GRUHD: A Greek Database of Unconstrained Handwriting

E.Kavallieratou, N.Liolios, E.Koutsogeorgos, N.Fakotakis, and G.Kokkinakis

Wire Communications Laboratory,
University of Patras, 26500 Patras, Greece.
Tel. ++30-61-991722, fax ++30-61-991855
(ergina,nliolios,junior,fakotaki,kokkinaki)@wcl.ee.upatras.gr

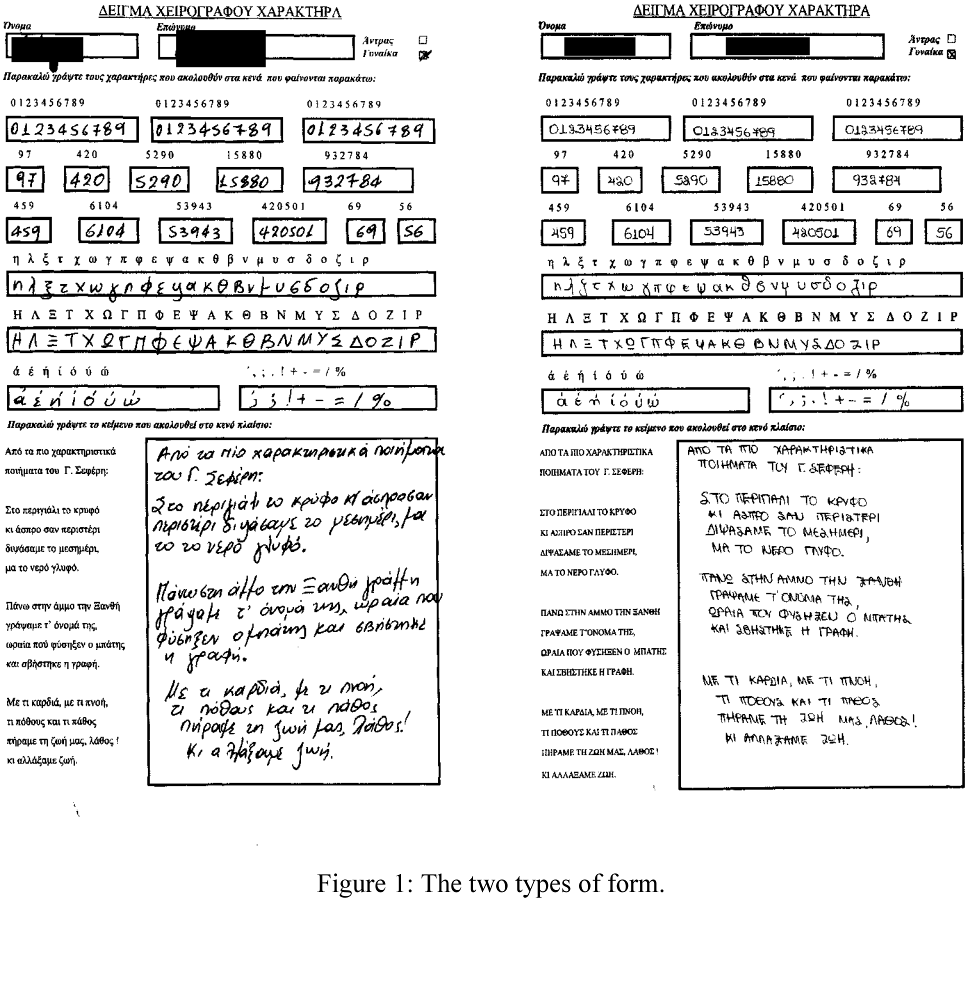
Abstract

In this paper we present the GRUHD database of Greek characters, text, digits, and other symbols in unconstrained handwriting mode. The database consists of 1,760 forms that contain 667,583 handwritten symbols and 102,692 words in total, written by 1,000 writers, 500 men and equal number of women. Special attention was paid in gathering data from writers of different age and educational level. The GRUHD database is accompanied by the GRUHD software that facilitates its installation and use and enables the user to extract and process the data from the forms selectively, depending on the application. The various types of possible installations make it appropriate for the training and validation of character recognition, character segmentation and text-dependent writer identification systems.

1. Introduction

The research in Optical Character Recognition (OCR) has started in the early 1960s. A very crucial and still open problem is the evaluation of the proposed systems on the basis of common resources. Indeed, the majority of the researchers use their own data for training and testing. Therefore, the extraction of useful conclusions regarding the contribution of the proposed systems is a very difficult, if not inapplicable, task.

Only recently pubic domain resources have become available. One of the most famous databases at the moment is NIST (Wilkinson, 1992) that contains isolated characters, while a more recent one is IAM-DB (Marti, 1999) that contains full English sentences. Moreover, there have been created databases of handwritten numerals (Suen, 1992) aiming at specialized applications, such as recognition of postal code. In addition to the English databases, there are databases of other languages (Kim, 1993; Saito, 1985).



ΔΕΙΓΜΑ ΧΕΙΡΟΓΡΑΦΟΥ ΧΑΡΑΚΤΗΡΑ

Όνομα [redacted]
Επώνυμο [redacted]
Άντρας □
Γυναίκα ☒

Παρακαλώ γράψτε τους χαρακτήρες που ακολουθούν στα κενά που φαίνονται παρακάτω:

0123456789 | 0123456789 | 0123456789
0123456789 | 0123456789 | 0123456789

97 | 420 | 5290 | 15880 | 932784
97 | 420 | 5290 | 15880 | 932784

459 | 6104 | 53943 | 420501 | 69 | 56
459 | 6104 | 53943 | 420501 | 69 | 56

η λ ξ τ χ ω γ π φ ε ψ α κ θ β ν μ υ σ δ ο ζ ι ρ
ηλξτχωγπφεψακθβνμυσδοζιρ

Η Λ Ξ Τ Χ Ω Γ Π Φ Ε Ψ Α Κ Θ Β Ν Μ Υ Σ Δ Ο Ζ Ι Ρ
ΗΛΞΤΧΩΓΠΦΕΨΑΚΘΒΝΜΥΣΔΟΖΙΡ

ά έ ή ί ό ύ ώ
ά έ ή ί ό ύ ώ

΄ , ; . ! + - = / %
; ; ! + - = / %

Παρακαλώ γράψτε το κείμενο που ακολουθεί στο κενό πλαίσιο:

Από τα πιο χαρακτηριστικά
ποιήματα του Γ. Σεφέρη:

Στο περιγιάλι το κρυφό
κι άσπρο σαν περιστέρι
διψάσαμε το μεσημέρι,
μα το νερό γλυφό.

Πάνω στην άμμο την ξανθή
γράψαμε τ' όνομά της,
ωραία πού φύσηξεν ο μπάτης
και σβήστηκε η γραφή.

Με τι καρδιά, με τι πνοή,
τι πόθους και τι πάθος
πήραμε τη ζωή μας, λάθος !
κι αλλάξαμε ζωή.

Από τα πιο χαρακτηριστικά ποιήματα
του Γ. Σεφέρη:
Στο περιγιάλι το κρύφο κι άσπρο σαν
περιστέρι διψάσαμε το μεσημέρι, μα
το το νερό γλυφό.

Πάνω στην άμμο την Ξανθή γράψαμε
τ' όνομά της, ωραία που
φύσηξεν ο μπάτης και σβήστηκε
η γραφή.

Με τι καρδιά, με τι πνοή,
τι πόθους και τι πάθος
πήραμε τη ζωή μας, λάθος!
Κι αλλάξαμε ζωή.

ΔΕΙΓΜΑ ΧΕΙΡΟΓΡΑΦΟΥ ΧΑΡΑΚΤΗΡΑ

Όνομα [redacted]
Επώνυμο [redacted]
Άντρας □
Γυναίκα ☒

Παρακαλώ γράψτε τους χαρακτήρες που ακολουθούν στα κενά που φαίνονται παρακάτω:

0123456789 | 0123456789 | 0123456789
0123456789 | 0123456789 | 0123456789

97 | 420 | 5290 | 15880 | 932784
97 | 420 | 5290 | 15880 | 932784

459 | 6104 | 53943 | 420501 | 69 | 56
459 | 6104 | 53943 | 420501 | 69 | 56

η λ ξ τ χ ω γ π φ ε ψ α κ θ β ν μ υ σ δ ο ζ ι ρ
ηλξτχωγπφεψακθβνμυσδοζιρ

Η Λ Ξ Τ Χ Ω Γ Π Φ Ε Ψ Α Κ Θ Β Ν Μ Υ Σ Δ Ο Ζ Ι Ρ
ΗΛΞΤΧΩΓΠΦΕΨΑΚΘΒΝΜΥΣΔΟΖΙΡ

ά έ ή ί ό ύ ώ
ά έ ή ί ό ύ ώ

΄ , ; . ! + - = / %
΄ , ; . ! + - = / %

Παρακαλώ γράψτε το κείμενο που ακολουθεί στο κενό πλαίσιο:

ΑΠΟ ΤΑ ΠΙΟ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ
ΠΟΙΗΜΑΤΑ ΤΟΥ Γ. ΣΕΦΕΡΗ:

ΣΤΟ ΠΕΡΙΓΙΑΛΙ ΤΟ ΚΡΥΦΟ
ΚΙ ΑΣΠΡΟ ΣΑΝ ΠΕΡΙΣΤΕΡΙ
ΔΙΨΑΣΑΜΕ ΤΟ ΜΕΣΗΜΕΡΙ,
ΜΑ ΤΟ ΝΕΡΟ ΓΛΥΦΟ.

ΠΑΝΩ ΣΤΗΝ ΑΜΜΟ ΤΗΝ ΞΑΝΘΗ
ΓΡΑΨΑΜΕ Τ'ΟΝΟΜΑ ΤΗΣ,
ΩΡΑΙΑ ΠΟΥ ΦΥΣΗΞΕΝ Ο ΜΠΑΤΗΣ
ΚΑΙ ΣΒΗΣΤΗΚΕ Η ΓΡΑΦΗ.

ΜΕ ΤΙ ΚΑΡΔΙΑ, ΜΕ ΤΙ ΠΝΟΗ,
ΤΙ ΠΟΘΟΥΣ ΚΑΙ ΤΙ ΠΑΘΟΣ
ΠΗΡΑΜΕ ΤΗ ΖΩΗ ΜΑΣ, ΛΑΘΟΣ !
ΚΙ ΑΛΛΑΞΑΜΕ ΖΩΗ.

ΑΠΟ ΤΑ ΠΙΟ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ
ΠΟΙΗΜΑΤΑ ΤΟΥ Γ. ΣΕΦΕΡΗ:

ΣΤΟ ΠΕΡΙΓΙΑΛΙ ΤΟ ΚΡΥΦΟ
ΚΙ ΑΣΠΡΟ ΣΑΝ ΠΕΡΙΣΤΕΡΙ
ΔΙΨΑΣΑΜΕ ΤΟ ΜΕΣΗΜΕΡΙ,
ΜΑ ΤΟ ΝΕΡΟ ΓΛΥΦΟ.

ΠΑΝΩ ΣΤΗΝ ΑΜΜΟ ΤΗΝ ΞΑΝΘΗ
ΓΡΑΨΑΜΕ Τ'ΟΝΟΜΑ ΤΗΣ,
ΩΡΑΙΑ ΠΟΥ ΦΥΣΗΞΕΝ Ο ΜΠΑΤΗΣ
ΚΑΙ ΣΒΗΣΤΗΚΕ Η ΓΡΑΦΗ.

ΜΕ ΤΙ ΚΑΡΔΙΑ, ΜΕ ΤΙ ΠΝΟΗ,
ΤΙ ΠΟΘΟΥΣ ΚΑΙ ΤΙ ΠΑΘΟΣ
ΠΗΡΑΜΕ ΤΗ ΖΩΗ ΜΑΣ, ΛΑΘΟΣ!
ΚΙ ΑΛΛΑΞΑΜΕ ΖΩΗ.

Figure 1: The two types of form.

An OCR database has to fulfill certain criteria depending on the application. However, the handiness as well as the completeness are major demands.

Concerning the Greek language, the alphabet includes 21 characters which are different from the character of the other Latin alphabet: 10 uppercase (Γ, Δ, Θ, Λ, Ξ, Π, Σ, Φ, Ψ, Ω) and 11 lowercase (γ, δ, ζ, θ, λ, ξ, π, σ, φ, ψ). Moreover, the greek character are very often met in documents of mathematics, physics and other sciences. These peculiarities, as well as the different style of writing that these characters drive make necessary the creation of a Greek character database.

In this paper we present a Greek Unconstrained Handwriting Database (GRUHD), which to the best of our knowledge, is the only existing database of Modern Greek in this domain. At present, the GRUHD database includes 1,760 forms written by 1000 persons, about 667,583 symbols and 102,692 words in total.

The GRUHD database is accompanied by the GRUHD software that enables the user to extract and use the data from the forms selectively, depending on the application. Thus, both the characters and the words can be classified according to various criteria (e.g., writer, sex) or extracted as a whole. Moreover, the characters can be classified in ASCII code.

The different types of data organization allowed by the presented database makes it appropriate for the training and testing of a large number of applications, such as character recognition, character segmentation, text-dependent writer identification or verification systems.

The database has been used for the training and testing of the character segmentation system described in (Kavallieratou, 2000) as well as in the OCR system developed in the framework of the European project ACCeSS (LE-1 1802) that combines spoken and written language in call center applications.

The structure of the paper is as follows: The data acquisition and processing procedures are presented in section 2.1 and 2.2, respectively, while the data organization is described in section 3. Finally, some conclusions are drawn in section 4.

2. Description of the Database

A team of 15 persons worked for four months (about 2,400 man-hours in total) for the design and the creation of the GRUHD database. More than 1,000 persons were asked to fill the forms of fig.1. However, no restriction was set to the writers concerning their style of writing (slanted, connected, or hand-printed characters etc.). Hence, the result is a compilation of unconstrained handwriting samples.

2.1. Data Acquisition

As already mentioned the acquisition of the data succeeded by asking more than 1,000 persons to fill the forms of fig.1. These forms were designed in accordance with those of the NIST database. Both forms are similar and contain 19 fields each. The writers were asked to copy in these fields the symbols shown above or next to each field.

As far as the fields are concerned, the first 14 of them include groups of digits (totally 72 digits). The next two fields contain the 24 Greek alphabet characters, the first in uppercase and the second in lowercase, but in random order. The 17th and 18th fields concern the seven stressed characters of the Greek alphabet and some other symbols (5 punctuation marks and 5 arithmetic symbols), respectively. The above fields are common in both forms. Finally, the last field contains a very familiar Greek poem of 205 characters by the awarded with the Nobel prize Greek poet G.Seferis, written in uppercase in the one form and entirely in lowercase in the other. This poem was selected in order to encourage the persons to copy it without paying much attention, thus giving a more natural style to the writing. Moreover the specific poem contains all the 24 characters of the Greek alphabet. The information of each field is also given in table1.

The writers were asked to use a black or a blue pen and copy everything inside the boxes. No more restrictions were set concerning either the kind of pen or the style of writing.

Field	Kind of Symbols	Comments
1	10 digits	Ascending order
2	10 digits	Ascending order
3	10 digits	Ascending order
4	2 digits	Random order
5	3 digits	Random order
6	4 digits	Random order
7	5 digits	Random order
8	6 digits	Random order
9	3 digits	Random order
10	4 digits	Random order
11	5 digits	Random order
12	6 digits	Random order
13	2 digits	Random order
14	2 digits	Random order
15	24 characters	Random order
16	24 characters	Random order
17	7 stressed char.	In order
18	10 other symbols	΄ , ; . ! + - = / %
19	205 characters	Poem

Table 1: The field information.

Special attention was paid in gathering data from writers of different age and educational level (fig.2). Moreover, we decided to accomplish the filling of forms in different places (homes, offices, schools, and public places) in order to include different styles of writing, i.e. relaxed, in a hurry etc.

Each writer was asked to fill up to two forms, one from each type. Finally, the forms that composed the GRUHD database were selected carefully to be legible and with the less mistakes possible. In total, 500 men and 500 women were selected (fig.2a).

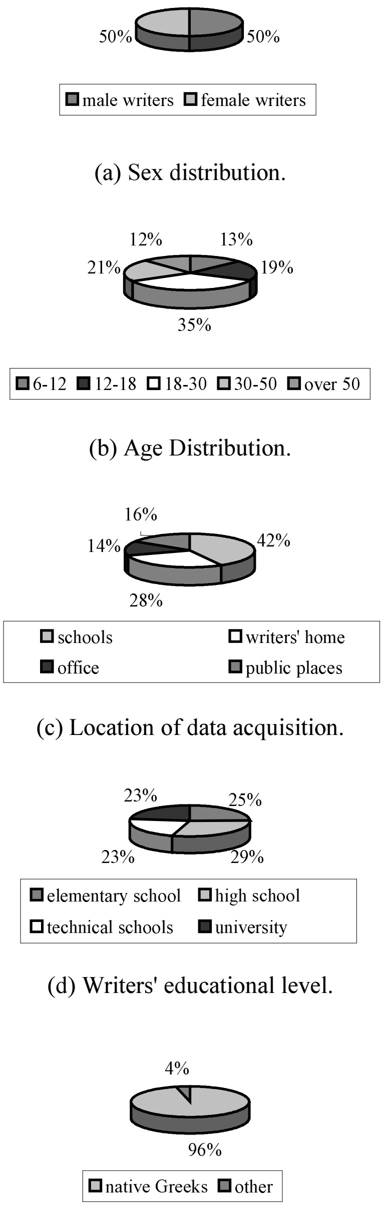
Specifically:

- The age distribution was as follows: 13% between 6-12 years, 19% between 12-18, 35% between 18-30, 21% between 30-50 and 12% over 50 years (fig.2b).

- 42% of the forms were filled in schools, 28% in writers’ homes, 14% in offices and 16% in public places (fig.2c).
- The educational level of the writers was: 26% elementary school, 32% general high school, 24% technical schools and 24% university (fig.2d).
- 96% of the writers were native Greeks (fig.2e).

2.2. Data Processing

The forms were scanned in grayscale and 300 dpi and labeled according to the writer and the sex of the writer as well as the type of form. Each form is stored in a file entitled XXXC.bmp, where XXX is a unique integer corresponding to the writer and C is a character that gives the type of form.



(a) Sex distribution.

(b) Age Distribution.

(c) Location of data acquisition.

(d) Writers' educational level.

(e) Nationality.

Figure 2: Writers’ distribution.

Character	Men writers	Women wr.	Total
Α	14875	14503	29378
Β	1452	1087	2539
Γ	3165	3258	6423
Δ	1326	2072	3398
Ε	6538	7998	14536
Ζ	1548	1805	3353
Η	8546	8231	16777
Θ	2468	2498	4966
Ι	8964	9423	18387
Κ	4352	4545	8897
Λ	2764	3093	5857
Μ	7457	7291	14748
Ν	4521	4620	9141
Ξ	2561	2395	4956
Ο	9478	9598	19076
Π	6385	6465	12850
Ρ	6795	6807	13602
Σ	8893	8898	17791
Τ	9006	9203	18209
Υ	3165	3388	6553
Φ	3267	2380	5647
Χ	1165	1253	2418
Ψ	1516	1826	3342
Ω	2468	2563	5031
α	10136	10352	20488
β	1274	1309	2583
γ	2497	2568	5065
δ	1643	1823	3466
ε	6941	6745	13686
ζ	1600	1711	3311
η	5763	5961	11724
θ	2671	2587	5258
ι	9354	9508	18862
κ	4693	4773	9466
λ	3056	3153	6209
μ	6942	6958	13900
ν	4873	5082	9955
ξ	1698	1762	3460
ο	8149	8364	16513
π	5943	6154	12097
ρ	7183	7406	14589
ς	2698	2774	5472
σ	5309	5378	10687
τ	9762	10099	19861
υ	2899	3023	5922
φ	2922	3024	5946
χ	1169	1248	2417
ψ	1623	1751	3374
ω	2695	2656	5351
ά	5964	5973	11937
έ	1896	1860	3756
ή	4268	4042	8310
ί	1269	1515	2784
ό	3159	3177	6336
ύ	1396	1502	2898
ώ	803	885	1688

Table 2: Amount of Greek alphabet characters included in the database.

Digit	Men writers	Women wr.	Total
0	7752	7651	15403
1	4962	5323	10285
2	5863	6136	11999
3	5023	5254	10277
4	7692	7721	15413
5	7794	7593	15387
6	5094	5153	10247
7	4163	4417	8580
8	5098	5174	10272
9	7563	7830	15393

Table 3: Amount of digits.

Character	Men writers	Women wr.	Total
΄	1269	1288	2557
,	3496	3590	7086
;	841	853	1694
.	3096	3152	6248
!	1564	1594	3158
+	853	851	1704
-	796	758	1554
=	842	855	1697
/	832	865	1697
%	796	890	1686

Table 4: Amount of symbols.

The forms written by men were given the numbers from 1 to 500 while the ones by female writers the numbers from 501 to 1,000. The character ‘l’ stands for the form with the poem in lowercase, while ‘u’ stands for the form with the poem in uppercase. There are no forms labeled with the same number and character, while forms with the same number but different character are written by the same person.

During data preprocessing, two annotation files were created for each form. The one keeps the coordinates for each word of the poem while the other the coordinates for each character, as well as the corresponding ASCII code. An additional check took place for each form, in order to verify the correctness of the above information. In tables 2-4 the amount of symbols included in the database are shown by sex and in total**.** The differences in the occurrence of the alphabet characters represent the differences distribution of the corresponding characters in Greek text.

3. Data Organisation

The GRUHD database is accompanied by the GRUHD software (fig.3) that uses the information acquired from the annotation and organizes the data according to the demanded requirements. As already mentioned two kinds of data can be extracted from the forms: symbols or words. Thus, there are many different ways of organizing these data.

The user is able to specify the data (fig. 4a) she/he wants to extract and select the forms, the writers (fig. 4b) or the sex (fig.5). According to the users’ selection the database can be organised in directory trees for the symbols or the words.

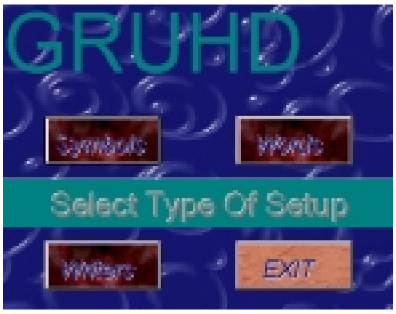
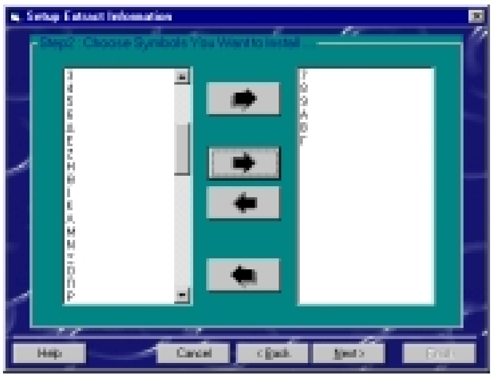
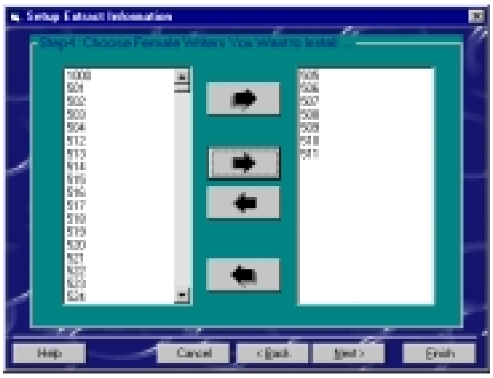


Figure 3: The GRUHD Software.

As far as the symbols are concerned we can select the organisation by ASCII code or writer. In the first case the organisation tree consists of directories named after the ASCII code of the symbols. By choice, the hierarchy tree can include or not the information of the sex of the writers. Finally, the bmp files with the symbols will have filenames of type XXX_YYY, where XXX is a number from 1-1000 representing the writer in accordance with the name of the corresponding form and YYY the serial number of the specific character in the corresponding form.


Setup Extract Information
Step2 : Choose Symbols You Want to Install
Help
Cancel
< Back
Next >
Finish

(a)



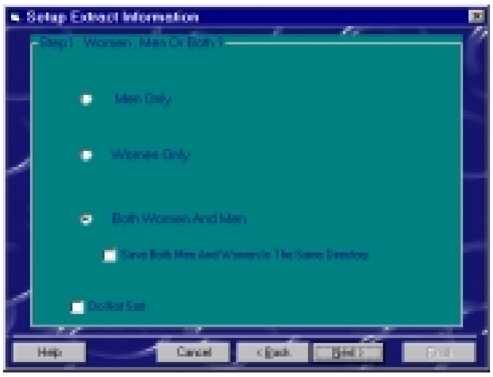
(b)

Figure 4: (a) selection of the data to be extracted,
(b) selection of the forms or the writers

In the latter case, organisation by writer, the data are classified in directories named after the serial number of the writer. The filenames are again of the type

XXX_YYY where XXX is the ASCII code of the included character and the YYY the serial number of the character with respect to the writer. Moreover, except of the sex information that may be included or not (fig.5), the data can also be recorded in the same directory, regarding the writer information (by ticking “Do not sort” in fig.5), and filename consisting of the ASCII code and the serial number of the character in the database..

Regarding the words, again we have several ways of organizing them. The filenames are of the type word_YYY, while YYY is a serial number according to the case. The sex information is optional here as well and the user can select to include the writer information in the directory tree or not. In the latter case the data are written in directories that bring the writer code as a name and the serial number is increased with respect to the writer.



Setup Extract Information

Step1 : Women , Men Or Both ?

○ Men Only

○ Women Only

● Both Women And Men

□ Save Both Men And Women In The Same Directory

□ Do Not Sort

Figure 5: Selection of the writers’ sex and directories organization

4. Conclusions

In this paper we presented the GRUHD database which consists of unconstained handwritten Greek characters, text, digits, and other symbols written from 1,000 writers, 500 men and equal number of women. Special attention was paid in gathering data from writers of different age and educational level. The GRUHD database is accompanied by the GRUHD software that facilitates its installation and use and enables the user to extract and use the data from the forms selectively, depending on the application. The various types of possible installations make it appropriate for the training and validation of character recognition, character segmentation and text-dependent writer identification and verification systems.

In particular, the words of the database have been used for the training and testing of the character segmentation system described in (Kavallieratou, 2000) as well as in the OCR system developed in the framework of the European project ACCeSS (LE-1 1802) that combines spoken and written language in call center applications.

The major problem we faced during the creation of the database was the selection of the most representative data. The form processing and the definition of the data bounding boxes were done manually, since no restriction was set to the writers in order to achieve unconstrained writing. Due to this fact many documents contained errors, missing characters etc, which didn’t allow the complete automation of processing.

References

Wilkinson, R., 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.

Marti, U., and H. Bunke, 1999. A full English sentence database for off-line handwriting recognition. *Proc. 5th Int. Conference on Document Analysis and Recognition, ICDAR'99*. Bangalore, 705 - 708.

Suen, C., C. Nadal, R. Legault, T. Mai, and L. Lam, 1992. Computer recognition of unconstained handwritten numerals. *Proc. Of the IEEE*, 7(80):1162-1180.

Kim, D., Y. Hwang, S.Park, E.Kim, S. Paek, and S. Bang, 1993. Handwritten korean character image database PE92. *Proc. Of the Second Int. Conf. On Document Analysis and Recognition*, 470-473.

Saito, T., H. Yamada, and K.Yamamoto, 1985. On the database ETL 9 of handprinted characters in JIS chinese characters and its analysis. *IEICE Transactions*, J68-D(4):757-764.

Kavallieratou, E., E. Stamatatos, N. Fakotakis, and G. Kokkinakis, 2000. Handwritten Character Segmentation Using Transformation-Based Learning. *Proceedings of ICPR 2000.*