Chapter 22
Online Arabic Databases and Applications

Houcine Boubaker, Abdelkarim Elbaati, Najiba Tagougui, Haikal El Abed, Monji Kherallah, and Adel M. Alimi

Abstract Large databases were developed for handwriting recognition in Latin script. In contrast, very few databases have been developed for Arabic script, and fewer have become publicly available. This paper describes a pilot study in which we present the nature of the Arabic handwritten language and the basic concepts behind the recognition process. An overview of online Arabic databases and applications presented in the literature is discussed in detail. We also present some related works using these databases.

22.1 Introduction

In the last few years, handwriting analysis and recognition has become a paramount subject of researchers' interest. The validation of the work done in this area was successfully established, thank to the databases used. Two sorts of databases are considered. One type is of interest to online studies (e.g., UNIPEN), and the other is of interest to offline studies (CEDAR, IRONOFF, NIST, IFN/ENIT, etc.). All these

H. Boubaker (✉) · A. Elbaati · N. Tagougui · M. Kherallah · A.M. Alimi
Research Group on Intelligent Machines (REGIM), National School of Engineers ENIS, University of Sfax, BP 1173, Sfax 3038, Tunisia
e-mail: houcine-boubaker@ieee.org

A. Elbaati
e-mail: abdelkarim.elbaati@ieee.org

N. Tagougui
e-mail: najiba.tagougui@ieee.org

M. Kherallah
e-mail: monji.kherallah@ieee.org

A.M. Alimi
e-mail: adel.alimi@ieee.org

H. El Abed
Institute for Communications Technology (IfN), Technische Universität Braunschweig, Schleinitzstrasse 22, 38116 Braunschweig, Germany
e-mail: elabed@tu-bs.de

V. Märgner, H. El Abed (eds.), *Guide to OCR for Arabic Scripts*,
DOI 10.1007/978-1-4471-4072-6_22,

databases are important for the research community in order to test new ideas and algorithms and to perform benchmarks and thereby measure progress and general tendencies.

Since 1997, Märgner et al. have discussed the general concept of benchmarking the output of each module of interest. Their method describes how to build a database with specific ground truth for document analysis systems (DASs) where they focus on the definition and generation of ground truth, especially for the image processing modules of a DAS. They also presented a method to build a database for benchmarking more generally with the use of synthetic data [18]. There exist no freely or commercially available (or Internet accessible) databases of Arabic characters, digits and words in either offline topics or in online topics. An interesting state of the art of Arabic language was established by [7], indicating the difference between the handwritten classical Arabic script as the courant and the modern standard one which represents the majority of use in Arabian countries. The author also gives a summary of techniques concerning Arabic handwriting recognition research. In online studies of the handwriting research topic, Alimi [2] presented a review of online Arabic handwriting recognition systems. He developed an evolutionary neuro-fuzzy approach to recognize Arabic handwritten characters. His experiment consists in testing the performance of the system to recognize Arabic handwritten characters segmented from cursive script. From this task the same writer was asked to write a text extracted from a newspaper containing about 1000 words. These words contain more than 3000 characters (almost all the possible combinations of the 117 Arabic letters were used). The written words were segmented manually into characters, and only the principal component of each character was kept [1, 2]. Mezghani et al. [19] have elaborated a set of 17 basic Arabic isolated letters with 432 samples of each character written by 18 writers. Their experience deals with an online recognition system carried out by a Kohonen neural network trained using an empirical distribution of features such as tangents and tangent differences at regularly spaced points along the character signal. El-Sana presented a recent work which deals with an on-line Arabic handwriting recognition field of the disclosed technique [8]. The method of recognition incorporates delayed strokes and uses a discrete hidden Markov model (HMM) to represent each of the letter shapes in the Arabic alphabet [3]. The dataset used contains between 30,000 and 40,000 Arabic words written without dots.

As a result, until now, there has been no robust standard comprehensive database online or offline for Arabic handwriting script recognition. However, some attempts have been realized, and one of the first databases that was publicly available and became the first standard databases for Arabic is the IFN/ENIT [7] which is an off-line database for Arabic words including 937 Tunisian town/villages names and postal codes written by 411 people. A Persian version of the IFN/ENIT was recently released, including city names handwritten in Farsi. The Persian version consists of 7271 binary images of 1080 Iranian province/city names, collected from 600 writers. For each image in the database, the ground truth information includes its zip code, and a sequence of characters and numbers.

The need to advance Arabic online handwriting recognition systems drives the research community to create and collect online Arabic databases. The validation of

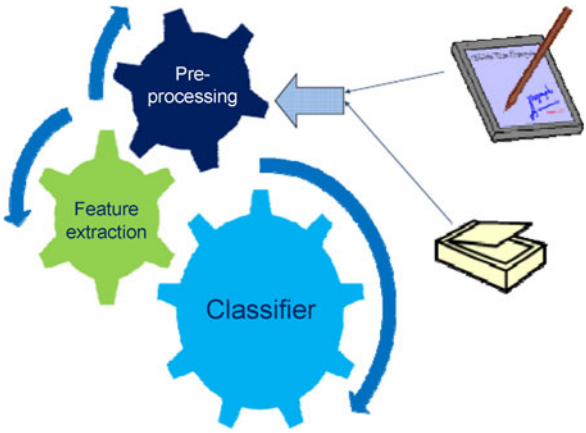


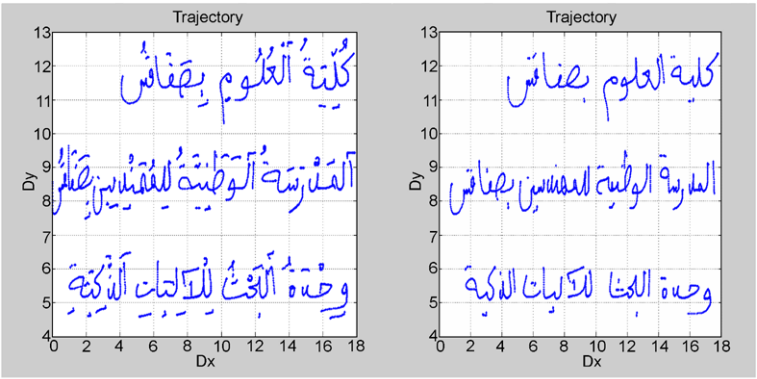
Fig. 22.1 Online/offline handwriting recognition process

the work done in this area cannot be successfully established without common international databases. The objective of this paper is to present standard online Arabic databases which will be important for the research community working in Arabic in order to test new ideas and algorithms and to perform benchmarks and thereby measure progress and general tendencies. This chapter is organized as follows. In Sect. 22.2, the state of the art of Arabic handwriting recognition will be presented. Section 22.3 briefly presents the Arabic script. Section 22.4 presents in detail our LMCA (Lettres, Mots et Chiffres Arabe) database formulation, and to prove the validity of LMCA's structure, some related works will be presented in subsections. In the same way, Sect. 22.5 will be devoted to the ADAB database formulation and to some related works proving its validity. In Sect. 22.6, a conclusion will be presented.

22.2 State of the Art of Arabic Handwriting Recognition

Two axes of research are available in handwriting recognition; the first one is called online, and the second offline. According to Fig. 22.1, using a digital tablet and a special pen offers an interactive dynamic information as a sequence of point coordinates. Using the scanner offers static information as pixels.

The recognition concerns handwritten characters or handwritten words. Three phases are needed for recognition system approval: pre-processing, feature extraction, and classification phases. The advantage of the IRONOFF database is that an offline image and an online trajectory are available. One interest concerns the evaluation of skeleton algorithms. Here, the online data could provide a way to compare the skeleton points (offline image) to an objective trajectory (online coordinates). One could also study the correlation that could exist between the speed of the pen and the gray level distribution or the width of the corresponding strokes. If the online data is jointly accessible with the offline images, it can be used to recover the temporal order of strokes from the offline images and thereby guide and train the segmentation to provide a relevant frame description.



(a) (b)

Fig. 22.2 Arabic handwriting

In that sense, these approaches bridge the gap between online and offline character recognition methods [11, 16], which is a very attractive concept since it has been shown that online handwriting exhibits superior results compared to offline recognition [20].

22.3 Arabic Script Description

Arabic script has been adopted for use in a wide variety of languages other than Arabic, including Persian, Kurdish, Malay, and Urdu. Arabic handwriting is a consonantal and cursive writing. This property is exhibited in two forms: printed or handwritten documents. There are no distinct upper and lower case letter forms. Some Arabic handwritten documents are written with some diacritics (see Fig. 22.2(a)), whereas in the majority of cases only points are considered in Arabic handwriting (see Fig. 22.2(b)). The Arabic alphabet is composed of 28 main characters (with diacritics and in isolated form) and is written from right to left. Most characters have four different shapes. The difference between these letters lies in their positions in the word, the number and the position of the diacritic dots, and the presence of the "Hamza" and vowels (see Table 22.1). In fact, the majority of letters change slightly in shape according to their position in the word (initial, medium, or final). This change occurs when letters are either joined to one another or isolated. There is also a big similarity between some letters [3]. In our work, we reduce the number of letters to 57 by eliminating all diacritics as points and vowels. If we do not consider the first letter of Table 22.1, "Hamza," the number will be reduced to 56 letters.

Table 22.1 represents the 28 letters of the Arabic script in their four different forms. Among them, six letters exist only in the isolated form and in the end form. They are marked with empty columns.

Table 22.1 The Arabic alphabet

Character	alone	end	middle	begin	Character	alone	end	middle	begin
Alif	ا	ـا			Dhad	ض	ـض	ـضـ	ضـ
Ba	ب	ـب	ـبـ	بـ	Taa	ط	ـط	ـطـ	طـ
Ta	ت	ـت	ـتـ	تـ	Dha	ظ	ـظ	ـظـ	ظـ
Tha	ث	ـث	ـثـ	ثـ	Ayn	ع	ـع	ـعـ	عـ
Jim	ج	ـج	ـجـ	جـ	Ghayn	غ	ـغ	ـغـ	غـ
Ha	ح	ـح	ـحـ	حـ	Fa	ف	ـف	ـفـ	فـ
Kha	خ	ـخ	ـخـ	خـ	Qaf	ق	ـق	ـقـ	قـ
Dal	د	ـد			Kaf	ك	ـك	ـكـ	كـ
The	ذ	ـذ			Lam	ل	ـل	ـلـ	لـ
Ra	ر	ـر			Mim	م	ـم	ـمـ	مـ
Zai	ز	ـز			Nun	ن	ـن	ـنـ	نـ
Sin	س	ـس	ـسـ	سـ	He	ه	ـه	ـهـ	هـ
Chin	ش	ـش	ـشـ	شـ	Waw	و	ـو		
Sad	ص	ـص	ـصـ	صـ	Ya	ي	ـي	ـيـ	يـ

22.3.1 Diacritic Symbols Influence

Some Arabic letters have the same form. However, they are distinguished from each other by the addition of dots in different positions relative to the main stroke. Some Arabic characters use special marks to modify the character accent. When diacritical symbols (dots, special marks) are used, they appear above or below the characters and they are drawn as isolated entities as shown in Fig. 22.2. Diacritical symbols are positioned at a certain distance from the character, which makes some difficulties in separating the border of a text line. Indeed, diacritical symbols can generate some redundant separate lines [3]. We count 15 among the 28 letters of the alphabet which contain dots. Some letters present a zigzag shape called "Hamza." It takes the same shape of the letter "Ayn" (see Table 22.1), but it is located above the letter "Alif." The letter "Hamza" is considered as an accent "vowel" in the Arabic alphabet [17].

22.3.2 Pre-processing Step

Pre-processing is primarily related to word processing operations such as normalization to remove handwriting irregularities. Most of the current Arabic word

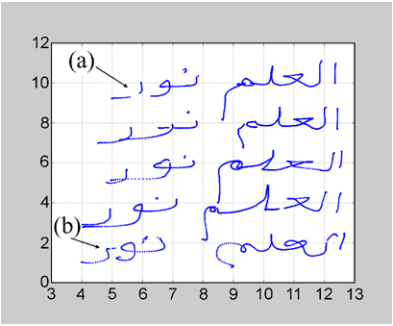


Fig. 22.3 Errors and multivariability existing between writers. (**a**) Stiction of pen-down switchers (bad contact). (**b**) Disconcerted and redundant points

recognition systems do not allow noisy data input. Therefore, current research must deal with the matters of multi-cultural handwriting styles and the adaptation method, which varies from one user to another. In a project on within-writer and between-writer variability, it was found that the number of stroke-shape interpretations in cursive script kept increasing with each new writer in a training system, and the existence of an asymptote was not apparent. The major problems caused by the multivariability are: The input may consist of discrete noise events, like dots or short lines resulting from inadvertently dropping the pen or tapping the pen on the writing surface unwillingly (see Figs. 22.3(a) and 22.3(b)). The input may consist of badly formed shapes, illegible to both humans and machines. Two successive points can be confused by a small segment (see Fig. 22.2(b)). Consequently, the handwriting input may contain device-generated errors: random noise, stiction of pen-down switches, unresponsiveness of switches, pen tilt errors, etc. These problems are shown in Fig. 22.3. In most cases, Arabic writing does not use vowels. The sense of the word is often determined by the context of the sentence. Thus vowels are not considered in our work. We corrected the trajectory by eliminating the majority of noise (diacritics as points and short segments). For this task we developed a simple filter based on distance measurement between successive points of the trajectory. The elimination of isolated points and small segments composed of a number of points is based on threshold optimization. In Figs. 22.4(a) and 22.4(b) we demonstrate the filter effectiveness.

22.4 LMCA Database and Related Works

22.4.1 Database Formulation

A database for character recognition algorithms is of fundamental interest for the training of recognition methods. We developed our own database which contains 30,000 digits, 100,000 Arabic letters, and 500 Arabic words. This database was developed in our laboratory, the REsearch Group on Intelligent Machines (REGIM).

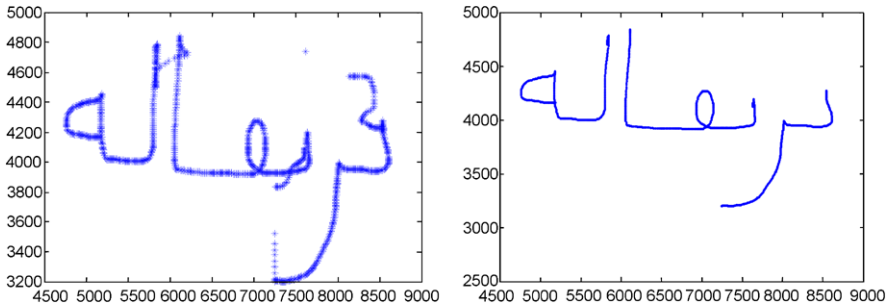


Fig. 22.4 The word "bortoukalaton" (orange) before (**a**) and after (**b**) smoothing

Fig. 22.5 (**a**) Examples of scanned written words. (**b**) Examples of scanned written digits

Both online/offline handwritten characters and words are considered. The online procedure is based on a collection of coordinates (x, y) of the handwritten trajectory, whereas the offline procedure is based on a collection of images of the handwritten trajectory. These two types of information should be available within the same coordinate system, with the same origin and the same resolution and orientation.

Fifty-five participants were invited to contribute to the development of the handwritten LMCA. The dataset of words of each participant is stored in one data file. When producing the data file, each participant was asked to write some Arabic words. We collected 500 words written by different writers. The data for each participant are stored in one data file. For the digits dataset construction, a participant was asked to write a set of all digits (1000 to 1500 samples of digits). We imposed that the writer should just write the same digit ten times, from 0 to 9, on the same page. One page contains 100 digits. The writer was asked to prepare only one page per day. We have collected 30,000 digits in total. More than half of them are regularly written. The remaining ones are those that have noise in the data, are poorly written, or are deliberately written in strange and unusual ways.

Figures 22.5(a) and 22.5(b) present an example of scanned words and digits. They are presented as an image in JPG format. The same procedure was applied to prepare 100,000 Arabic letters. About two-thirds of the writers were male, about 90 percent were right handed, the youngest writer was 8 years old, and the oldest was

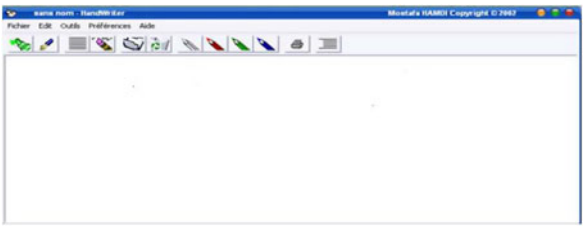


Fig. 22.6 The graphical user interface "Handwriter"

X	Y	Z
9009	6395	0
9012	6400	0
9014	6405	0
9017	6409	0
9019	6414	0
9020	6418	1
9020	6418	1
9020	6422	1
9020	6422	1
9020	6422	1

Fig. 22.7 The x and y coordinates from the digital tablet

66. In the online domain, the forms were sampled with a spatial resolution of 200 dpi and a sampling rate of 100 points/s (Wacom UltraPad A4) and were stored using the UNIPEN format. To collect data, a graphical user interface called "Handwriter" has been developed on a PC/NT window environment (see Fig. 22.6). The online information of the handwriting is kept in a text file. The pen position up and down is detected, respectively, by 0 and 1 values. The trajectory of the handwritten script is collected as coordinates of x and y from the digital tablet (see Fig. 22.7).

22.4.2 Related Works

Online Digit Handwriting Recognition System Based on Trajectory and Velocity Modeling

Digit recognition was studied ten years ago, and it was found that the fuzzy approach enhanced the classification performance. In this study, the feature extraction system was based on the "beta-elliptical" representation [14]. One of the main classification problems is the variability of the feature vector size depending on each digit number of strokes. The recognition process is divided into pre-processing steps and a subsequent classification. To face the complex problems of handwriting recognition, the use of multiple, hybrid and an association of classifier systems has attracted increasing interest during recent years. Based on their complementarities, an association of

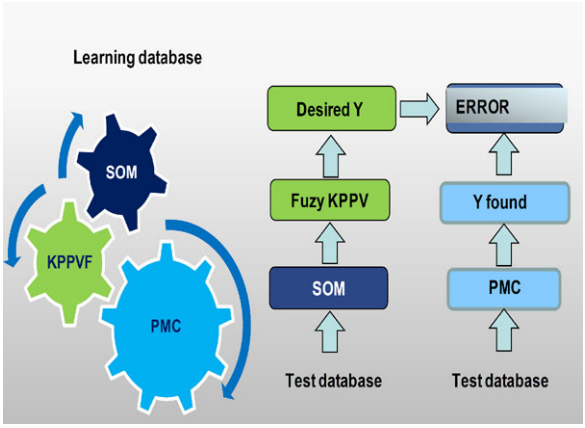


Fig. 22.8 Multiple classifier system

Table 22.2 Comparative study between UNIPEN digit dataset and LMCA digit dataset

Classifier	Modeling system	Dataset	Recognition rate
MLP	Beta-elliptical	30,000 digits LMCA	94.14 %
SVM	Beta-elliptical	Digit set of UNIPEN	94.78 %

classifiers increases the performance of the recognition system while limiting the error bound to the use of a unique classifier. The use of the multiple classifier systems benefits from the strong points of every classifier. In [14], the recognition system is based on the use of neural networks developed in a fuzzy concept. The desired outputs of a multilayer perceptron neural network (MLPNN) are formed using a self-organizing map (SOM) and a fast Kohonen neural network (FKNN) algorithm (see Fig. 22.8). Therefore, this system involves neuro-fuzzy networks based on a SOM and FKNN algorithm association used in the learning process [14]. The global recognition rate obtained is about 95.08 %. When testing our system, the global average squared error obtained is about 0.065. In this study, our aim is to validate the use of a digit set extracted from the LMCA dataset. It is known that the MLP and support vector machine (SVM) techniques give the same performance. The first experience was based on the use of the LMCA digit dataset, whereas the second one was based on the use of the UNIPEN digit dataset. The results obtained were similar, which proves that the developed LMCA digit dataset has a correct format benchmark (see Table 22.2).

Recognition of a Handwritten Arabic Word Based on Visual Encoding and GA

A handwritten word is represented by a continuation of visual codes of Arabic letters. In this case the order of these letters is considered. We attribute N the number of basic letters extracted from a cursive word [15]. Therefore, every gene of the

Table 22.3 Fitness value calculation

Visual indices	Va "1"	Lo "2"	Po "3"	Al "4"	Des "5"	As "6"	Roc "7"	Loc "8"	Loa "9"	Roa "10"	Ain "11"	Sad "12"	# "13"
Va "1"	0	1	1	1	1	1	1	1	1	1	1	1	1
Lo "2"	1	0	1	1	1	1	0.5	0.5	1	1	0.5	0.5	1
Po "3"	1	1	0	1	0.5	1	1	1	1	1	1	1	1
Al "4"	1	1	1	0	1	0.5	1	1	0.5	0.5	1	1	1
Des "5"	1	1	0.5	1	0	1	1	1	1	1	1	1	1
As "6"	1	1	1	0.5	1	0	1	1	0.5	0.5	1	1	1
Roc "7"	1	0.5	1	1	1	1	0	1	1	1	0.5	0.5	1
Loc "8"	1	0.5	1	1	1	1	1	0	1	1	0.5	0.5	1
Loa "9"	1	1	1	0.5	1	0.5	1	1	0	0.5	1	1	1
Roa "10"	1	1	1	0.5	1	0.5	1	1	0.5	0	1	1	1
Ain "11"	1	0.5	1	1	1	1	0.5	0.5	1	1	0	0.5	1
Sad "12"	1	0.5	1	1	1	1	0.5	0.5	1	1	0.5	0	1
# "13"	1	1	1	1	1	1	1	1	1	1	1	1	0

population has N chromosomes and every chromosome has one of the 58 possible values (1 to 57 for the basic Arabian characters and the value 0 for characters with more than one visual indication) numbered from the right to the left. The extraction rate obtained is about 72 %. However, in the second stage, which consists in correcting the weaknesses of the previous method, we developed a genetic algorithm (GA) in order to select the best combination of visual codes extracted from a word by the heuristic method [13]. The GA approach permits the recognition of cursive handwriting without the limitation of a lexical dictionary [12]. Therefore, the convergence of the GA is ensured by the technique given in the fitness function which consists in the use of the visual codes of Arabic words and the comparison method established between the visual indices strings according to Table 22.3. The number of generations (500) and the fitness value (0.5) were fixed as a convergence condition criterion. If the population size was fixed to 100 individuals, the recognition rate was about 99.85 %. These results are encouraging. In this experiment we used the 500 words and the 57 Arabic letters extracted from the LMCA dataset. 200 words were used as data prototypes for the selection of the initial population of the GA, and the others were used for testing our system.

Order Temporal Reconstruction from Arabic Image Word

The word image captured in gray level with a resolution of 300 dpi will be preprocessed in four stages: binarization, filtering, extraction of the skeleton, and elimination of the diacritical signs (see Fig. 22.9(a)). A suitable algorithm segments the skeleton in three types of segments: segments of connection, occlusion, and segments of end of stroke. The starting segment is localized by sweeping the image of the skeleton from the right to the left, and more tests are applied. Another algorithm makes it possible to order these segments based on heuristic rules. These rules count on the fact that Arabic script is written from right to left, and they take into account the natural order of stroke generation [10]. To validate this approach we tested it on a whole of the words extracted from the LMCA dataset. The temporal order signal which is reconstructed will be compared with its original online trajectory signal (see Fig. 22.9(b)) [9].

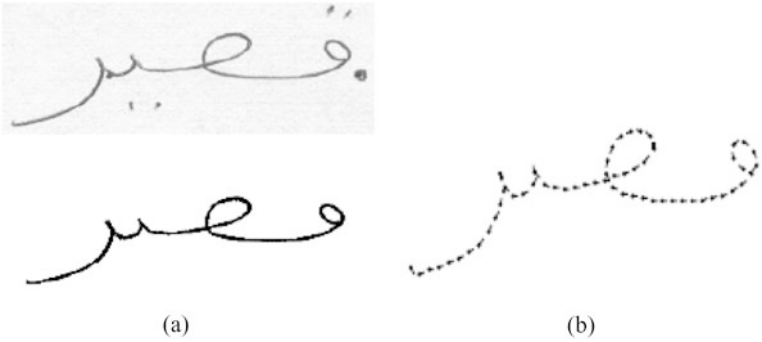


Fig. 22.9 (**a**) The Arabic word "kassiron" before and after pre-processing. (**b**) Restoration of the temporal order of the offline word "kassiron"

22.4.3 Conclusion

The different research works and their results prove that our LMCA database can be used in both modeling and recognition systems of Arabic handwriting. The related works presented prove also that LMCA is a standard database and it has the same format of the common UNIPEN or IRONOFF database. Our perspective is to increase the number of writers of LMCA to help make it perform for any techniques of modeling and classification of handwritten Arabic script.

22.5 ADAB Formulation and Related Works

22.5.1 ADAB Formulation

The Arabic DAtaBase (ADAB) was developed to advance the research and development of Arabic online handwritten text recognition systems. This database is developed as a cooperative effort between the Institut fuer Nachrichtentechnik (IfN) and the National School of Engineers of Sfax (ENIS), Research Group on Intelligent Machines (REGIM) [7]. The database consists of 19,575 Arabic words handwritten by more than 150 different writers, most of them selected from the narrower range of ENIS. The text that is written is from 937 Tunisian town/village names. We plan to extend this database with other Arabic writing styles. For this reason, we have developed special tools for the collection of the data and verification of the ground truth, which will be available for other groups for the collection of their own data in the same form of the ADAB. These tools allow one to record the online written data, to save some writer information, to select the lexicon for the collection, and to rewrite and correct wrong written text. Ground truth was added to the text information automatically from the selected lexicon and verified manually. The ADAB is freely available for noncommercial research

Table 22.4 ADAB sets

Set	Files	Words	Characters	Writers
1	5037	7670	40,500	56
2	5090	7891	41,515	37
3	5031	7730	40,544	39
4	4417	6786	35,832	25
Sum	19,575	30,077	158,420	157

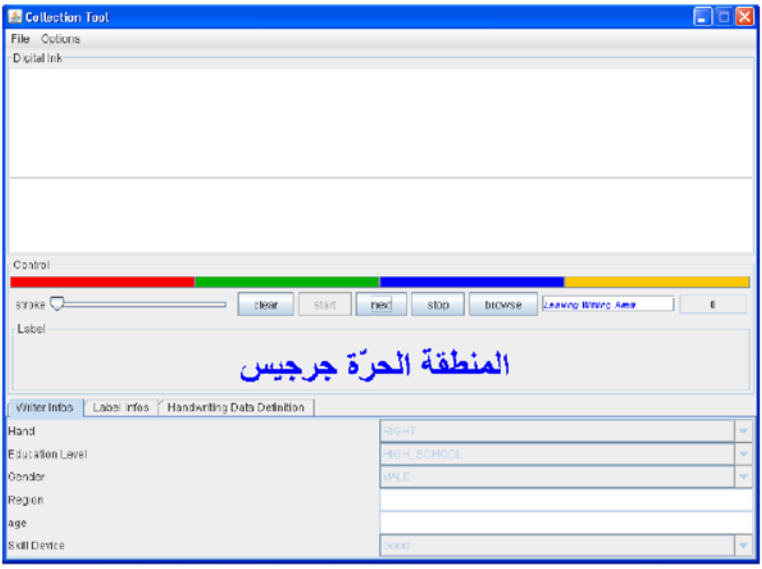


Fig. 22.10 The ADAB's collection tool

(www.regim.org) [7]. Our aim was to collect a database of handwritten town names written in a quality similar to that on a mobile phone with a digital input device. The collection process starts when the writer clicks on start bottom. The collection tool generates a town name randomly from 937 Tunisian town/village names, and the writer must write the displayed word (see Fig. 22.10). A pre-label will be automatically assigned to each file. It consists of the postcode in a sequence of numeric character references which will be stored in the UPX file format. An InkML file including trajectory information and a plot image of the word trajectory is also generated. Additional information about the writer can also be provided (see Fig. 22.11).

The ADAB is divided into four sets. Details about the number of files, words, characters, and writers for each set 1 to 4 are shown in Table 22.4.

Nom	Taille	Type
1267267556453	2 Ko	Fichier UPX
1267267583812	2 Ko	Fichier UPX
1267268596625	2 Ko	Fichier UPX
1267268790859	2 Ko	Fichier UPX
1267268813984	2 Ko	Fichier UPX
1267268877562	2 Ko	Fichier UPX
1267268895250	2 Ko	Fichier UPX
1267268926328	2 Ko	Fichier UPX

Nom	Taille	Type
1267267556453.inkml	3 Ko	Fichier INKML
1267267583812.inkml	5 Ko	Fichier INKML
1267268596625.inkml	4 Ko	Fichier INKML
1267268790859.inkml	5 Ko	Fichier INKML
1267268813984.inkml	5 Ko	Fichier INKML
1267268877562.inkml	6 Ko	Fichier INKML
1267268895250.inkml	5 Ko	Fichier INKML
1267268926328.inkml	6 Ko	Fichier INKML

Nom	Taille
1267267556453	2 Ko
1267267583812	6 Ko
1267268596625	4 Ko
1267268790859	3 Ko
1267268813984	4 Ko
1267268877562	4 Ko
1267268895250	6 Ko
1267268926328	6 Ko

Fig. 22.11 Samples of UPX files and their corresponding InkML and image files

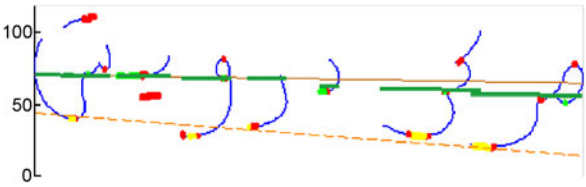


Fig. 22.12 Example of detected baseline correction (*green*) obtained from the consideration of topological conditions

22.5.2 Related Works

Online Arabic Handwriting Modeling System Based on Grapheme Segmentation

An online Arabic handwriting modeling system based on grapheme segmentation is presented. The system consists of three modules: detection of the baseline, grapheme segmentation, and feature extraction. The method developed in the first module is distinguished by the consideration of geometrical and topological features for the baseline detection and correction. In the second module, we use the detected baseline to check particular points: the back of the valleys and the angular points for the segmentation of the cursive handwriting trajectory in graphemes. The third module extracts parameters to model the position, the shape, and the fuzzy affectation rate of diacritics associated to each segmented grapheme. Figure 22.12 shows an example of a correct result of baseline detection obtained from the consideration of the topologic conditions compared with the results of the basic stage [5].

In the evaluation phase, the system is applied on the online database ADAB of Tunisian town names using the HMM Toolkit (HTK) as the classification module. The following recognition results for three ameliorated versions of the system were obtained (Table 22.5):

- Version 1: without diacritics detection (ICDAR 2009 competition) [4].
- Version 2: after adjusting the filters and without diacritics detection.
- Version 3: after adjusting the filters and with the extraction and fuzzy affectation of diacritics.

Table 22.5 Recognition rate obtained on ADAB sets 1 and 2

System version	ADAB set 1		ADAB set 2	
	Top 1	Top 5	Top 1	Top 5
Version 1	57.87	72.89	54.26	66.38
Version 2	86.38	96.43	83.55	94.68
Version 3	82.33	93.47	80.61	91.53

ICDAR Competition

The first international online Arabic handwriting recognition competition was held in 2009 [7]. The International Conference on Document Analysis and Recognition (ICDAR) is an international scientific conference in the field of document processing and image analysis. Occurring every two years, it brings together researchers from universities and businesses from all around the world. To compare the performance of the participants' systems, the ADAB was used. As part of the competition, seven systems have been benchmarked by independent leading domain experts. Among the tested recognition systems, several systems have been developed by specialized university laboratories such as the REGIM laboratory. Among the industry players, the Vision Objects MyScript® system proposes natural handwriting recognition technology for Arabic. The systems were tested on known data (sets 1 to 3) and on one test dataset which was unknown to all participants (set 4). The accuracy rate and the recognition speed were measured. With 99 % accuracy rates, the MyScript® system has the highest recognition rates, almost 4 % higher than the second best system participating in the competition. Regarding the recognition speed, Vision Objects won over its competitors hands down with an average processing time of 69 ms per word: more than 25 times faster than the second fastest system in the competition.

The competition results, presented in Table 22.6, show that Arabic handwritten word recognition systems have further made remarkable progress within recent years. Most of the participating systems show a very high accuracy, and some also perform at very high speed [6].

22.5.3 Conclusion

Online recognition of cursive Arabic handwritten words aims to contribute in the evolution of online Arabic handwriting recognition research. Since 2009 the freely available database ADAB is used by some research groups all over the world to develop online Arabic handwriting recognition systems. This database was the basis for the competition ICDAR 2009 for systems that are specialized in online recognition of cursive Arabic handwritten words, which confirms that it is a database with reliable matter.

Table 22.6 Recognition results in % of correct recognized images on reference datasets 1, 2, and 3 and on a subset of dataset 4

System	set 1			set 2			set 3			set 4*		
	top 1	top 5	top 10	top 1	top 5	top 10	top 1	top 5	top10	top 1	top 5	top10
MDLSTM-1	99.36	99.94	99.96	99.42	99.96	100.00	99.52	99.94	99.94	95.70	98.93	100
MDLSTM-2	98.55	99.60	99.66	98.77	99.88	99.92	98.89	99.64	99.70	95.70	98.93	100
VisionObjects-1	99.46	99.70	99.70	99.82	99.94	99.96	99.58	99.76	99.76	**98.99**	100	100
VisionObjects-2	99.29	99.60	99.60	99.51	99.74	99.74	99.26	99.56	99.56	98.99	100	100
REGIM-HTK	57.87	72.89	77.03	54.26	66.38	71.06	53.75	72.31	76.22	52.67	63.44	64.52
REGIM-CV	100	100	100	94.39	96.06	96.06	96.28	97.14	97.52	13.99	31.18	37.63
REGIM-CV-HTK	28.85	51.92	55.77	35.75	58.30	64.26	30.60	52.80	62.80	38.71	59.07	69.89

22.6 Discussion and Future Work

This work was about creating a good quality database with support for online handwritten Arabic script recognition. This type of database has many uses, including training and testing a recognition system. Two databases were created: the first one is the LMCA database which contains 30,000 digits, 100,000 Arabic letters and 500 Arabic words, and the second one is the ADAB, which contains 19,575 samples of 937 Tunisian town/village names. These databases were created by collecting writing contributions from more than 200 Arab persons of different age and sex. The input information can be digits, characters, or words. Many of the contributors expressed that it was unnatural to write on a Wacom or a Genius tablet because during writing the pen's trace isn't visible directly on the tablet as it is on a piece of paper. This could have affected the quality of the collected material and thus the final database. But it is not a real major limit, since the information collected will be used by recognition systems devoted to the recognition of handwritten script on small mobile devices where it is impossible to treat the writing style. To confirm the efficiency of these databases, many works were evaluated, and we have reported the results. The use of the ADAB to test and compare the participating systems in the ICDAR competition proves that it is really a standard database with consistent content. We plan to extend this database with other Arabic writing styles. This database will be important for the research community in order to test new ideas and algorithms and to perform benchmarks and thereby measure progress and general tendencies. We plan to expand these databases so that they became a reference in the field. Our perspective is also to try to expand this data by considering other writers of the Eastern countries.

Acknowledgements The authors thank all participants' contributions to the LMCA and ADAB databases formulation. In addition, they acknowledge the financial support of this work by grants from the General Direction of Scientific Research and Technological Renovation (DGRST), Tunisia, under the ARUB program 01/UR/11/02.

References

1. Alimi, A.M.: A neuro-fuzzy approach to recognize on-line Arabic handwriting. In: Proc. of Int. Conf. on Neural Networks, vol. 3, pp. 1397–1400, August 1997
2. Alimi, A.M.: An evolutionary neuro-fuzzy approach to recognize on-line Arabic handwriting. In: Proc. of the International Conference on Document Analysis and Recognition (ICDAR), pp. 382–386 (1997)
3. Biadsy, F., El-Sana, J., Habash, N.: Online Arabic handwriting recognition using hidden Markov models. In: Proc. of the Tenth International Workshop on Frontiers in Handwriting Recognition (2006)
4. Boubaker, H., Kherallah, M., Alimi, A.M.: New algorithm of straight or curved baseline detection for short Arabic handwritten writing. In: Proc. of the 10th International Conference on Document Analysis and Recognition (ICDAR) (2009)
5. Boubaker, H., El Baati, A., Kherallah, M., El Abed, H., Alimi, A.M.: Online Arabic handwriting modeling system based on the grapheme segmentation. In: Proceedings of the 20th International Conference on Pattern Recognition (ICPR), pp. 2061–2063 (2010)
6. El Abed, H., Kherallah, M., Märgner, V., Alimi, A.M.: ICDAR 2009—Arabic online handwriting recognition competition. In: Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR), vol. 3, pp. 1388–1392, July 2009
7. El Abed, H., Kherallah, M., Märgner, V., Alimi, A.M.: On-line Arabic handwriting recognition competition—ADAB database and participating systems. Int. J. Doc. Anal. Recognit. **14**(1), 15–23 (2011)
8. El-Sana, J., Biadsy, F.: On-line Arabic handwriting recognition field of the disclosed technique (2010)
9. Elbaati, A., Boubaker, H., Kherallah, M., El Abed, H., Ennaji, A., Alimi, A.M.: Arabic handwriting recognition using restored stroke chronology. In: Proc. of the International Conference on Document Analysis and Recognition (ICDAR) (2009)
10. Elbaati, A., Kherallah, M., Alimi, A.M., Ennaji, A.: De l'hors-ligne vers un système de reconnaissance en-ligne: application à la modélisation de l'écriture Arabe manuscrite ancienne. In: Proc of the Semaine du Document Numerique, SDN (ANAGRAM) (2006)
11. Jäger, S.: Recovery dynamic information from static, handwritten word images. Ph.D. thesis, Daimler-Benz AG Research and Tech., Verlag Dietmar Fölbach (1998)
12. Jouini, B., Kherallah, M., Alimi, A.M.: A new approach for on-line visual encoding and recognition of handwriting script by using neural network system. In: Proc. of the International Conference on Artificial Neural Networks and Genetic Algorithms (ICANNGA), Roanne, France, pp. 161–167 (2003)
13. Kherallah, M., Bouri, F., Alimi, A.M.: Toward an online handwriting recognition system based on visual coding and genetic algorithm. In: Proc. of the International Conference on Adaptive and Natural Computing Algorithms, Coimbra, Portugal, pp. 502–505 (2005)
14. Kherallah, M., Hadded, L., Mitiche, A., Alimi, A.M.: On-line recognition of handwritten digits based on trajectory and velocity modeling. Pattern Recognit. Lett. **29**, 580–594 (2007)
15. Kherallah, M., Bouri, F., Alimi, A.M.: On-line Arabic handwriting recognition system based on visual encoding and genetic algorithm. Eng. Appl. Artif. Intell. **22**(1), 153–170 (2009)
16. Lallican, P.M., Viard-Gaudin, C.: Off-line handwriting modeling as a trajectory tracking problem. In: Proc. of the 6th International Workshop on Frontiers in Handwriting Recognition (IWFHR), Taejon, Korea, August 1998, pp. 347–356 (1998)
17. Lorigo, L.M., Govindaraju, V.: Offline Arabic handwriting recognition: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **28**(5), 712–724 (2006)
18. Märgner, V., Karcher, P., Pawlowski, A.-K.: On benchmarking of document analysis systems. In: Proc. of the 4th International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 331–336 (1997)

19. Mezghani, N., Mitiche, A., Cheriet, M.: On-line recognition of handwritten Arabic characters using a Kohonen neural network. In: Proc. of the 8th International Workshop on Frontiers in Handwriting Recognition (IWFHR), Niagara-on the-Lake, Ontario, Canada, pp. 490–495. (2002)
20. Seiler, R., Schenkel, M., Eggimann, F.: Off-line cursive handwriting recognition compared with on-line recognition. In: Proc. International Conference on Pattern Recognition (ICPR), Vienna, pp. 505–509 (1996)