Novel Databases for Arabic Online Handwriting Recognition System

Mohamed AbdElNafea
Electrical Engineering Department,
Faculty of Engineering,
Aswan University, Egypt
mabdelnafea@gmail.com

Samia Heshmat
Electrical Engineering Department,
Faculty of Engineering,
Aswan University, Egypt
samia.heshmat@aswu.edu.eg

***Abstract*—**In this paper, novel databases for Arabic handwriting recognition, we call them Aswan online Arabic handwritten (**AOLAH**) databases, are presented. They are designed to help researchers in the field of online Arabic handwriting recognition. The first proposed database consists of 2,520 files, representing 28 Arabic characters written by 90 different writers. Those writers are volunteers from Aswan faculty of Engineering. The other database is extracted from the main database by taking only the strokes from characters. This second database contains 1,530 files for 17 strokes. The uniqueness of the recommended algorithm for data collection is the dealing with every stroke in the character separately. Stroke capturing is done by collecting data points along the path of an input device (stylus pen or mouse) same time those characters are drawn.

Keywords*—Handwritten database; Online Arabic Handwriting Recognition; Image processing.*

I. INTRODUCTION

Arabic language is the language spoken by Muslims in many countries and one of the most common languages because its characters are used in other languages like Kurdish, Persian, and Urdu. Researchers face a challenge when dealing with Arabic language in testing a recognition system they built because of the nature of Arabic language in cursive, overlapping, handwriting variability, different writing styles and other challenges. Furthermore, any proposed recognition system for Arabic handwriting recognition requires a real and substantial database [1, 2]; hence, we created novel databases for Arabic characters to help researchers in the field of online Arabic handwriting recognition. The databases are developed based on an algorithm that uses stroke capturing to facilitate recognition of Arabic characters [3]. The proposed databases (**AOLAH**) are typical formats of online handwritten data which is a sequence of coordinate points of the moving pen point. Connected parts of the pen trace, in which the pen point is touching the writing surface, are called strokes.

The remainder of this paper is organized as follows: explanation about existing Arabic databases is shown in Section 2, while Section 3 presents the basic idea and procedures of the proposed databases (**AOLAH**). Finally, the conclusion is given in Section 4.

978-1-7281-4801-4/20/$31.00 ©2020 IEEE

II. EXISTING ARABIC HANDWRITTEN DATABASES

Some of the main databases used in online Arabic handwriting recognition researches are described in this section. Table 1 shows a summary of these databases.

LMCA (2008) [4, 5]: The On/Off (LMCA) Dual Arabic Handwriting Database; this abbreviation is from the French sentence "Lettres, Mots et Chiffres Arabe,". This database contains 30,000 digits, 100,000 Arabic letters and 500 Arabic words, 55 participants were invited to contribute. This database was developed in REGIM (REsearch Group on Intelligent Machines) laboratory. Both on/off line handwritten characters and words were considered. LMCA database is limited to a small set of words, and the letters are collected separately (not segmented from cursive text).

OHASD (2010) [6]: The first on-line Arabic sentence database handwritten on tablet PC. The final version of this dataset is composed of 154 paragraphs, selected from public daily news, written by 48 writers, having a total of 3,825 words and 19,467 characters, after excluding erratic/illegible handwritings. This database has a limited lexicon, limited data, and a limited number of writers.

ADAB (2011) [7, 8]: The Arabic DAtaBase was developed by the institut fuer Nachrichtentechnik and the research group on intelligent machines (REGIM). It contains online samples of 937 Tunisian city names that consist of 33,164 Arabic words (174,690 characters) written by approximately 166 writers. It is used in competitions. The data are available in isolated word samples (not a natural Arabic online handwriting), and no segmentation of the words into letters is provided.

AOD (2012) [9]: A database of Arabic online digits was collected from 300 writers. Each writer was asked to write an average of 10 sample per digit with no constraints on the number of strokes for each digit or the writing style. 30,000 samples were collected. This database may be demanded from AUCEgypt. This database is limited to only digits.

ALTEC (2014) [10]: An online Arabic text database with a large lexicon is produced by the Arabic language technology center (ALTEC). It consists of 152,680 samples of 39,945 unique words, including 325,477 samples of 14,740 unique parts of a word, the database is collected from approximately 1,000 writers where samples are complete sentences that include digits and punctuation marks and the collected data is available on sentence, word and character levels. The main drawback of this database is that the data are collected by using a device digitally captures and stores everything written or drawn with ink on ordinary paper.

QHW (2014) [11]: The Qur'anic handwritten words database is the most commonly used words in the holy Quran. Handwritten words were chosen as the most common words repeated in the holy Quran. The initial version of QHW database includes 120 handwritten words and divided equally into two sets written by 200 writers in total. The QHW database contains 12,000 sample including more than 42,800 characters and 23,300 sub words. This database is a closed vocabulary database and has samples of a limited number of words.

Online-KHATT (2018) [12]: The Online-KHATT database contains more than 80,000 Arabic words written by 623 writers with approx. 801,421 characters using a source text that covers several domains to ensure a wide range of topics. Online-KHATT database may be considered as the largest Arabic online text database in terms of the number of lines written with electronic pens using natural Arabic text, however it ignored dealing with characters on the base of its strokes.

Table 1 - Summary of some Arabic online handwritten databases [13].

Database	Digits	Chars	Words	Writers
LMCA (2008) [4]	30000	100000	500	55
OHASD (2010) [6]	-	19467	3825	48
ADAB (2011) [7]	-	174690	33164	166
AOD (2012) [9]	30000	-	-	100
ALTEC (2014) [10]	-	106433	152680	1001
QHW (2014) [11]	-	42800	12000	200
Online-KHATT (2018) [12]	-	801421	80931	623

III. PROPOSED (AOLAH) DATABASES

We created novel databases for Arabic characters and strokes to help researchers in the field of online Arabic handwriting recognition. The databases name is Aswan online Arabic handwritten (**AOLAH**) databases. These proposed databases are developed based on an algorithm that uses stroke capturing to facilitate recognition of Arabic characters. We used these databases to test the validity of an algorithm developed for preprocessing online Arabic handwritten characters [3]. This algorithm, written by MATLAB, provides a graphical user interface to display the collected data from pen movements. It collects input pen movements and further stores these pen movements in a csv file. The csv file storage is required to retain original pen movements that are required at later stages beginning with preprocessing: the next stage of recognition process, csv file may also be used to verify the input stroke shape. Figure 1 shows the graphical user interface of developed application to collect the databases, with "sheen" character written in 4 strokes, while figure 2 depicts the screenshot of the data stored in the csv file. Figure 3 shows the shape of the collected stroke.

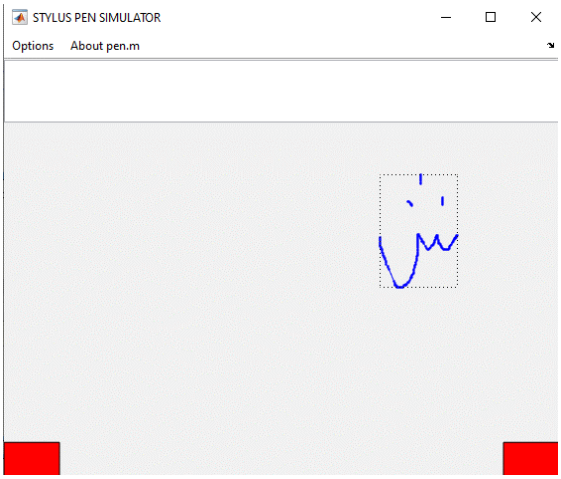


Figure 1 - GUI for stylus pen simulator.

	A	B	C
1	x-co	y-co	stroke
140	0.677679	0.651261	1
141	0.677679	0.654062	1
142	0.677679	0.656863	1
143	0.677679	0.659664	1
144	0.677679	0.662465	1
145	0.677679	0.668067	1
146	0.677679	0.670868	1
147	0.677679	0.673669	1
148	0.677679	0.676471	1
149	0.727679	0.777311	2
150	0.729464	0.777311	2
151	0.73125	0.77451	2
152	0.733036	0.771709	2
153	0.733036	0.768908	2
154	0.734821	0.768908	2
155	0.734821	0.766106	2
156	0.790179	0.788515	3
157	0.790179	0.785714	3
158	0.790179	0.782913	3
159	0.790179	0.780112	3
160	0.790179	0.777311	3
161	0.790179	0.77451	3
162	0.790179	0.771709	3
163	0.790179	0.768908	3
164	0.750893	0.852941	4
165	0.750893	0.85014	4
166	0.750893	0.847339	4
167	0.750893	0.844538	4

Figure 2 - Data stored in csv files.

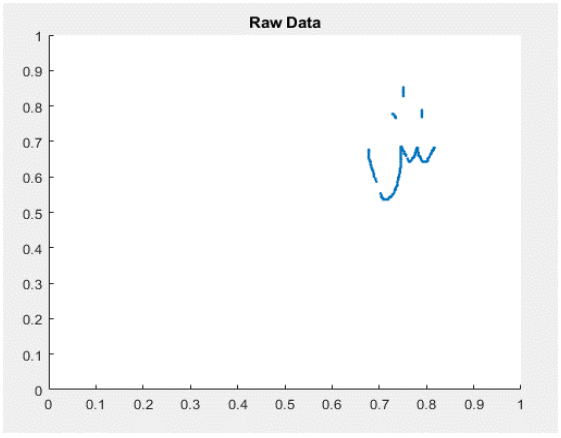
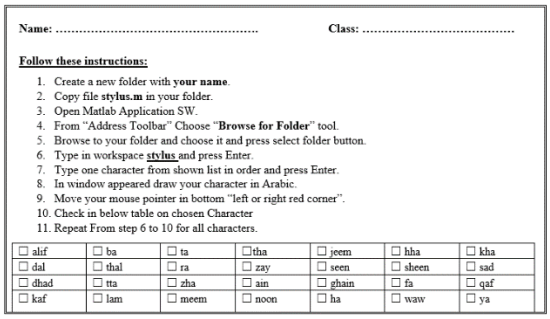


Figure 3 - Shape of the collected strokes.



Name: ..... Class: .....

Follow these instructions:

1. Create a new folder with **your name**.
2. Copy file **stylus.m** in your folder.
3. Open Matlab Application SW.
4. From "Address Toolbar" Choose "**Browse for Folder**" tool.
5. Browse to your folder and choose it and press select folder button.
6. Type in workspace **stylus** and press Enter.
7. Type one character from shown list in order and press Enter.
8. In window appeared draw your character in Arabic.
9. Move your mouse pointer in bottom "left or right red corner".
10. Check in below table on chosen Character
11. Repeat From step 6 to 10 for all characters.

☐ alif	☐ ba	☐ ta	☐ tha	☐ jeem	☐ hha	☐ kha
☐ dal	☐ thal	☐ ra	☐ zay	☐ seen	☐ sheen	☐ sad
☐ dhad	☐ tta	☐ zha	☐ ain	☐ ghain	☐ fa	☐ qaf
☐ kaf	☐ lam	☐ meem	☐ noon	☐ ha	☐ waw	☐ ya

Figure 4 - Data collection form.

Figure 5 - Volunteers folders with data collected inside.

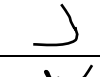
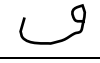
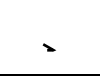
The volunteers are selected from students of Aswan faculty of Engineering, ages from 18 to 20 years old. A form with indications is given to each student without any constraints on the writing style, this form is shown in figure 4, while figure 5 shows examples of folders names, which are created with each student name and are used to save the collected data they write.

All these files are reviewed to guarantee the accepted files for the database. A total of 2,520 files are accepted from 97 different writers, representing 90 files for each character. A second database is extracted from the previous accepted database by extracting strokes from characters. 17 strokes are separated from 28 characters and a database of 1,530 files representing strokes was created, strokes shapes selected with their IDs are shown in table 2. The databases will be available on Aswan university site for free.

IV. CONCLUSION

This paper presented novel Arabic handwritten characters and strokes databases. These databases are focused only on Arabic handwritten characters with Naskh style. A lot of work is needed from researchers to supply Arabic society with this kind of strokes databases; Ruqaa, Thuluth, Diwani are some styles of Arabic language that are needed to be part of the future databases. Furthermore, collecting databases for shapes of Arabic characters depending on their locations in the word is quite needed. Moreover, databases of diacritics will be also of great importance for more advanced character recognition. More volunteers from different ages are needed to make a powerful database.

Table 2 - Arabic strokes shapes with IDs.

Stroke ID	Stroke Shape	Stroke ID	Stroke Shape
100		110	
120		130	
140		150	
160		170	
180		190	
200		210	
220		230	
240		250	
260			

ACKNOWLEDGEMENT

The authors thank all participants' contribution to **(AOLAH)** databases formulation. They sincerely appreciated Dr. Omar Abdel-Reheem and Eng. Fatma Gamal, from Aswan faculty of engineering, Aswan university, for their help to facilitate the data collecting process.

REFERENCES

[1] Hussein AlMuallim and Shoichiro Yamaguchi, "**A Method of Recognition of Arabic Cursive Handwriting**," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vols. PAMI-9, no. 5, pp. 715-722, 1987.

[2] Mustafa Ali Abuzaraida, Akram M. Zeki and Ahmed M. Zeki, "**Recognition Techniques for Online Arabic Handwriting Recognition Systems**," in *2012 International Conference on Advanced Computer Science Applications and Technologies (ACSAT)*, Kuala Lumpur, Malaysia, 2012.

[3] Mohamed AbdElNafea and Samia Heshmat, "**Efficient Preprocessing Algorithm for Online Handwritten Arabic Strokes**," in *2019 International Conference on Innovative Trends in Computer Engineering (ITCE)*, Aswan, Egypt, 2019.

[4] Monji Kherallah, Abdelkarim Elbaati, Haikal El Abed, and Adel M. Alimi, "**The On/Off (LMCA) Dual Arabic Handwriting Database**," in *REGIM: Research Group on Intelligent Machines*, University of Sfax, 2008.

[5] Abdelkarim Elbaati, Monji Kherallah, Abdellatif Ennaji, and Adel M. Alimi, "**Temporal Order Recovery of the Scanned Handwriting**," in *2009 10th International Conference on Document Analysis and Recognition*, Barcelona, Spain, 2009.

[6] Randa I. M. Elanwar, Mohsen A. Rashwan, and Samia A. Mashali, "**OHASD: The First On-Line Arabic Sentence Database Handwritten on Tablet PC**," *International Journal of Computer and Information Engineering,* vol. 4, no. 12, pp. 1907-1912, 2010.

[7] Haikal El Abed, Monji Kherallah, Volker Märgner and Adel M. Alimi, "**On-line Arabic handwriting recognition competition 'ADAB database and participating systems'**," *Document Analysis and Recognition,* vol. 14, pp. 15-23, 2011.

[8] Monji Kherallah, Najiba Tagougui, Adel M. Alimi, Haikal El Abed and Volker Margner, "**Online Arabic Handwriting Recognition Competition**," in *2011 International Conference on Document Analysis and Recognition*, Beijing, China, 2011.

[9] Sherif Abdel Azeem, Maha El Meseery and Hany Ahmed , "**Online Arabic Handwritten Digits Recognition**," in *2012 International Conference on Frontiers in Handwriting Recognition*, Bari, Italy, 2012.

[10] Ibrahim Abdelaziz and Sherif Abdou, "**AltecOnDB: A Large-Vocabulary Arabic Online Handwriting Recognition Database**," *ArXiv,* 2014.

[11] Mustafa Ali Abuzaraida, Akram M. Zeki and Ahmed M. Zeki, "**Online database of Quranic handwritten words**," *Journal of Theoretical and Applied Information Technology,* vol. 62, no. 2, pp. 485-492, 2014.

[12] Sabri A. Mahmoud, Hamzah Luqman, Baligh M. Al-Helali, Galal BinMakhashen and Mohammad Tanvir Parvez, "**Online-KHATT: An Open-Vocabulary Database for Arabic Online-Text Processing**," *The Open Cybernetics & Systemics Journal,* vol. 12, no. 1, pp. 42-59, 2018.

[13] Baligh M. Al-Helali and Sabri A. Mahmoud, "**Arabic Online Handwriting Recognition (AOHR): A Survey**," *ACM Computing Surveys (CSUR),* vol. 50, no. 3, pp. 1-35, 2017.