MAYASTROUN- A Multilanguage Handwriting Database

Sourour Njah Badreddine Ben Nouma Hala Bezine Adel M. Alimi
REGIM: REsearch Group on Intelligent Machines, University of Sfax, National School of Engineers (ENIS), BP 1173, Sfax, 3038, Tunisia
{sourour.njah, hala.bezine, adel.alimi } @ieee.org ; badreddine_001@hotmail.fr

***Abstract*— To test the performance of handwriting on-line or offline recognition system, to measure and experimenting handwriting, to valid psychologist's laws, to compare and to make competitions between existing handwriting system, we must use databases.**
In this paper, we present the new dual handwriting database: "MAYASTROUN".
The developed database contains a large lexicon of unconstrained cursive Arabic and Latin texts, words, characters, digits, mathematical expressions and signatures. The collected data contains more than 67825 and written by 355 writers.
To facilitate experiments and research, this database offers four different extensions of each file and interactive software. The database architecture is described in details in this paper. The MAYASTROUN-database is available for the purpose to improve handwriting research filed.

Keywords: MAYASTROUN-database; DB-CREATOR; Handwriting analysis.

I. INTRODUCTION

Experimenting handwriting is one of the important researches filed. In order to test new theories, algorithms, and systems, it is necessary to use databases. Actually there are many databases available for on-line and offline handwriting, or researchers can develop their own database to test their systems such as presented in this works [1] [2] [15] [16] [28]. Reviewing the existing databases, we note that none of them treats both on/off line Arabic and Latin cursive handwriting. So the development of a new database is useful for the scientific community and actually the interest towards Arabic and Latin script recognition is continuously growing, we have developed a multilanguage and multivariate database named "MAYASTROUN-database".

A. On-line versus offline handwriting recognition

Handwriting recognition can be broken into two fields which differ in the form in which the data is represented to the system. The field of on-line handwriting recognition requires that the user write on a digitizing tablet using a special stylus, so that the user's written strokes are captured as they are being formed by sampling the pen's (x, y) coordinates at eventually spaced time intervals. In contrast, in offline handwriting recognition, the user writes on paper which is later digitized by a scanner. The data is presented to the system as an image, requiring a segmentation of the writing from the image background before recognition can be done. The on-line case deals with a spatio-temporal representation of the input; whereas the off-line case involves analysis of the spatio-luminance of an image. Some of already proposed database in the literature are presented below in [11] [12] [13] [17] [18] [19] [20] [21] [22] [23] [24] [26] [27] [29] [30].

B. Review of on-line database

- ADAB (Arabic DAtaBase) database, in 2009. It contain up to 15000 Arabic words corresponding to Tunisian town and village names, handwritten by more than 130 different writers [12][13].
- The IAM handwriting database, since 2005, contains forms of unconstrained handwritten English text, more than 1700 acquired forms and 221 writers [19].
- The TUAT Nakagawa Lab Online Handwriting database, since 1997, consists of handwritten character patterns from 10 writers out of 120 writers of the latter for a sample text excerpted from Japanese newspapers, covering Japanese, Chinese, Latin/numeric characters, and other symbols [20].
- The UNIPEN handwriting database, since 1993, contains over 5 millions Western characters, from more than 2200 writers [18].

C. Review of on/off line databases

- The IRONOFF dual handwriting database, since 1999. For each handwritten character or word, the on-line and off-line signals are available. It contains 4086 isolated digits, more than 10679 isolated lower and upper case letters, and 31346 isolated words from a 197 word lexicon (French, English) [21].

D. Review of offline databases

- The Rimes database composed of mails sent by individuals to companies or administrations; it contains 12,723 pages corresponding to 5605 mails [27].
- The IFN/ENIT database of handwritten Arabic words contains more than 2200 binary images of handwriting

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.230

samples forms from 411 writers, 26000 binary city word images and 212211 characters and ligatures [25].
- The IAM handwriting database contains forms of unconstrained handwritten English text. Since 1999, it contains 1539 pages of scanned text, 5685 isolated and labelled sentences, and 13353 isolated and labelled text lines, 115320 isolated and labelled words [24] [17].
- The CEDAR CDROM 1 database contains 5632 city/states handwritten words, 4938 states handwritten words, 9454 ZIP codes, 21179 handwritten digits and 27837 alphanumeric characters. A CDROM 2 is a database of machine-printed Japanese character images [22].
- The CENPARMI contains 17000 isolated digits obtained from postal ZIP codes [21] [11].
- The MNIST database of handwritten digits. It is a subset of a larger set available from NIST, includes a training set of 60000 examples, and a test set of 10000 examples [26].
- The NIST hand printed forms and characters database. Special database 19 contains samples forms from 3600 writers, 810000 character images isolated from their forms. It replace NIST special database 3 and 7 [23].

This paper presents a MAYASTROUN-database for both on-line and off-line unconstrained handwritten data.
In the second part of this paper, we expose database details. Then we present the related works. Finally, we present conclusions and future works.

II. MAYASTROUN-DATABASE DETAILS

The lexicon of MAYASTROUN-database was developed in the REGIM laboratory. It was collected by using five digital tablets connected to computers, 355 scripter mainly of Tunisian nationality, and the DB-CREATOR software. This database contains: digits, Western characters in lower and upper case, Arabic texts and characters, mathematical expressions, symbols, words in different languages and signatures. Note that there are no constraints according to the style, speed or size are imposed to writers, so a free style of handwriting is presented. Different samples of handwritten forms can be retrieved with or without pen ups, with or without dots and with different speed and size. Fig.1 presents the general schema of database construction.

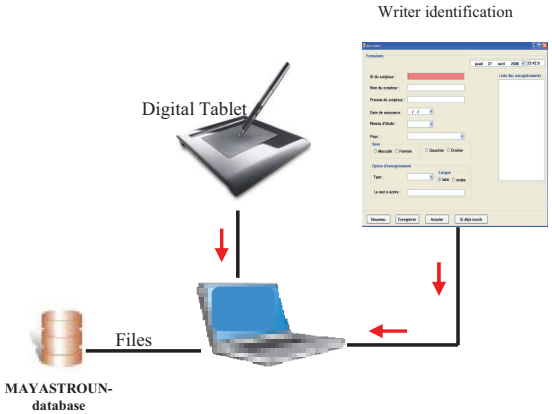

Figure 1. General schema of MAYASTROUN-database construction.

MAYASTROUN database is presented in two sets. In its first version (MAYASTROUN set1) contains: 2600 letters, 1000 words, 1000 digits and 200 Arabic texts acquired by 100 writers [7].

In set 2, we go for an extension on acquired samples with other different scripter. So, we obtain 67825 samples, detailed in the Table 1. Note that the Arabic alphabet is composed of 28 basic, some of these letters can appear in four different forms, the form of the letter changes if it is in the beginning, middle, in the end or in isolated form letters in addition to the appearance of dots [14].

TABLE I. MAYASTROUN-database description.

Acquired data	Numbers
Arabic words	1500
Arabic characters	5600
Arabic texts	200
Latin text	600
Latin words	38000
Latin characters	14675
Digits	6500
Mathematical expression	550
Signature	200
Total	67825

Table 2 presents a part of collected Arabic words.

TABLE II. LEXICON OF 68 ARABIC WORDS.

إيناس	واحد	يامن	بسم الله الرحمان الرحيم
دخل	إثنان	أكل	لا اله الا الله
محمد أمين	ثلاثة	طائر	لا اله الا الله محمد رسول الله
مريم	أربعة	حائر	محمد رسول الله
معرفة	خمسة	منزل	الله
يوسف	ستّة	غابة	آية
ضحك	سبعة	طاهر	رسالة
الرّحمان	ثمانية	مثير	اناقة
مكث	تسعة	جبان	سلام
نجاح	عشرة	لؤلؤ	هو
ميروك	عشرون	مؤسسة	النجاح
بيت	ثلاثون	رسول	ياسين
تلميذ	أربعون	قياس	ياسر
قذف	خمسون	عالم	وليد
قاعدة	ستّون	وظيفة	أمين
تونس	مائة	مسلم	سوار
نجاح	ألف	صحيح	محمد

To collect data and information from writers, software was developed in **.Net** language named DB-CREATOR. The home page of this software is shown in Fig.2.

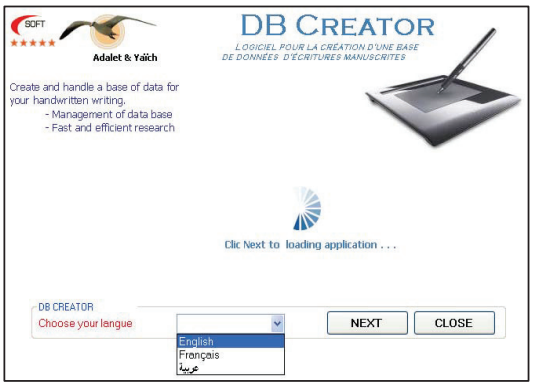

Figure 2. Home page of the DB-CREATOR.

According to the chosen language (Arabic, French, English) from the home page of the DB-CREATOR an electronic form with corresponding language is presented to the writer. So, each writer has an identifier number (ID) which will be used when acceding to the database. The writer fills an electronic form to collect personal information such as first and last name, country, date of birth, gender, the type of the script and other information, as illustrated in Fig.3.

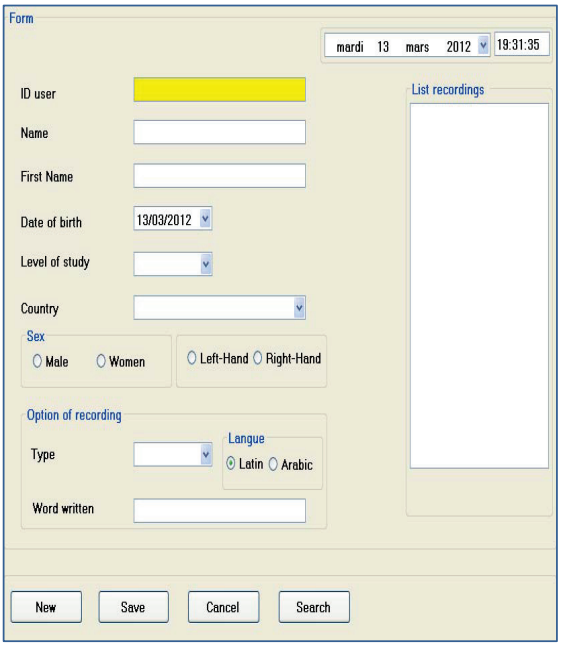

Figure 3. Electronic form for writer identifications.

The collected data files will be stored in repertories constructing the global architecture of the database as presented in Fig.4. Each input file is named with the handwritten script.

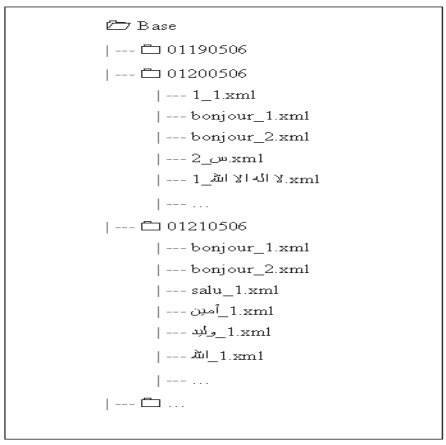

Figure 4. Architecture of MAYASTROUN database.

Each sample in the database, have four complementary files and stored with different extension. The first one has **.out** extension for the on-line handwritten database. The second one has **.bmp** extension for the off-line handwritten database. The third one has **.inkml** extension. And finally **.m** extension for Matlab software.

Fig.5 presents some examples of acquired data.

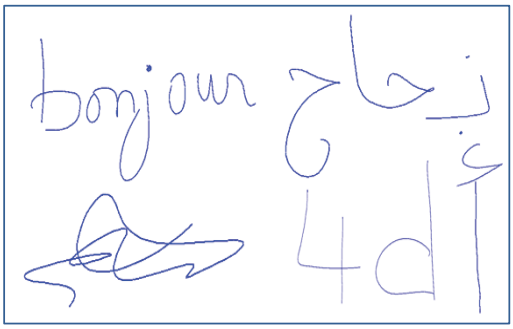

Figure 5. Some examples from MAYASTROUN-database.

Fig.6 presents detailed format of XML file.

Date creation of the file	<Date>mardi 13 mars 2012</Date>
Characteristics of the digital tablet	<Xdim>X_DIM 3600</Xdim> <Ydim>Y_DIM 3600</Ydim> <Point>POINTS_PER_SECOND 100</Point> <Cord>COORD X Y P</Cord>
Informations of the writer	<Id>01210506</Id> <Name>njah</Name> <fName>sourour</fName> <Country>Tunisie</Country> <dBirth>01/09/1976</dBirth> <lStudy>Universitaire</lStudy> <Hand>Droitier</Hand> <Sex>Feminin</Sex>
Type of acquired data	<WordType>Mot</WordType> <SegmentWord>bonjour</SegmentWord>
Acquired data	<Data> <zero> <Position>2915 3882 0</Position> <Position>2842 3903 0</Position> <Position>2765 3924 0</Position> ... </zero> <one> <Position>1177 5197 1</Position> <Position>1176 5209 1</Position> <Position>1176 5217 1</Position> ... </one> <zero> <Position>3624 4679 0</Position> <Position>3623 4673 0</Position> </zero> </Data>

Figure 6. Example of XML file.

III. MANIPULATING MAYASTROUN-DATABASE

The conception of database is important for handwriting community, so to facilitate acquisition of large lexicon, a DB-CREATOR was developed in three languages: Arabic, French and English. It is an interactive software, which permits to navigate in the database.

As mentioned, MAYASTROUN-database contains a variability of data, so we need a function to access at these data. One of the important utility of this software is to facilitate data research.

Simple search: if we search stored files with their associated names. For example, if we want to know all stored scripts beginning with the letter “a”, the user of the interface writes it in the text zone. The result obtained is all files written by different scripter, in addition of the total number. Fig.7. illustrates the interface for simple search for examples of scripts beginning with the letter “a”, note that that the total number is equal to 7 as indicated at the bottom of the window.

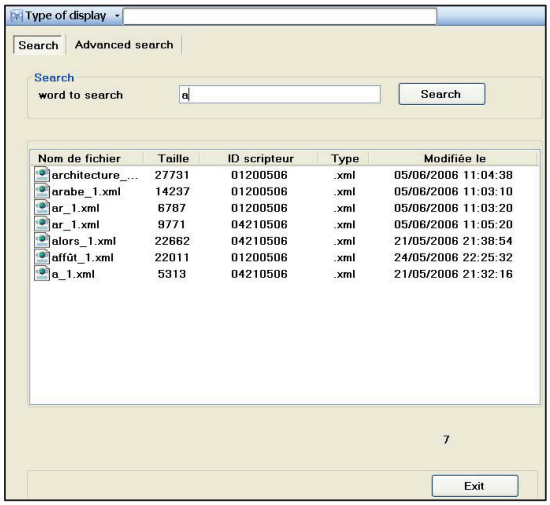

Figure 7. Interface of simple research for scripts beginning with the letter "a".

Advanced search: if we want to search files with other criterion such as the first or last name, date of birth, study level of the scripter, or the date of creation of the file. Fig.8, shows the interface used in advanced search.

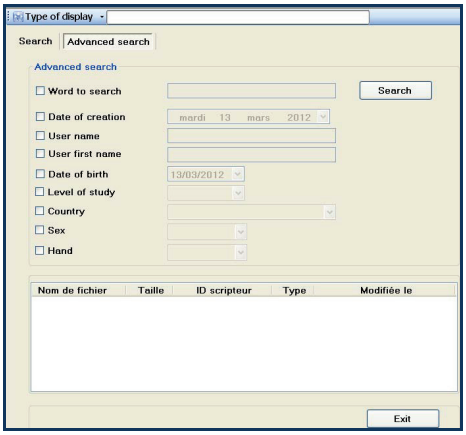

Figure 8. Interface of advanced research.

IV. WORKS RELATED TO MAYASTROUN-DATABASE

The collected data of the set1 are used to validate different works. In [7] [8] [9] [10], it was used to validate the PerTOHS theory for handwriting segmentation via perceptual codes (Elementary and global ones), where we have obtained encouraged results. These results exploited for handwriting segmentation are similar to those produced by the human perceptual system during the writing process, in addition to the obtained important data reduction rate.

In [3], we have developed an on-line handwriting identification system for Arabic texts. These texts were written in different five fonts, where good recognition rates are obtained using neural networks and the Beta-elliptic model [4] [5] [6].

In order to communicate with the computer in the simplest and fastest way, an Ink-enabled handwriting editor was developed in [5]. It deals with the transformation of the XML file to another editable format (Word, Latex, html…) using the XSLT language for the conversion of files.

These different experiments on acquired data with different forms and language, show successful results on segmentation, transformation, recognition and generation.

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we present MAYASTROUN-database, which contains 67825 data samples, written by 355 writers.

This enormous varied lexicon is acquired by digital tablets in different languages. The large number of writers guarantees a multiplicity of writing styles due to the differences in educational levels and the use of left or right hand to write.

To facilitate acquisition and manipulation of data, a DB-CREATOR software was developed in .Net language.

The first set of this database is used in works related to handwriting studies, such as segmentation, identification, recognition and development of editor.

This database is important for the handwriting research community in order to test and to validate new ideas and algorithms for both on-line and offline handwriting case.

Our perspectives are:

- To enlarge this database by acquiring other data in different languages by various writers.
- To extend the filed of handwriting analysis in graphology research, where we try to find the connection between handwriting and a person's behaviour.
- Development of educative editors for persons who are interested to learn Arabic or other languages.
- Organizing competitions based on collected data of MAYASTROUN-database in order to compare and test new systems in handwriting filed.
- To progress in "CODEMA" project, we will acquire more mathematical expressions.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the financial support of this work by grants from General Direction of Scientific Research (DGRST), Tunisian, under the ARUB 01/UR/11/02 program.
They also would mention that this paper is integrated in the Tuniso/Maroco project "CODEMA" under reference number 11/TM36.
Thanks for everyone who have participated in acquiring data.

REFERENCES

[1] A.M. Alimi, "Evolutionary Computation for the Recognition of On-Line Cursive Handwriting". IETE Journal of Research, Special Issue on "Evolutionary Computation in Engineering Sciences" edited by S.K. Pal et al., vol. 48, no. 5, pp. 385-396, 2002.

[2] A.M. Alimi, "Neuro-fuzzy approach to recognize Arabic handwritten characters", Proceedings of the IEEE International

Conference on Neural Networks, ICNN, Vol 3, pp. 1397-1400 , 1997.

[3] H. Bezine, S. Njah, A.M. Alimi, " Identification des fonts arabes anciennes par réseaux de neurones", Atelier sur la Numérisation de l'Ecrit Ancien et des GRAndes Masses de données, Fribourg, Swiss 2006.

[4] H. Bezine, A.M Alimi, N.Derbel, ''An explanation of the feature of a handwriting trajectory movement controlled by a Bêta-Elliptic Model'', Proc.Int.Conf., ICDAR, Edinburgh UK, pp. 1228-1232, 2003.

[5] H. Bezine, M. Kefi and A.M Alimi, "On the Beta-elliptic model for the control of the human arm movement", IJPRAI, vol 21, n°1, 2007.

[6] H. Bezine, S. Njah, A.M. Alimi, "Towards an Ink-Enabled Handwriting Editor", International Graphonomics Society, Cancun, Mexique, pp.213-216, 2011.

[7] S. Njah, H. Bezine, A.M Alimi, "On-line Arabic Handwriting Segmentation via Perceptual Codes: Application to MAYASTROUN database", IEEE SSD, Sousse, Tunisia, pp. 1-5 2011.

[8] S. Njah, H. Bezine, A.M Alimi, "A fuzzy-genetic system for segmentation of on-line handwriting: application to ADAB Database", SSCI-GEFS, Paris-France, pp.95-102, 2011.

[9] S. Njah, H. Bezine, A.M Alimi, " A new encoding system: Application to on-line Arabic handwriting", ICFHR, Kolkata-India, pp. 451-456, 2010.

[10] S. Njah, H. Bezine, A.M Alimi, "A new approach for the extraction of handwriting perceptual codes using fuzzy logic", ICFHR, Canada, pp. 302-307, 2008.

[11] C.Y Suen, C. Nadal , R. Legault, and T.A .Mai and L.Lam, '' Computer recognition of unconstrained handwritten numerals'', Proc of the IEEE, vol 80 , n°7, pp.1162-1180, 1992.

[12] H. El Abed, V. Märgner, M. Kherallah and A.M Alimi, " ICDAR 2009 Online Arabic Handwriting Recognition Competition", ICDAR, Spain, pp. 1388-1392, 2009.

[13] H. El Abed, V. Märgner, "Comparison of Different Preprocessing and Feature Extraction Methods for Offline Recognition of Handwritten Arabic Words", ICDAR, Brazil, pp. 974-978, 2007.

[14] K. Daifallah, N. Zark., H. Jamous, " Recognition-based Segmentation Algorithm for On-line Arabic Handwriting", ICDAR, Spain, pp. 886-890, 2009.

[15] M. Kherallah, S. Njah, A.M Alimi, N. Derbel, "Recognition of on-line handwritten digits by neural networks using circular and beta approaches". Proc.Int. Conf. IEEE International Conference on Systems, Man and Cybernatics, Hammamet, Tunisie, 2002.

[16] Plamondon R., Srihari S.N., "On-line and off-line handwriting recognition: A comprehensive survey", IEEE Trans. Pattern Anal. Machine Intell, vol 22, pp; 63–84, 2000.

[17] U. Marti, and H. Bunke, ''A full English sentence database for off-line handwriting recognition '', Proc.Int.Conf., ICDAR, India, pp.705-708, 1999.

[18] http://hwr.nici.knu.nl/unipen.

[19] http://iamwww.unibe.ch/~fkiwww/iamondb/

[20] http://www.tuat.ac.jp/%Enakagwa/ipdb/index_e.htm

[21] C. Viard-Gaudin, P M. Lallican, S. Knerr, and P. Binter, ''The IRESTE On/Off (IRONOFF) Dual Handwriting Database '', ICDAR, India, pp. 455-458, 1999.

[22] http://www.cedar.buffalo.edu/Databases/CDROM1/

[23] http://www.nist.gov/srd/nistsd19.htm

[24] http://www.iam.unibe.ch/~zimmerma/iamdb/iamdb.htm

[25] http://www.ifnenit.com/

[26] http://yann.lecun.com/exdb/mnist/

[27] http://www.rimes-database.fr

[28] S. Njah, "Reconnaissance en ligne des chiffres manuscrits par réseaux de neurones en utilisant l'approche " bêta-circulaire", Master thesis, University of Sfax, 2003.

[29] T.Saito, H.Yamada, K.Yamamoto, "On the data base ETL 9 of hand-printed characters in JIS chinese characters and its analysis", IEICE Trans. Fund. Electr. Comm. and Comp. Sciences,Vol J68-D(4), pp.757-764, 1985.

[30] G. Dimauro, S.Impedovo, R. Modugno, G. Pirlo, "A new database for research on bank-check processing", 8th International Workshop on Frontier in Handwriting Recognition , Nagara on the Lake (Canada), IEEE Computer Society Press, USA, pp.524-528,2002.