Greek Handwritten Character Recognition

G. Vamvakas, N. Stamatopoulos, B. Gatos, I. Pratikakis and S. J. Perantonis

Computational Intelligence Laboratory,
Institute of Informatics and Telecommunications,
National Center for Scientific Research "Demokritos",
GR-153 10 Agia Paraskevi, Athens, Greece
http://www.iit.demokritos.gr/cil/,
{gbam, nstam, bgat, ipratika, sper,}@iit.demokritos.gr

Abstract

In this paper, we present a database and methods for off-line isolated Greek handwritten character recognition. The Computational Intelligence Laboratory (CIL) Database consists of 35,000 isolated and labelled Greek handwritten characters. This database was tested with an existing structural approach for Greek handwritten characters as well as with a novel approach based on a hybrid feature extraction scheme. According to this approach, two types of features are combined in a hybrid fashion. The first one divides the character image into a set of zones and calculates the density of the character pixels in each zone. In the second type of features, the area that is formed from the projections of the upper and lower as well as of the left and right character profiles is calculated. For the classification step, Support Vectors Machines (SVM) and Euclidean Minimum Distance Classifier (EMDC) are used.

Keywords: Handwritten OCR, Feature Extraction, Greek Handwritten Character Database

1. Introduction

Optical Character Recognition (OCR) systems attempt to facilitate the everyday use of computers aiming at the transformation of large amounts of documents, either printed or handwritten, into electronic form for further processing. Many applications, such as postal address reading, bank checking recognition, write identification, reading for the visually handicapped etc, use OCR systems. Nowadays, the recognition of printed isolated characters is performed with high accuracy. However, the recognition of handwritten characters still remains an open problem in the research arena. In general, character recognition procedure consists of two steps: (a) feature extraction, where each character is represented as a feature vector and (b) classification of these feature vectors into a number of classes [Brito 2004].

Selection of a feature extraction method is probable the single most important factor in achieving high recognition performance [Trier 1996]. Due to the nature of handwriting with its high degree of variability and imprecision, obtaining these features is a difficult task. A feature extraction algorithm must be robust enough that for a variety of instances of the same symbol, similar feature sets are generated,

thereby making the subsequent classification task less difficult [Fitzgerald 2004]. In the literature, feature extraction methods have been based on two types of features: statistical and structural [Covindan 1990]. Representation of a character image by statistical distribution takes care of style variations to some extent. This method is used for reducing the dimension of the feature set providing high speed and low complexity [Arica 2001]. On the other hand, characters can be represented by structural features with high tolerance to distortions and style variations.

The most common statistical features used for character representation are: (a) zoning, where the character is divided into several zones and features are extracted from the densities in each zone [Luiz 2002] or from measuring the direction of the contour of the character by computing histograms of chain codes in each zone [Mohiuddin 1994], (b) projections and (c) crossings and distances. In [Koerich 2003], the projection profiles at four directions (top, bottom, left, right) are used and in [Kim 2000], two sets of features are extracted using crossings and distances. The first one consists of the number of transitions from background to foreground pixels along vertical and horizontal lines through the character image and the second one calculates the distances of the first image pixel detected from the upper and lower boundaries, of the image, along vertical lines and from the left and right boundaries along horizontal lines.

Structural features are based on topological and geometrical properties of the character, such as reference lines, ascenders, descenders, cross points, branch points, strokes and their directions, etc. One of the most popular techniques for structural feature extraction is coding which is obtained by mapping the strokes of a character into o 2-D parameter space, which is made up of codes. In [Kavallieratou 2002], a structural approach for recognizing Greek handwritten characters is introduced. A 280-dimension vector is extracted consisting of histograms and profiles. The well known horizontal and vertical histograms are used in combination with the radial histogram, out-in radial and in-out radial profiles. The radial histogram is defined as the sum of foreground pixels on a radius that starts from the centre of the image and ends up at the border. The value of the out-in radial profile is defined as the position of the first foreground pixel found on the radius that starts from the periphery and goes to the centre of the character image forming an angle φ with the horizontal axis. Similarly, the value of the in-out profile is defined as the position of the first foreground pixel found on the radius that starts from centre of the character image and goes to the periphery forming an angle φ with the vertical axis. This is also, to the best of our knowledge, the only work for recognizing Greek handwritten documents.

An essential part of the development and evaluation of every off-line character recognition technique is the comparison of results by using the same standard database as other researchers [Guyon 1997]. There are many examples of widely used databases for handwriting recognition, such as NIST [Wilkinson 1992], CEDAR [Hull 1994], CENPARMI [Suen 1992] and UNIPEN [Guyon 1994].As far as Greek

characters are concerned the only database one can find in the literature is the GRUHD database [Kavallieratou 2001] that includes Greek characters, texts, digits and other symbols in unconstrained handwriting mode.

In this paper we present the Computation Intelligence Laboratory (CIL) Database consisting of 35,000 isolated Greek Handwritten Characters. Furthermore, an off-line OCR methodology for these letters is proposed based on a hybrid feature extraction scheme. The remaining of the paper is organized as follows. In Sections 2 and 3 the data acquisition procedure and the proposed feature extraction method are presented respectively. Experimental results are discussed in Section 4 and finally conclusions are drawn in Section 5.

2. The CIL Database

This database was created by handling a number of forms (Fig. 1) that include 56 Greek characters: 24 uppercase, 24 lowercase, 7 accented vowels and the final 'ς'. Initially, 125 writers are requested to fill these forms and afterwards, a scanner converts them to digital binary images. Then, we detect the vertical and horizontal lines. The form is being scanned horizontally and vertically and when the consecutive black pixels exceed a predefined number, then it is presumed that a line has been detected. However, a problem that may occur, if the image is skewed, is that a line may not be detected (Fig. 2).

Specifically, if D_{min} is the minimum number of consecutive black pixels and d the width of the line, then, the maximum angle of skew, for which the line can be detected, is

$$\phi_{\max} = \tan^{-1} \frac{d}{D_{\min}} \tag{1}$$

So, we note, that the maximum acceptable angle of skew depends from the line's width and D_{min}.

When the lines' detection has been completed, then the cells in which the characters are written can be defined. A simple method is to define the co-ordinates of the cells from the lines' and columns' intersections. But if the lines are slightly skewed, then the co-ordinates of the intersections are incorrectly calculated with, perhaps, severe effects at the extraction of the character from the cell (e.g. part of the character, which will be located at the boundaries of the defined "cell", can be fragmented). To avoid this problem, we redefine the co-ordinates for every cell separately, starting from the centre of the initial cell and scanning upwards until a line is detected (similarly downwards).

When the co-ordinates of the cell have been defined, then the bounding box containing the character inside the cell is extracted and stored in the database as binary image.

Every writer filled 5 forms, resulting in a database of 35,000 isolated and labelled characters. Each class represents a character and consists of 625 variations of this character (56x625=35,000).

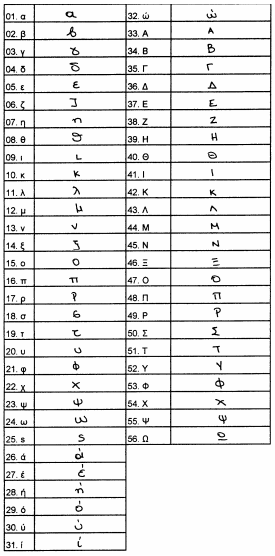


Figure 1: *Sample of the forms used for the CIL Greek Characters Database.*

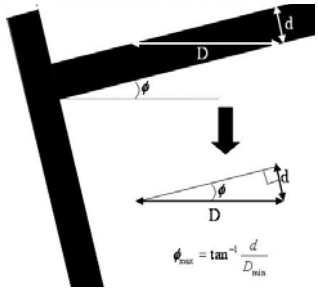


Figure 2: *Error in line detection*

3. Proposed Feature Extraction Method

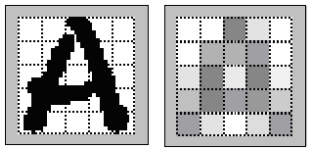
3.1 Pre-processing

Before the feature extraction algorithm takes place, we first normalize all binary character images to a NxN matrix. After normalization, slope correction is performed, which is based on [Buse 1997]. The dominant slope of the character is found from the slope corrected character which gives the minimum entropy of a vertical projection histogram. The vertical histogram projection is calculated for a range of slope correction angles, α_i, where the angle is given in $\pm\ \theta$. A slope correction range of θ=60^o appears to cover all writing styles.

3.2 Feature Extraction

In our approach we employ two types of features. The first one divides the character image into a set of zones and calculates the density of the character pixels in each zone. In the second type of features, the area that is formed from the projections of the upper and lower as well as of the left and right character profiles is calculated.

Let *im(x,y)* be the character image array having 1s for foreground and 0s for background pixels, x_{max} and y_{max} be the width and the height of the character image.



(a) ***(b)***

Figure 3. *(a) The normalized character image, (b) Features based on zones. Darker squares indicate higher density of character pixels.*

In the case of features based on zones, the image is divided into horizontal and vertical zones. For each zone, we calculate the density of the character pixels (Fig. 3). Let Z_H and Z_V be the total number of zones formed in both horizontal and vertical direction. Then, features based on zones $f^z(i)$, $i=0\ldots Z_H Z_V - 1$ are calculated as follows:

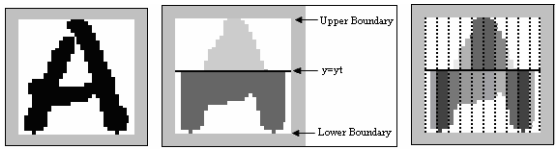
$$f^z(i) = \sum_{x=x_s(i)}^{x_e(i)} \sum_{y=y_s(i)}^{y_e(i)} im(x,y) \tag{2}$$

where,

$$x_s(i) = (i - \left\lfloor \frac{i}{Z_H} \right\rfloor Z_H)\frac{x_{\max}}{Z_H},\ x_e(i) = (i - \left\lfloor \frac{i}{Z_H} \right\rfloor Z_H + 1)\frac{x_{\max}}{Z_H},\ y_s(i) = \left\lfloor \frac{i}{Z_H} \right\rfloor \frac{y_{\max}}{Z_V},\ y_e(i) = (\left\lfloor \frac{i}{Z_H} \right\rfloor + 1)\frac{y_{\max}}{Z_V}$$

In case of features based on character (upper/lower) profile projections, the character image is divided into two sections separated by the horizontal line $y = y_t$ (Eq. 3):

$$y_t = \frac{\sum_x \sum_y im(x,y) \cdot y}{\sum_x \sum_y im(x,y)} \tag{3}$$



(a) ***(b)*** ***(c)***

Figure 4. *(a) The normalized character image. (b) Upper and lower character profiles. (c) The extracted features. Darker squares indicate higher density of zone pixels*.

Upper/lower profiles (Eq. 4, 5) are computed by considering, for each image column, the distance between the horizontal line $y=y_t$ and the closest pixel to the upper/lower boundary of the character image (Fig. 4):

$$y_{up}(x) = y_t - y_0, \tag{4}$$

$$\text{where } y_0 = \begin{cases} y_t, \text{ if } \sum_{y=0}^{y_t} im(\mathrm{x},\mathrm{y}) = 0 \\ y : (im(\mathrm{x},\mathrm{y}) = 1 \,\&\, \mathrm{y} = \min(\mathrm{y_i})), \mathrm{y_i} \in [0, \mathrm{y_t}], \text{else} \end{cases}$$

$$y_{lo}(x) = y_0 - y_t, \tag{5}$$

$$\text{where } y_0 = \begin{cases} y_t, \text{ if } \sum_{y=y_t}^{y_{max}} im(x,y) = 0 \\ y : (im(x,y) = 1 \,\&\, y = \max(y_i)), y_i \in [y_t, y_{max}], \text{else} \end{cases}$$

Let P_V be the total number of blocks formed in each produced zone (upper, lower). For each block, we calculate the area of the upper/lower character profiles denoted as in the following:

$$f^P{}_{up\_ar}(i) = \sum_{x=x_s(i)}^{x_e(i)} y_{up}(x) \tag{6}$$

$$f^P{}_{lo\_ar}(i) = \sum_{x=x_s(i)}^{x_e(i)} y_{lo}(x) \tag{7}$$

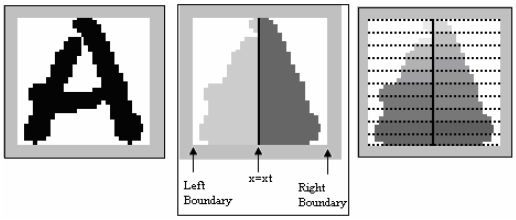
where,

$$x_s(i) = (i - \left\lfloor \frac{i}{P_V} \right\rfloor P_V) \frac{x_{\max}}{P_V} \;,\; x_e(i) = (i - \left\lfloor \frac{i}{P_V} \right\rfloor P_V + 1) \frac{x_{\max}}{P_V}$$

and $i=0 .. P_V-1$. Fig. 4 illustrates the features extracted from a character image using projections of character profiles.

In case of features based on character (left/right) profile projections, the character image is divided into two sections separated by the vertical line $x = x_t$ (Eq. 8):

$$x_t = \frac{\sum_x \sum_y im(x,y) \cdot x}{\sum_x \sum_y im(x,y)} \tag{8}$$



(a) **(b)** **(c)**

Figure. 5: *(a) The normalized character image. (b) Left and right character profiles; (c) The extracted features. Darker squares indicate higher density of zone pixels.*

Left/right character profiles (Eq. 9,10) are computed by considering, for each image column, the distance between the vertical line $x=x_t$ and the closest character pixel to the left/right boundary of the character image (Fig. 5):

$$x_{le}(y) = x_t - x_0, \tag{9}$$

$$\text{where } x_0 = \begin{cases} x_t, \text{ if } \sum_{x=0}^{x_t} im(x,y) = 0 \\ x : (im(x,y) = 1 \,\&\, x = \min(x_i)), x_i \in [0, x_t], \text{else} \end{cases}$$

$$x_{ri}(y) = x_0 - x_t, \tag{10}$$

$$\text{where } x_0 = \begin{cases} x_t, \text{ if } \sum_{x=x_t}^{x_{\max}} im(x,y) = 0 \\ x : (im(x,y) = 1 \,\&\, x = \max(x_i)), x_i \in [x_t, x_{\max}], \text{else} \end{cases}$$

Let R_V be the total number of blocks formed in each produced zone (left, right). For each block, we calculate the area of left/right character profiles denoted as in the following:

$$f^{R}{}_{le\_ar}(i) = \sum_{y=y_s(i)}^{y_e(i)} x_{le}(y) \tag{11}$$

$$f^{R}{}_{ri\_ar}(i) = \sum_{y=y_s(i)}^{y_e(i)} x_{ri}(y) \tag{12}$$

where,

$$y_s(i) = (i - \left\lfloor \frac{i}{R_V} \right\rfloor R_V) \frac{y_{\max}}{R_V} \quad , \quad y_e(i) = (i - \left\lfloor \frac{i}{R_V} \right\rfloor R_V + 1) \frac{y_{\max}}{R_V}$$

and $i=0 .. R_V-1$. Fig. 5 illustrates the features extracted from a character image using projections of character profiles.

The overall calculation of the proposed feature vector is given in Eq. 13. The corresponding feature vector length equals to $Z_H Z_V + 2P_{V+} 2R_V$.

$$f(i) = \begin{cases} f^{c}(i) = \sum_{x=x_s(i)}^{x_e(i)} \sum_{y=y_s(i)}^{y_e(i)} im(x,y), i = 0..Z_H Z_V - 1 \\ f^{P}{}_{up\_ar}(i) = \sum_{x=x_s(i-Z_H Z_V)}^{x_e(i-Z_H Z_V)} y_{up}(x), i = Z_H Z_V .. Z_H Z_V + P_V - 1 \\ f^{P}{}_{lo\_ar}(i) = \sum_{x=x_s(i-Z_H Z_V + P_V)}^{x_e(i-Z_H Z_V + P_V)} y_{lo}(x), i = Z_H Z_V + P_V .. Z_H Z_V + 2P_V - 1 \\ f^{R}{}_{le\_ar}(i) = \sum_{y=y_s(i-Z_H Z_V + 2P_V)}^{y_e(i-Z_H Z_V + 2P_V)} x_{le}(y), i = Z_H Z_V + 2P_V .. Z_H Z_V + 2P_V + R_V - 1 \\ f^{R}{}_{ri\_ar}(i) = \sum_{y=y_s(i-Z_H Z_V + 2P_V + R_V)}^{y_e(i-Z_H Z_V + 2P_V + R_V)} x_{ri}(y), i = Z_H Z_V + 2P_V + R_V .. Z_H Z_V + 2P_V + 2R_V - 1 \end{cases} \tag{13}$$

4. Experimental Results

For our experiments the CIL Database was used. After the size normalization step some characters such as the upper-case ‘O’ and the lower-case ‘o’, are considered to be the same. So, for having meaningful results we merged these two classes into one, by randomly selecting 625 characters from both classes. This was done to a total of 10 pair of classes. Table 1 shows which classes are merged. This concluded in having 46 classes with 625 patterns in each class and the database now has 28,750 characters. Moreover, 1/5 of each class was used for testing and the 4/5 for training. So the used database was split into a training set of 23,000 characters and a testing set of 5,750 characters.

As it has already been described in Sections 4 we have used a size normalization step followed by a slope correction step before feature extraction. During the normalization step, the size of the normalized character images used is x_{max}=60 and y_{max}=60. In the case of features based on zones, the character image is divided into 5 horizontal (Z_H =5) and 5 vertical (Z_V =5) zones forming 25 blocks with size 12x12 (Fig. 3). Therefore, the total number of features is 25. In the case of features based on character (upper/lower) profile projections we keep the same size of the normalized image, while the image is divided into 10 vertical zones (P_V =10) (Fig. 4). Consequently, the total number of features equals to 20. Similarly, the normalized image divided into 10 horizontal zones (R_V =10) (Fig. 5). Therefore, the total number of features equals to 20. Combination of features based on zones and features based on character profile projections led to the feature extraction model (Eq. 13) that uses a total of 65 features.

Table 1: *Merged Classes*

	Upper-case	**Lower-case**
1	Ε	ε
2	Θ	θ
3	Κ	κ
4	Ο	ο
5	Π	π
6	Ρ	ρ
7	Τ	τ
8	Φ	φ
9	Χ	χ
10	Ψ	ψ

In the particular classification problem classification step was performed using two well-known classification algorithms, Euclidean Minimum Distance Classifier (EMDC) [Theodoridis 1997] and Support Vector Machines (SVM) [Cortes 1997] with Radial Basis Function (RBF).

Table 2 depicts the recognition rates (%) for the proposed hybrid method, the single feature methods and the method described in [Kavallieratou 2002], which is the only method reported in the literature for the recognition of Greek handwritten characters. These recognition rates achieved after combining slope correction with either single features of both feature extraction schemes. We can draw several conclusions. First, as it can been easily seen the best performance is achieved in the case of using an additive fusion resulted after the combination of slope correction preceding the hybrid feature extraction scheme. Moreover, our approach seems to have better results although the total number of features used (65) is smaller than the one in [Kavallieratou 2002] (280). Finally, the SVM achieved considerable higher recognition rates than the Euclidean Minimum Distance Classifier (EMDC).

Table2: *Experimental Results*

Pre-processing	Types of Features			Number of features	Classifier		Recognition Rate (%)
Slope Correction	Method in [Kavallieratou 2002]	Proposed Method			EMDC	SVM	
		Zones	Projections				
	√			280	√		81.36%
√	√			280	√		81.20%
		√		25	√		85.94%
			√	40	√		76.80%
		√	√	65	√		83.44%
√		√		25	√		85.36%
√			√	40	√		78.46%
√		√	√	65	√		84.55%
	√			280		√	87.52%
√	√			280		√	88.62%
		√		25		√	88.29%
			√	40		√	87.56%
		√	√	65		√	90.12%
√		√		25		√	88.48%
√			√	40		√	87.75%
√		√	√	65		√	**91.61%**

5. Conclusions

This paper presents a new database of isolated Greek handwritten characters (CIL Database). This database was tested with an existing structural approach for Greek handwritten characters [Kavallieratou 2002] as well as with a novel approach based on a hybrid feature extraction scheme. The proposed approach seems to have better results although it uses smaller feature vector.

Our future research, apart from expanding the database, will focus on exploiting new features as well as fusion methods to further improve the current performance.

References

A. S. Britto, R. Sabourin, F. Bortolozzi, C.Y.Suen (2004), *Foreground and Backround Information in an HMM-Based Method for Recognition of Isolated Characters and Numeral Strings*, 9th IWFHR , pp 371-376.

O. D. Trier, A. K. Jain, T. Taxt (1996), *Features Extraction Methods for Character Recognition – A Survey*, Pattern Recognition, 29(4): 641-662.

J. A. Fitzgerald, F. Geiselbrechtinger, and T. Kechadi (2004), *Application of Fuzzy Logic to Online Recognition of Handwritten Symbols*, 9th IWFHR, pp. 395-400.

V.K. Covindan, A.P. Shivaprasad (1990), *Character Recognition – A Review*, Pattern Recognition, 23(7), pp. 671- 683.

N. Arica and F. Yarman-Vural (2001), *An Overview of Character Recognition Focused on Off-line Handwriting*, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 31(2), pp. 216 - 233.

Luiz S. Oliveira, F. Bortolozzi, C.Y.Suen (2002), *Automatic Recognition of Handwritten Numerical Strings: A Recognition and Verification Strategy*", IEEE Transactions on Pattern Recognition and Machine Intelligence, Vol. 24, No. 11.

K. M. Mohiuddin and J. Mao (1994), *A Comprehensive Study of Different Classifiers for Handprinted Character Recognition*. Pattern Recognition, Practice IV, pp. 437-448.

A. L. Koerich (2003), *Unconstrained Handwritten Character Recognition Using Different Classification Strategies*. IAPR International Workshop on Artificial Neural Networks in Pattern Recognition (ANNPR 2003).

J. H. Kim, K. K. Kim, C. Y. Suen (2000), *Hybrid Schemes Of Homogeneous and Heterogeneous Classifiers for Cursive Word Recognition*, 7th IWFHR, Amsterdam, pp 433 - 442.

E. Kavallieratou, N. Fotakis, G Kokkinakis (2002), *Handwritten Character Recognition Based on Structural Characteristics*, ICPR, p. 30139, 16th International Conference on Pattern Recognition (ICPR'02) – Vol. 3.

I. Guyon, R. Haralick, J. Hull, and I. Phillips (1997), *Database and benchmarking*, In H. Bunke and P. Wand, editors, Handbook of Character Recognition and Document Image Analysis. World Scientific, Chapter 30, pp. 779-799.

R. Wilkinson, J. Geist, S. Janet, P. Grother, C. Burges, R. Creecy, B. Hammond, J. Hull, N. Larsen, T. Vogl, and C. Wilson, (1992). The first census optical character recognition systems conf. #NISTIR 4912, The U.S. Bureau of Census and the National Institute of Standards and Technology, Gaithersburg, MD.

J. Hull (1994), *A database for handwritten text recognition research*, IEEE Trans. on Pattern Analysis and Machine Intelligence, Volume 16, Issue 5.

C. Y. Suen, C. Nadal, R. Legault, T. Mai, and L.Lam (1992),*Computer recognition of unconstrained handwritten numerals*, Proc. of the IEEE,Volume 7,Issue 80,pp. 1162-1180.

I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, and S. Janet (1994), *Unipen project of on-line data exchange and benchmarks*, Proc. of the 12th IAPR Int. Conf on Pattern Recognition, Jerusalem, Israel, Oct. 1994, pp.29-33.

E.Kavallieratou, N.Liolios, E.Koutsogeorgos, N.Fakotakis, G.Kokkinakis (2001), *The GRUHD database of Modern Greek Unconstrained Handwriting*, In Proc. ICDAR.

R. Buse, Z.Q. Liu, and T. Caelli (1997), *A Structural and Relational Approach to Handwritten Word Recognition*, IEEE Trans.Systems,Man, and Cybernetics, Part B, 27(5), pp. 847-861.

Theodoridis, S., and Koutroumbas (1997), K., Pattern Recognition, (Academic Press).

Cortes C., Vapnik, V(1997),*Support-vector network*,Machine Learning,vol. 20,pp. 273-297.