gr
http://www.iit.demokritos.gr/cil

Abstract

Recognition of handwritten manuscripts is essential for efficient content exploitation of the valuable Old Greek historical collections. In this paper, we focus on the problem of recognizing Old Greek handwritten manuscripts and propose a novel recognition technique that can be applied to a large number of important historical manuscript collections which are written in lower case 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 recognition 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.

1. Introduction

In the field of handwriting recognition a great progress has occurred during the past years [1]. In methodology, two general approaches can be identified: the segmentation approach [2][3] and the global or segmentation-free approach [4][5]. The segmentation approach requires that each word has to be segmented into characters while the global approach entails the recognition of the whole word. Although the global approach is referred to the literature as "segmentation-free" approach, it involves 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 [6]. 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 [7][8], peephole method [9], n-tuple feature extraction [10], and hit-or-miss operator [11].

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. These constitute 95% of the complete character set used in a typical old Greek manuscript. (see Figure 1b, 1c).

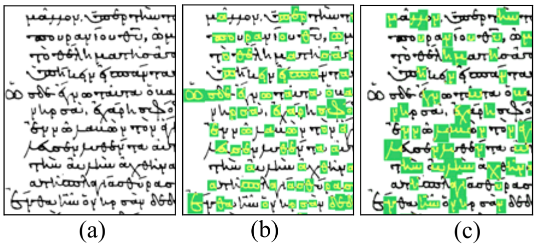


Figure 1. (a) Old Greek manuscript; (b) Identified characters or character ligatures that contain closed cavities; (c) Identified characters or characters ligatures that contain open cavities.

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, the open and closed cavities of the skeletonized characters are detected and a feature extraction scheme is applied. Finally the individual cavities are recognized on the basis of their features.

2 Image Binarization, Enhancement and Skeletonization

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. The proposed scheme for image binarization and enhancement is fully described in [12].

For the skeletonization process, we use an iterative method presented in [13]. This method is an extension of Zhang and Suen's method [14].

3 Open and closed cavities detection

In this step, open and closed cavities are detected in the skeletonized image. For the closed cavities detection, we used a novel fast algorithm based on processing the white runs of the binary image, described in [15]. For the open cavities detection the water reservoir principle is used [16]. 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 (see Figure 2).

Some of the detected cavities 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 1/3 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:

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

where $mean(y_i^u)$ is the mean value of all y-coordinates of the pixels that compose the open cavity and $mean(y_j^o)$ is the mean value of all y-coordinates of the pixels that compose the neighbor closed cavity. In Figure 2 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:

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

where, min(.), max(.), mean(.) denote the minimum value, the maximum value and the average value, of the set *X* or *Y*, respectively. Figure 2 shows the skeletonized components with the corresponding bounding box *W* around each open and closed cavity.

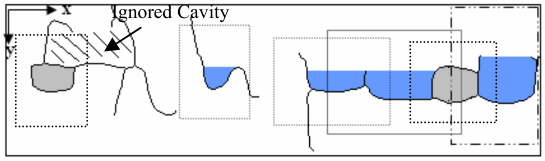


Figure. 2 The skeletonized components with the respective bounding box around each open and closed cavity.

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 $\mathcal{F} = \{f_1, f_2, \ldots, f_{15}\}$. 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} and f_{12} denote the slope of upper and lower protrusible segments, respectively. Features f_{13} and f_{14} denote the length of segments that appear in the block R_{13} and R_{14} as depicted in Figures 3a and 3b, while f_{15} denotes the opening angle of the cavity. This angle is constructed as in the following: We first determine points *A, B* denote the intersection points at the cavity and the horizontal line at a height n/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 $A\hat{P}B$ angle (see Figure 4).

Feature estimation is employed in the following two steps.

- *Step 1: Bounding box division into blocks*

Let $\mathcal{H}$={($x_i^{\mathcal{H}}, y_i^{\mathcal{H}}$), $i \in [1,n]$ }, be the set of the skeleton 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 horizontal lines $F_1(x)=min(y_i)$ and $F_2(x)=max(y_i)$ as it shown in Figure 3a, resulting in 9 blocks from which only 7 ($R_1,\ldots,R_6$ and R_{13}) are taken under consideration since in these areas distinguishing protrusions are expected. Furthermore, for the horizontal mode we divide *W* into two horizontal areas equal width and assign two divide lines $F_1(y)=min(x_i)$ and $F_2(y)=max(x_i)$ as it is shown in Figure 3b, resulting in extra 6 blocks from which only 5 ({R_7, R_8, R_9, R_{10}, R_{14}}) are taken under consideration.

- *Step 2: Block-based feature computation*

In this step, we estimate the length of a protrusion by a clockwise 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 skeleton pixel coordinates depicted in block R_i and meanwhile they don't comprise pixel of the cavity. For each pixel *j* of H_{Ri} 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 clockwise tracing. Once the directions are evaluated the proposed feature f_i is defined as follows:

$$f_i = \frac{1}{D}\sum_{j=1}^{m_i} g_i(s_j^i),\ i \in [1,14] \qquad (3)$$

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. They determine a unique pixel template for each block R_i that expresses the expected local orientation of the corresponding pixels.

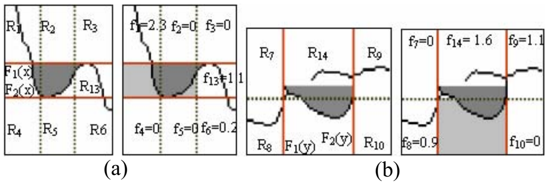


Figure 3. (a)Vertical mode for the Greek letter "η". Left: Blocks $R_1 \ldots R_6$ and R_{13} delimited by F_1(x) and F_2(x). Right: An example for features $f_1 \ldots f_6$ and f_{13} valuation. For this character f_{11}=-0.13, f_{12}=0 and f_{15}=126° (b). Horizontal mode for the Greek letter "σ". Left: Blocks $R_7 \ldots R_9$ and R_{14} delimited by F_1(y) and F_2(y). Right: An example for features $f_7 \ldots f_{10}$ and f_{14} valuation. For this character f_{15}=153°.

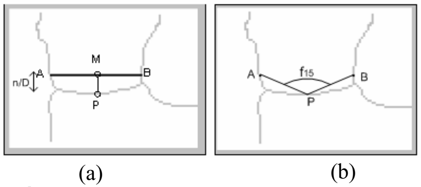


Figure 4. Feature f_{15}. (a) Points determination. (b) The opening angle $A\hat{P}B$.

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

	E	NE	N	NW	W	SW	S	SE
$g_{1\text{-}3}$	0	1	1	1	0	0	0	0
$g_{4\text{-}6}$	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

↖	N	↗
W	●	E
↙	S	↘

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 to 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 not to be 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. All the skeleton pixels that are involved in this feature extraction phase are marked in order not to be 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 not marked pixels (see Figure 5).

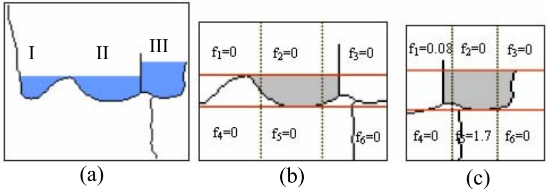


Figure 5. (a) A connected component with three open cavities. (b) The features of cavity II. Although at the blocks R_3 and R_6 there are protrusible segments the respective features are 0 because these segments belong to cavity III. (c) The features of cavity III.

4.4 Cavity Merging

In this stage, two or more closed cavities are merged when they share a common boundary. The merged closed cavities can be characterized as a character or character ligature having (i) two, three or four closed cavities arranged in an horizontal axes, (ii) two closed cavities arranged in a vertical axes and (iii) closed cavities arranged in both horizontal and vertical axes (see Table 2). Therefore, when two closed cavities i and j, have common pixels, the merged character is characterized as horizontal if Eq. 4 is true, otherwise it is characterized as vertical.

$$\begin{aligned}&\min\left(\left|\max(y_i)-\max(y_j)\right|,\left|\min(y_i)-\min(y_j)\right|\right)<\\&\min\left(\left|\max(y_i)-\min(y_j)\right|,\left|\min(y_i)-\max(y_j)\right|\right)\end{aligned} \tag{4}$$

where $max(y_i)$ and $min(y_i)$ is the maximum and minimum y-coordinate of ith closed cavity and $max(y_j)$ and $min(y_j)$ is the maximum and minimum y-coordinate of jth closed cavity.

Moreover, in this stage two or more open cavities that have a common boundary and they don't 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, pattern ID 8,9). After merging, the features of the resulting cavity are re-estimated in the single merged cavity.

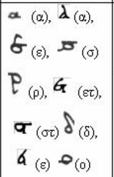
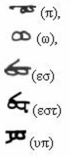
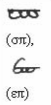
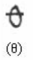
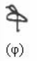
5 Character Recognition

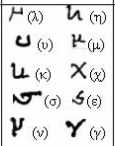
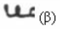
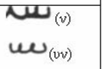
The character recognition process consists of two basic stages. In the first stage each character is classified into a pattern by their spatial configuration as shown in Table 2. For example, the characters that have one closed cavity are classified to the pattern with ID 1 and the characters with one 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. An example decision tree for ID 1, where all conditions upon which a tree traversal is progressing, can be shown in Figure 6. The corresponding threshold values T_i, which support the required conditioning is computed in the following:

$$T_i = \underset{f_i^j \neq 0}{mean}(f_i^j) \quad i \in [1, 15],\ j \in C \tag{5}$$

where C is the set of cavities.

Table 2. The proposed dictionary for open and closed cavity patterns.

Pattern ID	1	2	3	4	5	6
Characters or character ligatures with closed cavities	 (α), (α), (ε), (σ) (ρ), (ετ), (στ) (δ), (ε) (ο)	 (π), (ω), (εσ) (εστ) (υπ)	 (σπ), (επ)	(απ)	 (θ)	 (φ)

Pattern ID	7	8	9
Characters or character ligatures with open cavities	 (λ) (η) (υ) (μ) (κ) (χ) (σ) (ε) (ν) (γ)	 (β)	 (ν) (υν)

6 Experimental results

Our experiments aim to test the classification performance of the proposed handwritten manuscripts recognition procedure. The overall experimental samples originate from different manuscripts of the Book of Job collection, manually labeled with the ground truth. We have built a set of open and closed cavity patterns that contains a total of 3886 characters and character ligatures. Table 3 shows the results obtained by applying the algorithm, in terms of recall and precision rates for each one of the most frequent character appearing in the documents. Recall is the correct number of open and closed cavities classified divided by the total number of the open and closed cavities. Precision is the number of correct open and open cavities classified divided by the total number of open and closed cavities classified. Our system recognizes all characters having open or closed cavities with an average recall of 90.74% and precision of 88.85%.

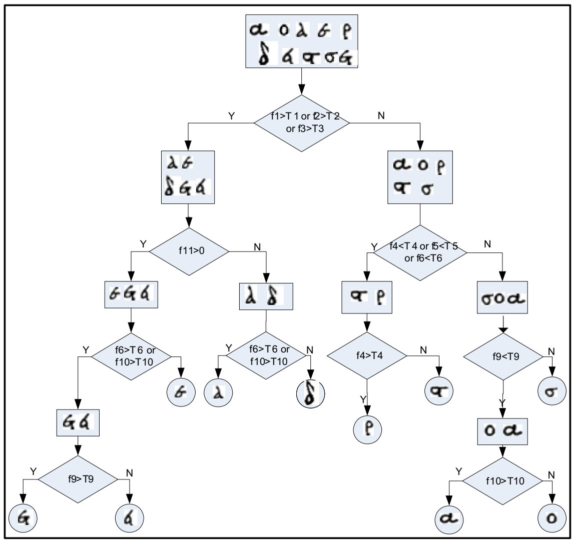


Figure 6. Classification tree for characters of pattern with ID 1.

Table 3. Precision and Recall for most frequent appearing characters

		Number of characters	Recall	Precision
	α1	557	95,33	98,94
	ε1	231	94,37	97,76
	ν	392	91,58	95,48
	ο	502	99,40	97,84
	σ	255	94,90	86,43
	υ	221	98,19	68,24

7 Conclusions

In this paper, we present a novel methodology for recognizing Old Greek handwritten manuscripts. 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 the proposed system combined with an efficient post-processing lexicon technique will further improve the recognition accuracy.

8 References

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

[2] B. Eastwood, A. Jennings A and A. Harvey, "A Feature Based Neural Network Segmenter for Handwritten Words", International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'97), Australia, 1997, pp. 286-290.

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

[4] D. Guillevic and C.Y. Suen, "HMM Word Recognition Engine", Fourth International Conference on Document Analysis and Recognition ICDAR97, 1997, pp. 544.

[5] C.Y. Suen, "Building a New Generation of Handwriting Recognition Systems". Pattern Recognition Letters, 14, 1993, pp. 303-315.

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

[7] A. Amin and G. Masini "Machine recognition of cursive Arabic words", Application of Digital Image Processing IV, San Diego, CA, Vol SPIE-359, August 1982,pp. 286-292.

[8] R. Duda and E. Hart "Pattern Classification and Scene Analysis", Wiley 1973.

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

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

[11] R.C. Gonzalez and R.E. Woods "Digital Image Processing", Addison-Wesley, 2003.

[12] B. Gatos, I. Pratikakis, S. Perantonis, "An adaptive binarisation technique for low quality historical documents", IAPR Workshop on Document Analysis systems (DAS'2004), Lecture Notes in Computer Science (3163), Florence, Italy, September 2004, pp. 102-113.

[13] H.J. Lee, B. Chen "Recognition of Handwritten Chinese Characters via Short Line Segments", Pattern Recognition 25 (5), 1992, pp. 543-552.

[14] M. Zhang and C. Suen, "Digital Image Processing", 2nd edition, 1987, pp. 398-402.

[15] B. Gatos, K. Ntzios, I. Pratikakis, 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), Florence, Italy, September 2004, pp. 63-74.

[16] U. Pal, A. Belaid and C. Choisy "Touching numeral segmentation using water reservoir concept", Pattern recognition Letters, vol. 24, 2003, pp. 261-272.