A NEW ARABIC (AHD/AMSH) HANDWRITTEN DATABASE

AMER AL-NASSIRI
Faculty of Computer Science & Eng., Ajman University of Science & Technology, Ajman – UAE, Email: amernassiri@yahoo.com

SHUBAIR A. ABDULLA
Faculty of Education & Basic Sciences, Ajman University of Science & Technology, Ajman – UAE, Email: shubaira@hotmail.com

ABSTRACT

This paper introduces new database for Arabic handwritten words. The Arabic handwritten database (AHD/AMSH) represents a utility to facilitate the experiments of the character recognition algorithms. It contains three types of images: word, isolated character, and digit images. The AHD/AMSH can be used for baseline detection, characters segmentation, normalization, thinning, training and testing purposes. The stages of construction of the AHD/AMSH database were planned carefully to ensure its excellence. 150 words, 35 courtesy amount and 20 digits were used to fill the form which has been filled by 82 writers in 5 different age groups. The results were 12300 words, 29028 sub-words, 56170 characters, 2870 courtesy amounts, 820 Indian digits, and 820 Arabic digits. After dividing the database into two categories, training and testing, it has been tested manually and systematically.

***Keyword*:** *Handwritten Arabic characters, Cursive writing, Database, Handwritten recognition*

1. INTRODUCTION

The off-line Arabic handwritten character recognition system comprises four main processes: preprocessing, segmentation, feature extraction, and recognition [1]. Its function is to automate the character recognition operation. The character recognition system experiment is one of the building phases that should be considered and performed before making any decision regarding the effectiveness. The testing of the off-line Arabic handwritten character recognition system is usually conducted by preparing the potential cases as an input list and then observing its output. Moreover, some systems involve Artificial Neural Network (ANN), Support Vector Machine (SVM), or HMM in the recognition stage [2, 3, 4]. The Arabic handwritten database is a list of images of words written by different writers and used to facilitate the training and testing of the off-line Arabic handwritten character recognition system by providing the training instances and the testing inputs. In other words, it is a test bench to experiment the off-line Arabic handwritten character recognition systems.

The availability of a standard and well-done database progresses the quality of the character recognition algorithms and this advancement is also applicable to the other scripts which use the Arabic characters in writing [5]. Unlike the Latin script, the Arabic script is suffering from the lack of a standard and solid database. There were no freely or noncommercially available standard databases for the Arabic handwritten word images till 1999 [6]. Compared to the Latin script where the databases, like CEDAR [7] and NIST [8] were designed many years ago, the situation in the Arabic script calls for reliable standard databases. From the outset, the researchers were testing their algorithms by using a small database of their own, like what has been conducted by T. Sheikh & Guindi [9] and H. Almuallim and S. Yamaguchi [10]. The situation was left unchanged until 1999 when the first database was published by N. Kharma et al. [6]. It has been collected from the students of Al-Isra' University in Amman, Jordan. Each student was asked to write words, digits, sentences, and finally to sign at the bottom of the form. The results were 37000 words, 10000 digits, and 2500 signatures. In spite of these results, the database has not been available publicly. In 2000, Y. Al-Ohali et al. [11] introduced a database for bank checks recognition. The database has been collected through scanning 7000 real world grey-level bank checks in collaboration with Al Rajhi Banking and Investment Corp. in Saudi Arabia. The collecting process involved scanning which was done at the bank center and removing the private information. Their database included 1547 courtesy amount, 23325 Arabic sub-words, and 9865 Indian digits, but it is clear, it is a special database and not suitable for a general Arabic recognition systems.

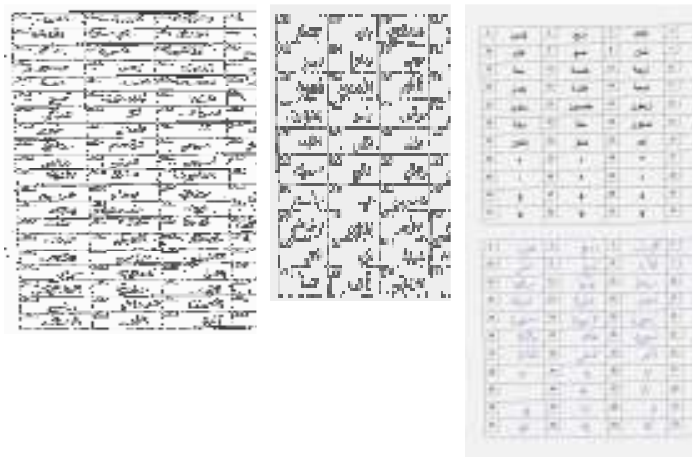
In 2002, S. Alma'adeed et al. [12] built an Arabic Handwritten Database (AHDB) database which contains words and numbers used in bank checks. A form of 6 pages was filled by 100 different writers. The writers were asked to fill four entries: 67 words corresponding to numbers that could be used in check writing, 29 words were from the most popular Arabic words like "في", "من", "على", etc., 3 sentences representing numbers and quantities. M. Pechwitz et al. [13] presented a new database for handwritten Arabic town/village name termed as IFN/ENIT**.** They were aiming to collect images of handwritten words similar to those written in letter addresses. 411 different writers filled 2265 forms with about 26400 names containing more than 21000 characters. Although the large number of writers gave a variety of writing styles, it did not guarantee the covering for all the shapes of the Arabic characters. This database is available on the internet. The literature of the well known Arabic handwritten database can be summarized as shown in table 1.

Table 1 The current Arabic handwritten databases

Database	Authors	Year	W/SW	Word Type	Availability
–	N. Kharma et al [6]	1999	3/000	General	--
–	Y. Al-Ohali et al [11]	2000	29325	Bank Check Filling	--
AHDB	S. Alma'sceed et al [12]	2002	–	Bank Check Filling	--
IFN/ENIT	M. Pechwitz et al [13]	2002	26459	Town/Village Name	On internet

This brief survey in the literature reveals that the community of the Arabic handwritten researches still needs to be enriched with more professional databases that cover all types of words and numbers.

The contribution of this paper is to design professional Arabic handwritten database (AHD/AMSH) that can be used for the Arabic character processes such as segmentation, recognition, baseline detection, etc. The database words have been chosen carefully to cover all the shapes of the Arabic characters. The filling forms (figure 1) have been distributed among writers of different ages (Table 2) to give the words an authentic reliability and readability. Professional computerized applets have been designed to capture word and number images and facilitate the building procedures. Table (3) shows the programmed applets. The outline of the forthcoming sections is: section 2 details the database construction stages, section 3 illustrates the features of the database, section 4 explains the experiments of the database, and section 5 concludes some notes.



Words Form Numbers Form

Figure 1 An example of AHD/AMSH filling forms

Table 2. The writers' groups

Group	Writers' Age
1	Between 5-15 years
2	Between 16-25 years
3	Between 26-35 years
4	Between 36-45 years
5	Above 45 years

Table 3. The programs that have been designed to facilitate the building processes

Program	Function
BMPWords	Ripping Word from the scanned pages
BMPPreprocessing	Preprocessing (smoothing and noise removing)

2 CONSTRUCTION STAGES OF AHD/AMSH

The AHD/AMSH database consists of 12300 Arabic handwritten words written by 82 different writers. 150 words have been chosen from the holy Quran, and then arranged in 5 page filling form. The choosing of the words is done cautiously and accurately in order to assure that all the shapes of the Arabic characters are covered.

After the forms have been checked and approved by a linguist, they have been distributed among writers in different age groups. Tables 4 and 5 show the words, courtesy, and numbers listed in the AHD/AMSH filling form. The writers are divided into 5 groups depending on their age (Table 3).

Table 4 The words of the AHD/AMSH filling form

الرحمن (SW1,L6)	بسم (SW1,L1)	الله (SW1,L2)	الرحيم (SW1,L3)	الحمد (SW1,L4)	لله (SW1,L5)
رب (SW2,L1)	العالمين (SW2,L2)	الرحمن (SW2,L3)	الرحيم (SW2,L4)	مالك (SW2,L5)	يوم (SW2,L6)
الدين (SW3,L1)	إياك (SW3,L2)	نعبد (SW3,L3)	وإياك (SW3,L4)	نستعين (SW3,L5)	اهدنا (SW3,L6)
الصراط (SW4,L1)	المستقيم (SW4,L2)	صراط (SW4,L3)	الذين (SW4,L4)	أنعمت (SW4,L5)	عليهم (SW4,L6)
غير (SW5,L1)	المغضوب (SW5,L2)	عليهم (SW5,L3)	ولا (SW5,L4)	الضالين (SW5,L5)	آمين (SW5,L6)
١ (SW1,L7)	٢ (SW2,L7)	٣ (SW3,L7)	٤ (SW4,L7)	٥ (SW5,L7)	٦ (SW6,L7)
٧ (SW7,L7)	٨ (SW8,L7)	٩ (SW9,L7)	١٠ (SW10,L7)	١١ (SW11,L7)	١٢ (SW12,L7)
١٣ (SW13,L7)	١٤ (SW14,L7)	١٥ (SW15,L7)	١٦ (SW16,L7)	١٧ (SW17,L7)	١٨ (SW18,L7)
١٩ (SW19,L7)	٢٠ (SW20,L7)	٢١ (SW21,L7)	٢٢ (SW22,L7)	٢٣ (SW23,L7)	٢٤ (SW24,L7)
٢٥ (SW25,L7)	٢٦ (SW26,L7)	٢٧ (SW27,L7)	٢٨ (SW28,L7)	٢٩ (SW29,L7)	٣٠ (SW30,L7)

2.1 DISTRIBUTING THE AHD/AMSH FILLING FORM

The AHD/AMSH Filling Form was easy-to-fill. It contained 6 pages, the first of which was filled with personal information about the writer: name, age, gender, and occupation. The second and third pages contained a table of 6 columns by 30 rows; each cell was numbered and contained one word. The writer wrote the word in its corresponding cell inside a blank table on pages four and five. The last page contained the most used courtesy amount words along with Indian and Arabic digits. In designing the form, it was taken into consideration that the writing of the words has to be unconstrained and the ripping of the words from the scanned pages has to be easy and takes a little time. Table 6 shows the number of words,

sub-words, characters, courtesy amount words, and digits in one form.

Table 5 The courtesy and numbers listed in the AHD/AMSH filling form

خمس	اربع	ثلاث	اثنان	واحد
عشر	تسع	ثمان	سبع	ست
ستة	خمسة	أربعة	ثلاثة	أثنتان
إحدى	عشرة	تسعة	ثمانية	سبعة
ستون	خمسون	أربعون	ثلاثون	عشرون
مائة	مئة	تسعون	ثمانون	سبعون
إثنتين	صفر	أحد	مليون	ألف
٥	٤	٣	٢	١
٠	٩	٨	٧	٦
5	4	3	2	1
0	9	8	7	6

Table 6 the content of the AHD/AMSH filling form

Words	150
Sub-words	354
Characters	685
Courtesy amount words	35
Digits	20 (10 Indian + 10 Arabic)

2.2 DATA GROUPING

The writers were classified into 5 groups according to their ages as shown in table 3. In each group, there were a random number of males and females. The writers were aged from 8 years to about 55 years. Their occupations were administration staff, faculty members, housewives, undergraduate students, and schoolboys. The writers who belonged to group 1 were from Fujairah Islamic and Scientific Academy FISA in the United Arab Emirates**,** while the faculty members and the undergraduate students were from Ajman University of Science and Technology–Fujairah Campus in the United Arab Emirates. The idea of filling the AHD/AMSH form by writers in different age groups was to give the words an authentic readability level and for testing the effect of age on the segmentation and recognition rates. The readability level started from a low level (group 1) and improved in the next higher groups (figure 2-a). Some abnormal cases which showed that the readability level was low because of the nonchalance of the writers were discovered (figure 2-b).

From Group 1 From Group 3 From Group 5

a. The Readability level of the Word "ظهر" Increased

"رواسي" Word From Group 5 "أراكهم" Word From Group 5

b. Some Abnormal Cases

Figure 2 Some Normal and Abnormal Cases

2.3 SCANNING AND DIGITIZATION

The forms were scanned by using HP Scan Jet 2400c scanner at 300dpi resolution. The scanner software was set to produce B/W BMP image which is easy to manipulate. A tagging scheme was suggested to facilitate the access to the BMP files. This scheme is described as follows:

Digit	**Value**
The group No	1..5
The gender of the writer	<F> or <M>
The writer No	1.. to number of the writers
The page No	<P4> or <P5>

Examples:

- 1F4P4.BMP
- 5M2P5.BMP

After the scanning operation, a process was started to rip each word in B/W BMP image file, using BMPWords program shown in table 3. A small Windows-based applet *"BMPWords*" (Figure 3 shows a snapshot of the BMPWords interface) was designed to automate the word extraction or ripping from the scanned image. Another tagging scheme was used in tagging the files of the words images, as follows:

Digit	**Value**
The group No	1..5
The gender of the writer	<F> or <M>
The writer No + < _ >	1.. to number of the writers
The word No	1..150

Examples:

- 1F40_120.BMP
- 5M23_11.BMP

The final step was to segment the words and produce the isolated characters' image. The characters were tagged according to the following scheme:

Digit	**Value**
The Serial No of the character	1..28
Letter	<C>
The shape No of the character plus < _ >	1..6
The serial No of the character's representation	1.. no of the representation

Examples:

- 27C2_3.BMP
- 5C1_11.BMP

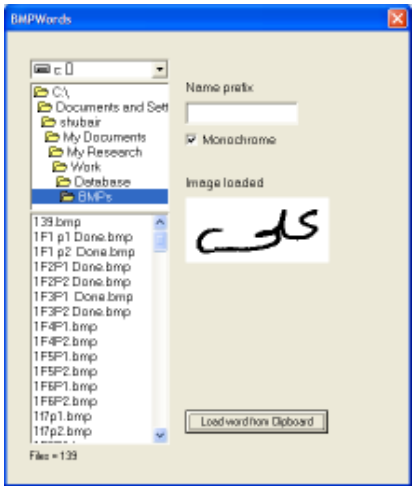


Figure 3. The BMPWord program interface.

Due to the erratic handwritten words error, a designed Windows-based applet "BMPPreprocessing" was used to automate the noise removing in order to obtain a clean word images. Figure 4 shows the interface of the BMPPreprocessing program. The words were saved in five sub directories which were allocated to the groups plus one sub directory to contain the documentation of the database.


Figure 4 the interface of the BMPPreprocessing program

3 THE AHD/AMSH FEATURES

The AHD/AMSH database possesses distinctive specifications which gave it a good rank among other databases. These specifications can be summarized as follows:

1. Different styles of handwriting and different levels of complication give the algorithms of handwritten character recognition an ideal experimental tool. The number of writers, words, characters, and numbers which contained in the database is shown table 7.
2. The database is divided into two sets: training set and testing set as shown in table 8.
3. The database contains five types of images: connected characters (words/sub-words), isolated characters, courtesy, and numbers. The connected characters images can be used by different types of algorithms such as baseline detection, characters segmentation, normalization, and thinning algorithms. The isolated character and number images are perfect for the features extraction and for training the recognizers such as NN and SVM.

Table 7 Number of Writers, Words, Sub-words, Letters, and Digits in Each Group

Categories	Writers		Database				
	M	F	Word	Sub-word	Letter	Courtesy	Digits
5-15 years	7	7	2100	4956	9590	490	280
16-25 years	12	10	3300	7788	15070	770	440
26-35 years	8	9	2550	6018	11645	595	340
36-45 years	9	8	2550	6018	11645	595	340
Above 45	5	7	1800	4248	8220	420	240
Total			**12300**	**29028**	**56170**	**2870**	**1640**

Table 8 the training and testing sets in the database

	Word	Sub-word	Letter	Courtesy	Digits
Training	1845	4354	8428	431	246
Testing	10455	24674	47745	2440	1394
Total	**12300**	**29028**	**56170**	**2870**	**1640**

4. Every Arabic character shape has been represented in the database at least 82 times; this means providing a large number of different instances of the character shape. These instances can be used to consolidate the trainer that leads to build accurate recognizer.
5. A lot of irregular handwriting styles are covered in the database which increases the challenge in solving the segmentation, features extraction, and the recognition problems. Figure 5 shows some of these styles, which can be categorized as difficult Arabic handwritten words.

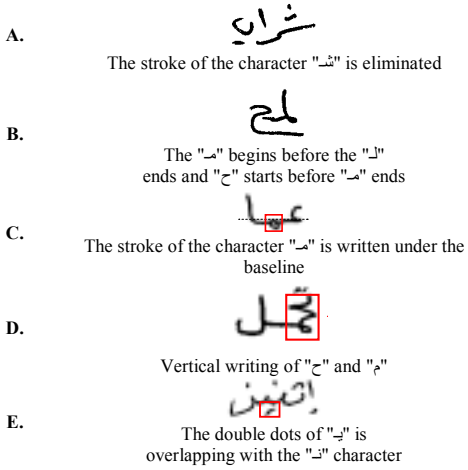


Figure 5. Some of the Irregular Handwriting Styles

4 EXPERIMENTS

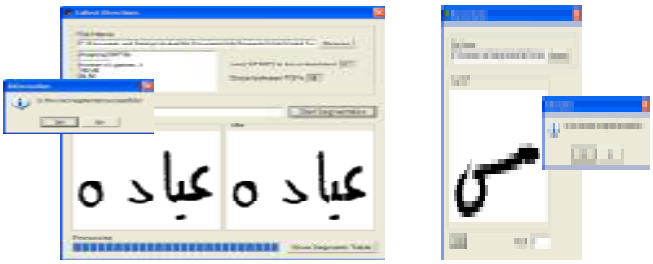
In order to ensure the quality of the database, it has been tested by two different algorithms:

1- Automatically by using the RISF segmentation algorithm suggested by [14] and the "Character Recognition of Arabic and Latin Scripts" recognition system suggested [15] which was used to recognize the isolated Arabic handwritten characters. All the words and characters in the training sets were used to train the algorithm. The testing set, shown in

table 8, was used to test the recognition system [15]. The results were promising and they show about 95.8%.

2. Manually by counting the correct results by watching the output for the RISF segmentation algorithm as shown in figure 6. The segmentation rate was about 98.5%.

The difference between the two recognition rates is mainly related to the proficiency of the recognition system which is mentioned in point 1.



The segmentation The recognition

Figure 6 The programs used in the experiments

5 CONCLUSIONS AND FUTURE DIRECTIONS

This paper has introduced new non-commercial Arabic handwritten database. Great efforts have been devoted to provide an experimental tool that helps the researchers in their endeavor toward building successful character recognition automation. The AHD/AMSH database can be used for different kinds of Arabic character recognition algorithms.

The future planning involves recommendations derived from the review of the literature of the Arabic handwritten character community. These recommendations are described in the following points:

1. Increasing the number of the words and the number of the characters is one of our interests. According to the planning, the AHD/AMSH Filling Form must be filled by new writers to increase the number of the objects in the database.
2. The database must be published on the internet to be available publicly. We have already started the publishing process, and the website will be accessible by the end of the 2007. The database can be requested freely by the author's email.
3. By publishing it on the internet, any suggestion by researches working in the field can be taken into consideration to improve the database.

REFERENCES

[1] N. Arica and F. T. Y. Vural. "An overview of character recognition focused on off-line handwriting". IEEE Trans. on Systems, Man, and Cybernetics - Part C: Applications and Reviews, V. 31, No 2, P 216-233, 2001.

[2] M. Altuwaijri and M. Bayoumi, "Arabic Text Recognition using Neural Networks", IEEE International Symposium on Circuits and Systems ISCAS'94, London, Vol. 6, pp. 415418, 30 May – 2 Jun 1994

[3] Fu Chang, Chin-Chin Lin, Chun-Jen Chen, "Applying a hybrid method to handwritten character recognition", Proceedings of the 17th International Conference on Volume 2, Issue , 23-26 Aug. 2004, pp.529 – 532.

[4] Mezghani, N., Mitiche A., Cheriet M., "On-line recognition of handwritten Arabic characters using a Kohonen neural network", Frontiers in Handwriting Recognition Proceedings, Eighth International Workshop on Volume, Issue , 2002, pp. 490 – 495.

[5] L. M. Lorigo, V. Govindaraju, Offline Arabic Handwriting Recognition: A Survey, IEEE Transactions on Pattern analysis and Machine Intelligence, V. 28, NO. 5, pp. 712-724, May 2006.

[6] N. Kharma, M. Ahmed, R. Ward, "A New Comprehensive Database of Hand-written Arabic Words, Numbers and Signatures used for OCR Testing", Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering, P. 766-768, 1999.

[7] CEDAR, http://www.cedar.buffalo.edu/Databases/CDROM1

[8] NIST, http://www.nist.gov/srd/nistsd19.htm

[9] T. Sheikh & Guindi, "Computer Recognition of Arabic Cursive Scripts" Pattern Recognition, V. 21, N. 4, P. 293-302, 1988.

[10] H. Almuallim and S. Yamaguchi, "A Method of Recognition of Arabic Cursive Handwriting," IEEE Trans. Pattern Analysis and Machine Intelligence, V. 9, P. 715-722, 1987.

[11] Y. Al-Ohali, M. Cheriet, and C. Suen, "Databases for Recognition of Handwritten Arabic Check", Proc. Seventh Int'l Workshop Frontiers in Handwriting Recognition, P. 601-606, 2000.

[12] S. Alma'adeed, D. Elliman, and C.A. Higgins, "A Data Base for Arabic Handwritten Text Recognition Research," Proc. Eighth Int'l Workshop Frontiers in Handwriting Recognition, P. 485-489, 2002.

[13] M. Pechwitz, S. S.Maddouri, V.Maergner, N. Ellouze, and H. Amiri. "IFN/ENIT – database of handwritten Arabic words", In Proc. of CIFED 2002, P. 129-136, Tunisia, 2002.

[14] Shubair Abdulla, Amer Al-Nassiri, Rosalina Abdul Salam. "Off-Line Arabic Handwritten Word Segmentation Using Rotational Invariant Segments Features (RISF)", Accepted and to be appeared in IAJIT, Vol.5, No.4, Oct.2008.

[15] F. Hussain and J. Cowell, "Character Recognition of Arabic and Latin Scripts", pp. 51-56, Fourth International Conference on Information Visualization (IV'00), 2000.