International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:4, No:12, 2010

OHASD: The First On-Line Arabic Sentence Database Handwritten on Tablet PC

Randa I. M. Elanwar, Mohsen A. Rashwan, and Samia A. Mashali

***Abstract*—**In this paper we present the first Arabic sentence dataset for on-line handwriting recognition written on tablet pc. The dataset is natural, simple and clear. Texts are sampled from daily newspapers. To collect naturally written handwriting, forms are dictated to writers. The current version of our dataset includes 154 paragraphs written by 48 writers. It contains more than 3800 words and more than 19,400 characters. Handwritten texts are mainly written by researchers from different research centers. In order to use this dataset in a recognition system word extraction is needed. In this paper a new word extraction technique based on the Arabic handwriting cursive nature is also presented. The technique is applied to this dataset and good results are obtained. The results can be considered as a bench mark for future research to be compared with.

***Keywords*—**Arabic, Handwriting recognition, on-line dataset.

I. Introduction

HANDWRITING is one of the most important ways in which civilized people communicate. It is used both for personal (e.g. letters, notes, addresses on envelopes, etc.) and business communications (e.g. bank checks, tax and business forms, etc.) between person and person and for communications written to ourselves (e.g. reminders, lists, diaries, etc.) [1].

Despite long standing predictions that handwriting, and even paper itself, would become obsolete in the age of the digital computer, both persist. The reason that handwriting persists in the age of the digital computer is the convenience of paper and pen as compared to keyboards for numerous day-to-day situations.

Computers are becoming ubiquitous as more people than ever are forced into contact with computers and our dependence upon them continues to increase, it is essential that they become more friendly to use. As more of the world's information processing is done electronically, it becomes more important to make the transfer of information between people and machines simple and reliable. Thus the daily increasing information written on paper has to be digitized.

For making editable digital copies of handwritten documents, Handwriting recognition is used to transform text represented in the spatial form of graphical marks into its symbolic representation.

The field of handwriting recognition can be split into two different approaches. The first of these, on-line, deals with the recognition of handwriting captured by a tablet or similar touch-sensitive device, and uses the digitized trace of the pen to recognize the symbol. In this instance the recognizer will have access to the x and y coordinates as a function of time, and thus has temporal information about how the symbol was formed. The second approach concentrates on the recognition of handwriting in the form of an image, and is termed off-line. In this instance only the completed character or word is available.

Extensive research has been carried out in terms of technical papers and reports by various researchers around the world. Research into handwriting recognition still continues to be intense after all these years. The motivation may be attributed to the challenging nature of the character recognition problem and the countless number of commercial applications that it may be applied to [2].

Until the beginning of the nineties, on-line handwriting recognition research was mainly academic and most results were reported in the open literature. The situation has changed in the past few years with the rapid growth of the pen computing industry. Because of the very harsh competition, many companies no longer publish in peer-reviewed literature and no recent general survey is available.

In the last few years, academic research has focused on cursive script recognition. Performances are reported on different databases and are difficult to compare [3]. This shows how essentially we need to have comparative evaluation.

Comparative evaluation consists of a set of participants that compare the results of their systems using the same or similar control tasks and related data with metrics that are agreed upon. The results are presented and compared, while the methods used are discussed and contrasted.

Evaluation has the following advantages: First, the comparative element persuades participants to deliver the best results possible. Second, the developers benefit from evaluation because complete evaluation toolkits and by-product data become available afterwards. Third, institutions receive the possibility to evaluate their own technology in

Manuscript received June 25, 2010.
Randa I. Elanwar is Assistant Researcher in computers and systems dept., Electronic research institute, Cairo, Egypt (phone: 202-33310515; e-mail: eng_r_i_elanwar@yahoo.com).
Mohsen A. Rashwan is Professor of Digital Signal Processing, Electronics and communication dept., Cairo University, Cairo, Egypt (e-mail: Mohsen_Rashwan@rdi-eg.com).
Samia A. Mashali is Professor of Digital Signal Processing, Computers and systems dept., Electronic research institute, Cairo, Egypt (e-mail: samia@eri.sci.eg).

relation to the state of the art.

It may happen that better results are obtained by some participants because they have used data of better quality. Therefore, evaluation will help identify better data, not only better techniques. It also contributes to assess the impact that the quality of data has on system performance.

A side effect of evaluation is often the production of high quality resources. Data are distributed to the participants in order to help them with training and testing their systems. As the participants need the data, there is an imperative to provide data of good quality and in due time. Consequently, we can claim that evaluation provides a partial solution to the problem through the production of standardized, annotated and validated linguistic resources at a low cost.

Dozens of handwriting datasets have been published since 1990s. Off-line datasets had been more publically available before on-line datasets for handwriting recognition. Starting from digit or letter datasets to city name, further to sentence and that application fields have expanded from small lexicon domains, such as bank check reading and address recognition, to large lexicon and general unconstrained domains. Constructing datasets is a costly and time consuming process but the availability of public datasets allows researchers to pay more attention to enhance their systems rather than collecting large datasets to work on.

Public on-line databases like UNIPEN and IRONOFF for English, HCL-2000 and HIT-MW for Chinese and other datasets for other languages are available for researchers to evaluate and compare handwriting recognition systems. Very high accuracies are achieved so far. Unfortunately the success obtained with these handwriting recognition systems has not readily transferred to Arabic. This may be attributed to the immense variation in language nature and the lack of human language resources (especially large public datasets) used to enhance recognition results.

For Arabic the only publically available dataset is the IFN/ENIT offline handwritten dataset. It is a dataset of isolated words representing a limited vocabulary lexicon of 937 Tunisian city names. For on-line Arabic datasets, the same institution introduced a new dataset called ADAB [4] at the tenth International Conference on Document Analysis and Recognition (ICDAR'2009). It has been presented in a handwriting recognition competition for systems output evaluation, but it is not yet launched for public use. The database in version 1.0 consists two parts: the training part and the test part. The training part is composed of of 15158 Arabic words handwritten by more than 130 different writers, most of them selected from the narrower range of the l'Ecole Nationale d'Ing`enieurs de Sfax (ENIS). The text written is from 937 Tunisian town/village names. The ADAB-database is split in 3 sets. Details about the number of files, words, characters, and writers for each set 1 to 3 are shown in Table 1. The test part is composed of 1562 files, 2418 words and 12648 characters written by 24 writers.

TABLE I
FEATURES OF ADAB-DATASETS 1, 2, AND 3

Set	Files	Words	Characters	Writers
1	5037	7670	40500	56
2	5090	7851	41515	37
3	5031	7730	40544	39
sum	15158	23251	122559	132

The recognition of isolated words in a large vocabulary is already a demanding problem. If we attempt to recognize phrases and sentences, linguistic constraints have to be used to limit the search space. The use of a language model brings great gains in recognition accuracy. But a lexicon which limits the search to a set of permitted words is not the only solution. Grammars can also be used to limit which words are permissible in a given context to account for the frequencies of different words. However, grammars are typically used at the word level in the recognition of phrases and sentences, but not on isolated words [1].

In the short term, to meet the accuracy requirements of industry applications, it is important to focus on simplified recognition tasks such as limited vocabulary hand printed character recognition. In the long term, however, research should be challenged by harder tasks, such as large vocabulary cursive recognition [3].

The rest of the paper is organized as follows. Section 2 describes the design of OHASD. In Section 3 the data acquisition process is presented. Section 4 gives an overview of the procedure for dataset word extraction. Experiments and results are presented in Section 5, and finally Section 6 draws some conclusions and gives an outlook for future work.

II. OHASD DATASET

Handwriting is an acquired skill and clearly one that is a complex perceptual-motor task, sometimes referred to as a neuromuscular task [5]. There are two fundamental fields of study pertaining to handwriting:

1. The study of handwriting as a neuromuscular activity, its development as a skill and the effect upon it of various internal and external factors.
2. The study of handwriting identification as a discriminatory process. The second uses knowledge acquired through the first, but is entirely independent of it [5].

Handwriting identification derives from the comparison of writing habits, and an evaluation of the significance of their similarities or differences.

Many studies pursued the correlation of writing features and various medical and mental conditions. Others sought to identify the affects of social status, self-esteem, and sex upon handwriting. Much work dealt with the pedagogy of writing and remedial approaches to improve its quality in the writing of children. Perhaps the greatest effort was devoted to the correlation of writing features and particular personality characteristics, commonly called graphology or

graphoanalysis. A large portion of this work sought to find evidence validating the claims of the vocation as a valid and reliable instrument for personnel selection, aptitude determinations, or as a psycho-diagnostic tool [5].

According to Huber and Headrick [5], the rate at which the speed of writing increases is greatest between the ages of 7 and 9 years. It tapers off to 13 years, when there is little further increase. Writing is a culture-bound activity, not only insofar as language and its orthography, but also in many motor aspects that are greatly influenced by culture and education. Studies have also shown that fatigue and weakness have their effects upon the control of the writing instrument and its performance may be altered in unpredictable fashions. They claim also that changes in health may affect the fluency, rather than the designs of writing. Its effect is usually observed in a loss of control and the introduction of more erratic movements. The same can be noted in case of persons under emotional stress or those experiencing special influences of medications, drugs or alcohol. Furthermore, the nature of a document can have an effect upon the writing of signatures applied to it. Certainly wills, mortgages, large contracts, and real estate transactions are significant events in the lives of many persons, and it is understandable that the signing of such documents will be a more conscious act than it is in signing many others.

Even the writing conditions and circumstances: like writing on a knee, writing in a moving vehicle, in bed or on desk, also have their effects. Similarly, concentration on the act of writing has its effect also as consciousness steals from fluency. The action becomes more deliberate and slower as concentration increases. The change will be noted in line quality unless the reason for concentration is due to uncertainty as to the letters or text to be written that may interrupt the writing process more profoundly.

To compose a large vocabulary dataset from people at different ages, different cultures or education levels, different emotional, health or mental conditions, different writing circumstances, etc. is extremely hard, costly and time consuming. It needs a team work with the help of multi-institutional supervision. So, as a start we take the initial step by collecting the first sentence dataset for on-line handwriting recognition of Arabic.

Our dataset is constructed by 56 volunteers of both genders ranging from 23 to 40 years thus excluding personal variations due to age. They are all working at research centers (governmental and private ones). They are well-educated people who represent the potential users of handwriting recognition, such as personal notes and manuscripts transcription.

To make them feel like note taking in some meeting, the sentences they are supposed to write are dictated to them. Some of them wrote on knee and some wrote on table. They were all not familiar to using tablets thus were highly concentrating on the writing event. To relief tension the sentences topics are chosen from four different subjects: Economic, Political, Law and Sports.

Our sentence dataset is inspired by IAM-Database [6] and IAM-OnDB datasets [7], thus complete sentences are carefully selected from public daily news, printed and dictated to writers. Each writer is asked to write paragraphs with sentences ranging from 15 to 46 words.

After excluding erratic/illegible handwritings, the final version of our dataset is composed of 154 paragraphs written by 48 writers, having a total of 3825 words and 19,467 characters and available with the corresponding ground truth text files. The details of dataset collection are illustrated in the next section.

III. Acquisition

The design of dataset collection tool layout is simple and clear as illustrated in figure 1. It is divided into two distinct blocks: actions buttons block and guideline block. The layout is supplied by a vertical slider to extend the space reserved for handwriting.

When opening a new layout for writing, the user presses the ‘Record’ button and the cursor switches to the writing status and any movement of the stylus on the top of tablet is seen as a dark line on the screen. Note that the label on the ‘Record’ button is changed to ‘Pause’. The pen positions together with the time stamps are simultaneously recorded. The guiding lines are used to avoid any tilt while writing.

When the writer finishes, he presses the ‘Pause’ button and the cursor returns to the idle status. The ‘Play’ button re-draws or re-writes the recorded handwritings in the same order they were written. The ‘Delete’ button deletes any specified stroke number.

The required dataset sentences are dictated to writers with few restrictions: writing in naskh font, not using punctuations or diacritics, not using digits.

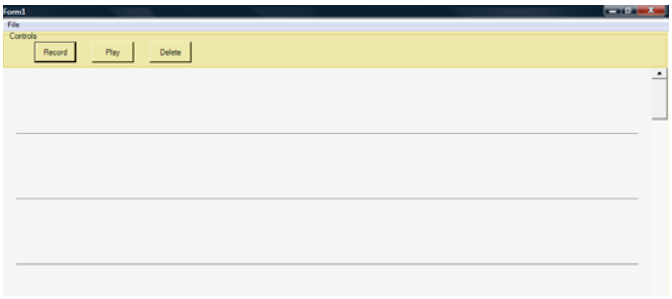

Fig. 1 An illustration of layout

The raw data stored usually includes several consecutive lines of text. For using our dataset by a recognizer, we need to segment the text into individual lines and words. Most word extraction techniques of on-line data are guided by heuristic rules as in [8 and 9]. Ratzlaff [8] uses a “bottom-up” clustering of discrete strokes into increasingly larger groups that eventually merge to complete text lines. The initial clustering is based on the strong evidence of spatiotemporal proximity. Subsequent merging is based on more sophisticated metrics that include dependencies on estimates of inter-line

distance and mean character height. The Y-axis projection histogram is generated for each stroke, then the initial bottom-up clustering began by creating Forward Projection (FP) groups. Strokes are merged into FP groups if they are temporally adjacent and have strongly overlapping Y-axis projections. A single unmerged stroke becomes an independent FP group. This procedure will face effective drawbacks if it is going to be applied to Arabic script due to the small complementary strokes "secondaries" occurring above and below text lines and having null overlapping Y-axis projections, creating a number of independent FP groups. In addition to this the writers tend to write in a very irregular pattern causing large base line skews among the text line and even within one word.

Loudon et al [9], presents a methodology that successfully works with English script due to limited cursive nature, i.e. the stroke (pen down/up movement) usually represents a single character. They calculate parameters like stroke width, and spaces between adjacent strokes, etc. as shown in figure 2.

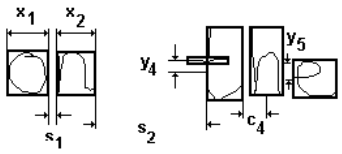

Fig. 2 Some parameters for the handwritten phrase “on the”.

This procedure will also face effective drawbacks if it is going to be applied to Arabic script due to the cursive nature of language, which means that the stroke in Arabic word usually represents more than one character (3-5 characters on average) which makes it impossible to estimate or even expect the Arabic stroke geometry (height, width, etc.). Don't forget also that the English characters usually have one or two delayed (complementary) strokes written immediately after the main stroke (e.g. t, i, j, A, etc.) which is not the case in Arabic strokes as it was noticed that most writers tend to put the secondaries after writing most or even the whole word. In addition to what was mentioned, the character size (scale) and writing sequence (order) varieties among writers, and sometimes for a single writer, make the problem more difficult.

Our new technique is a compromise between the two previous ideas. We do what best suits the Arabic writing nature. The same bottom-up clustering concept in [8] using the spatiotemporal relations between strokes and using parameters similar to those in [9] are used to build the smallest possible FP groups instead of separating the whole text line then splitting it to smaller groups. This helps us to overcome any base line skew and have more accurate estimate about the stroke height. The whole method is described in the following section.

IV. Word Extraction Technique

By examining the states of successively written Arabic strokes (either main-type strokes or complementary-type like dots for example), they are found related spatially to each other by one of the following relations:

1. *Touching*: then the two strokes should belong to the same word (strokes 7,8 in fig 3).
2. *Not touching but having an x-axis histogram overlap:* then the two strokes should belong to the same word (strokes 1,3 in fig 3).
3. *Neither touching nor overlapping on x-axis*: If the inter-stroke distance is less than the average stroke width, then the two strokes should belong to the same word (strokes 1,2 in fig b). Otherwise, the two strokes should belong to two different words (strokes 2,5 in fig 3).

Following these relations results in several independent groups of strokes. Each group of strokes contains the main and complementary strokes of the same word regardless the sequence/order by which they were written.

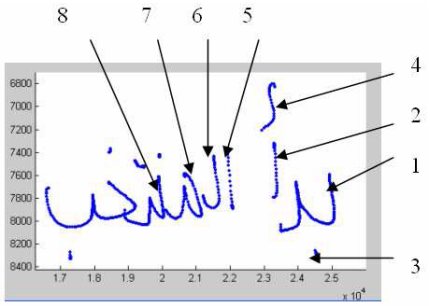

Fig. 3 The successive Arabic strokes states.

The inter-stroke distance is taken to be equal to the average stroke width of the previously written strokes of the same FP group. This estimate works quite well especially that the very small size of secondaries compensate for the presence of long strokes (consisting of 3 or more characters) resulting in a suitable estimate.

V. Experiments and Results

The handwritten Arabic sentence dataset we discuss in this paper addresses several important aspects not covered by other datasets. It is the first sentence dataset. It is naturally written by multiple writers. As a result, there are real text lines and real handwriting phenomena, such as miswriting and erasing.

The word extraction experiments are conducted for the purpose of getting a first impression of how difficult the word separation is, considering that writers don't pay enough attention while writing to make the intra-word spaces less than inter-word spaces. This may be attributed to the fact that when a normal pen on paper, the writer’s paw/palm usually rests on a rigid surface, while writing on electronic surface makes the writer raise his hand which puts much more stress on the writers’ hand. Therefore we must expect more noise and

distortions in handwritings.

Although our separation techniques does not correctly extract all the dataset words meaningfully, but the results were satisfying, as shown in figure 4. You will find some FP groups containing one complete word, others split the word in two or more portions (due to too large inter-stroke distances) and others include more than one word (due to too small inter-stroke distances). The method succeeds to separate 53% of the dataset words meaningfully. 40.7% of the words were under-segmented (stuck) and 7.4% are over-segmented. This result can be considered as a bench mark for any future research to be compared with.

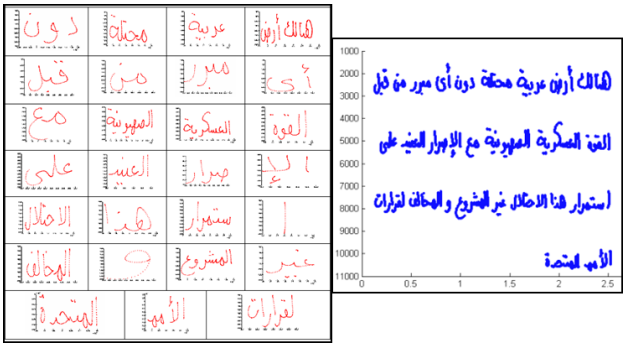

Fig. 4 The result of the word extraction process.

Additionally, our technique solves many significant problems facing the text line extraction in other on-line systems:

1. The problem of "secondaries" having null overlapping Y-axis projections, that are usually separated as an independent text line.
2. The text line separation difficulties due to the presence of base line skew.
3. The delayed stroke problem.
4. The estimation of the used parameters (heuristics).

VI. Conclusions and Future Work

Handwriting is a skill that is personal to individuals. It consists of artificial graphical marks on a surface; its purpose is to communicate something. Newer technologies such as personal digital assistants (PDAs) and digital cellular phones have an impact on handwriting. On-line recognition of handwriting has been an active subject of research since years.

Language engineering research that has the potential to lead to commercial success in the near future is based on data. Pubic datasets for on-line handwriting recognition are essential for researchers, developers and Institutions for results enhancement and product evaluation. Sentence datasets are more beneficial than isolated words datasets because they allow systems to use language resources and achieve better results.

Our dataset is the first sentence dataset for Arabic language. It is real, natural, simple and clear. It consists of more than 3800 words and more than 19,400 characters written by 48 writers of ages 23-40 years, both genders and well-educated. The dataset is collected on a layout described in details. Each file contains the data of a whole paragraph so words had to be extracted.

On-line English handwriting text line extraction techniques, depending on the y-axis histogram projection and character geometry (width, height, etc.), do not function well when applied to Arabic handwriting due to the special characteristics of Arabic handwriting. Thus, we have thought of a new technique, more suitable to the Arabic writing nature. Its performance is considered satisfactory even with presence of some under-segmented words and some over-segmented words. The main reason beyond this is that the writers didn't pay enough attention to make the spaces within one word less than those between words but we can consider our results as a bench mark for any future research to compare with.

Following the work in [10], we determine our future volunteers to be college students, government clerk graduated from higher school because according to the handwriting theory, the handwriting goes into a stable and consistent state at 25 years old, and after that there is little change. Secondly, the college students are enrolled throughout the country, so the handwriting by them can be seen as samples from the whole country.

Additionally, we want to classify writers' datasets according to gender, handedness, age and education to study the influence of variations in these categories. We aim to enlarge the dataset to cover almost all the vocabulary in Arabic lexica and then we can upload it together with the corresponding ground truth on the internet for public use.

Acknowledgment

The authors deeply thank everyone who took the time for participating in the recordings.

References

[1] A. L. Koerich, Large Vocabulary Off–line Handwritten Word Recognition, PHD Thesis, Ecole de Technologie Sup´erieure, Universit´e du Qu´ebec, 2002.

[2] M. Blumenstein, Intelligent Techniques for Handwriting Recognition, PHD Thesis, Faculty of Engineering and Information Technology, Griffith University, 2000

[3] R. Cole, Survey of the state of the art in human language technology, Cambridge University Press, New York, USA,1997, Ch. 2, Pages: 513-537, ISBN:0-521-59277-1.

[4] H. El Abed, V. Märgner, M. Kherallah, A. M. Alimi, "ICDAR 2009 On-line Arabic Handwriting Recognition Competition", 10th International Conference on Document Analysis and Recognition, 2009, ISBN: 978-0-7695-3725-2.

[5] R. A. Huber and A. M. Headrick, Handwriting Identification: Facts and Fundamentals, CRC Press LLC, New York, 1999.

[6] U.-V. Marti and H. Bunke, "The IAM-database: an English sentence database for offline handwriting recognition", International Journal of Document Analysis and Recognition, 2002, vol. 5, pp. 39 – 46.

[7] M. Liwicki and H. Bunke, "IAM-OnDB - an On-Line English Sentence Database Acquired from Handwritten Text on a Whiteboard", Proceedings of the Eighth International Conference on Document Analysis and Recognition table of contents, 2005, pp. 956 - 961, 2005, ISBN ~ ISSN:1520-5263 , 0-7695-2420-6

[8] E.H. Ratzlaff, "Inter-line Distance Estimation and Text Line Extraction For Unconstrained On-line Handwriting", Proceedings of the Seventh

International Workshop on Frontiers in Handwriting Recognition, 2000, pp 33-42.
[9] G. Loudon, O. Pellijeff, LI Zhong-Wei, "A Method for Handwriting Input and Correction on Smartphones", Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition, 2000, pp 481-485.
[10] T. Su, T. Zhang, and D. Guan, "HIT-MW Dataset for Offline Chinese Handwritten Text Recognition", Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition, 2006.