QUWI: An Arabic and English Handwriting Dataset for Offline Writer Identification

Somaya Al Maadeed, Wael Ayouby, Abdelâali Hassaïne, Jihad Mohamad Aljaam
Computer Science and Engineering Department
College of Engineering, Qatar University
Doha, Qatar
$\{s\_alali, wael.ayouby, hassaine, jaam\}$@*qu.edu.qa*

Abstract

This paper presents a new offline dataset called the Qatar University Writer Identification dataset (QUWI). This dataset contains both Arabic and English handwritings and can be used to evaluate the performance of offline writer identification systems. It consists of handwritten documents of 1017 volunteers of different ages, nationalities, genders and education levels. The writers were asked to copy a specific text and to generate a random text, which allows the dataset to be used for both text-dependent and text-independent writer identification tasks. We describe the gathering and processing steps and define several evaluation tasks regarding the use of this dataset.

1. Introduction

It is difficult for forensic detectives to manually analyze and identify writers from handwritten documents, especially when there are many complicated and similar samples. Intelligent writer identification systems, which are gaining popularity, can automatically recognize the authors of such documents and can help investigators to achieve their objectives quickly and obtain more accurate results. The performance of such systems should be evaluated using a large, standard dataset that includes hundreds of documents by different writers. As far as we know, there is no sufficiently comprehensive, well-designed standard dataset that is annotated and publicly available for handwritten Arabic documents. In this paper, we present the QUWI dataset which contains handwritten documents of 1017 writers in both Arabic and English. This paper is organized as follows: Section 2 presents a quick review of the previous datasets which are used in the field of writer identification. Section 3 gives a detailed description of the QUWI dataset. In section 4, characteristics of this dataset are discussed. Finally, the paper ends with some conclusions and remarks about the availability of the new dataset.

2. Existing datasets

Several handwriting datasets that can be used for writer identification exist, in this section, we give an overview of such datasets.

The "IRONOFF" database [19] contains 1000 digitized documents in English and French written by 700 different writers and can be used for both offline and online tasks.

The most commonly used dataset in writer identification is probably "IAM" [14] which contains 1539 digitized English offline documents written by 657 writers. An online version of this dataset has been created as well [12], it contains more than 1700 digitized documents for 221 different writers.

In 2007, "CASIA Handwriting Dataset" has been created, it contains online Chinese and English handwritings and it has later been extended [3].

In that same year, the French offline "RIMES dataset" [10] has been created, it contains 12723 pages of more than 1300 different writers.

Two relatively small offline datasets have been created in the Indian Statistical Institute, one in Bengali language consisting of 80 digitized documents by 40 writers [8] and one in Telugu, consisting of 110 handwritten documents from 22 writers [16].

The CVC-MUSCIMA is another interesting dataset that can be used for the identification of musicians based on their written music scores. It contains 1000 music sheets written by 50 different musicians [7].

The CEDAR dataset contains a handwritten English

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.256

letter copied by as many as 1500 writers representative of the US population. This dataset can be used for text-dependent writer identification, but is unfortunately not publicly available [17].

In 2009, the Arabic MADCAT dataset [18] has been created, it contains about 10000 handwritten pages of about 325 writers. This dataset is unfortunately not publicly available. Moreover, the number of writers is not high enough when compared with other state-of-the-art datasets.

The BBN technologies also holds an offline Arabic dataset [2]. It includes 39500 documents by 259 writers. Each document contains an average of 20 lines and 100 words, but this dataset is not publicly available and the number of writers is also not high enough in this dataset.

The Spanish forensic laboratory also collected a Spanish language dataset of 30 different writers in which each writer has 300 character samples [5].

A team at the AmirKabir University of Technology in Iran created two Farsi handwriting datasets. The first one contains handwriting of 40 writers and the second one contains handwritings of 180 persons [9, 15].

The IFN/ENIT dataset [4] contains approximately 2200 handwritten binary images of 411 Tunisian individuals. This dataset has mainly been used for Arabic handwriting recognition, but can also be used for writer identification.

The AHDB dataset [1] contains Arabic words used in filling out the numbers on checks. In addition, it contains some sentences that are used in writing checks in Arabic. It also includes the most popular words in Arabic writing and a free handwriting page from each writer's imagination. The dataset contains 105 folders including 315 documents written by only 105 writers; thus, it is mainly used for handwritten text recognition and is publicly available.

Several writer identification contests have been organized recently, each of these contests have provided a benchmarking dataset. The first one includes 208 handwritings for 26 different writers in Latin languages [13]. The second one used the CVC-MUSCIMA dataset previously mentioned [6]. The last one used an Arabic dataset "AWIC2011" of 54 writers, each writer produced three paragraphs, among which two have been used for training and one for testing [11].

Table 1 sums up this section by giving a comparison between all these datasets and the QUWI dataset.

In the next section, we give a detailed description of the QUWI dataset, its corresponding collection and acquisition processes as well as its structure.

3. QUWI dataset description

The QUWI dataset contains documents for 1017 writers. The development of this database is significant because of its size, the number of writers, their diversity (nationality, age, background, etc.), and the variety of the dataset in terms of the included languages as well as pens and pencils used and the varying colors and thicknesses of the handwritings. Note that blue and black pens are the most often used by volunteers.

Volunteers were first asked to fill out an information page that includes the name, age, gender, handiness, writer's profession, educational level, and nationality (Figure 1). This page is used to create an anonymous Excel file that contains this information regarding all the writers. The name is however kept confidential for privacy reasons.

Volunteers were then instructed to copy in their natural handwriting four pages such that: The first one contains approximately six handwritten lines in the Arabic language from the writer's imagination (or copied from a newspaper or from whatever source). The second page contains an Arabic text of three paragraphs to be copied by all the writers. Similarly, the third page contains about six handwritten lines in English from the writer's imagination and the fourth page contains an English text to be copied by all the writers. The first and the third pages are to be used for text-independent writer identification tasks, whereas the second and fourth page are to be used for text-dependent writer identification tasks. Figure 2 shows an examples of such pages.

Note that a few writers have used French instead of English when producing the third page, but this occurred with only 34 writers. Also, a few writers left the English pages blank as they do not master it (yet).

Each writer needed approximately 35 minutes to complete the forms because of the significant number of lines of writing required. Some individuals completed the writing tasks in approximately 20 minutes whereas others required more than 40 minutes depending on their writing speed. It was important to assist the volunteers to ensure that they are writing the correct phrases, using the correct pens and generating the correct number of pages. There were some people who know Arabic but had not practiced writing for some time. There were also some volunteers who were beginners in English.

To ensure more diversity, we tried to ensure that each writer wrote each page with a different pen or pencil. Sometimes, we gave a second volunteer the same pen that the first volunteer used for one page and required the first volunteer to use another pen for his or her next page. Thus, the same color was used on various pages

Table 1. Comparison between datasets in offline mode

Name of the dataset	Language	Writers	Documents	Availability
CEDAR	English	1500	1500	Proprietary
IAM	English	657	1539	Public
IRONOFF	English & French	700	>1000	Proprietary
RIMES dataset	French	1300	12723	Through competitions
ISI dataset1	Bengali	40	80	Upon request
ISI dataset2	Telugu	22	110	Upon request
BBN dataset	Arabic	259	39500	Proprietary
Spanish forensic dataset	Spanish	30	-	Proprietary
LDC dataset	Arabic	70	7447	Only to LDC members
AmirKabir dataset	Farsi	180	540	Unknown
IFN/ENIT-dataset	Arabic	411	2200	Public
AHDB dataset	Arabic	105	315	Public
IIT-Demokritos	Latin languages	26	208	Public
CVC-MUSCIMA	Music Scores	50	1000	Public
AWIC2011	Arabic	54	161	Public
QUWI	Arabic & English	1017	5085	Through competitions & commercially

by multiple writers, and the same writer was able to use different pens or the same pen on different pages. The pens were selected randomly by page number and/or by writer. The aim of this step is to prevent algorithms from identifying the correct writer based on the pen or pencil used.

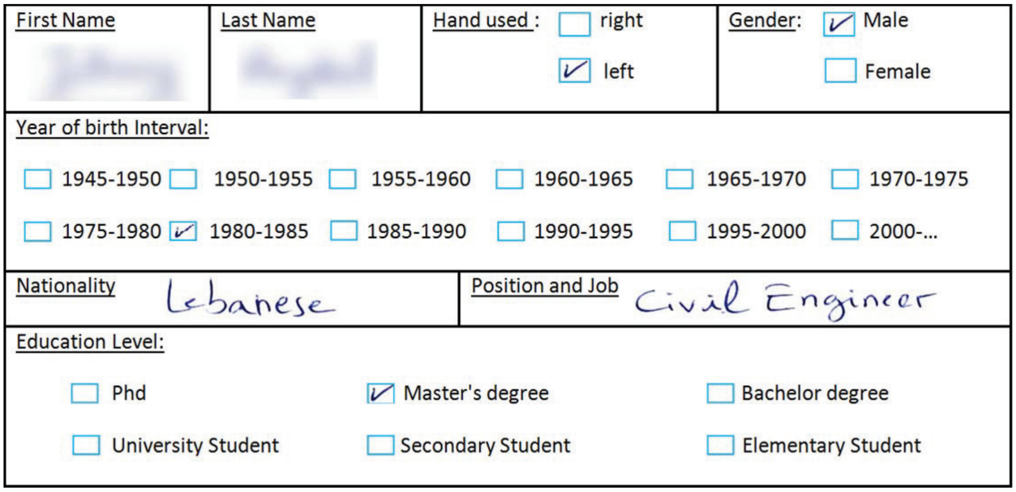
The dataset is structured in 1017 folders. Each folder contains four scanned documents, the dataset thus contains a total number of 4068 documents. Digitization has been performed with an EPSON GT-S80 autofeeder scanner, using a 600 DPI resolution and a lossless color JPEG format. Each folder contains the handwriting data for the same writer using the same pen or possibly different pens or pencils.

4. Dataset Analysis

The QUWI dataset is diverse. The novelty and real advantage of the dataset are the diversity of writers, of languages and all criteria. The analysis of the dataset indicates that 306 of the volunteers are Qataris, approximately 190 are Lebanese, 101 are Palestinians, 104 are Egyptians, and 68 are Jordanians. There are also many Sudanese, Yemenis, Syrians, Iranians, Iraqis nad Saudis individuals represented in the dataset. The variation in the educational levels of the participants is also interesting: the dataset includes not only highly educated but also less educated people. Indeed, the volunteers included elementary school students, secondary school students and university students as well as workers, employees, engineers, doctors, professors, accountants, and secretaries in industrial, administrative and academic environments. The volunteers also differed in age. There are volunteers younger than 12 years, teenagers, and adults and older than 40 years. The dataset was 52% written by females (530 writers) and 48% written by males (487 writers). 953 of the volunteers (93.7%) are right handed whereas 64 volunteers (6.3%) are left handed. These factors (nationality, age range, handiness, educational level, and gender) are important elements of our dataset. Figure 3 shows the distribution of writers in this dataset with respect to each of these factors.

In total, the dataset contains 4068 digitized pages. It contains approximately 60,000 words written in Arabic by 1017 writers (around 60 words per writer) for text-independent analysis and more than 100,000 Arabic words written by the same 1017 writers for text-dependent analysis. Similarly, it contains around 60,000 words for text-independent analysis and more than 100,000 words for text-dependent analysis in English.

In addition to writer identification, the dataset might be useful for many other research areas including the



First Name	Last Name	Hand used :	Gender:
		☐ right ☑ left	☑ Male ☐ Female

Year of birth Interval:

☐ 1945-1950 ☐ 1950-1955 ☐ 1955-1960 ☐ 1960-1965 ☐ 1965-1970 ☐ 1970-1975
☐ 1975-1980 ☑ 1980-1985 ☐ 1985-1990 ☐ 1990-1995 ☐ 1995-2000 ☐ 2000-...

Nationality Lebanese | Position and Job Civil Engineer

Education Level:

☐ Phd ☑ Master's degree ☐ Bachelor degree
☐ University Student ☐ Secondary Student ☐ Elementary Student

Figure 1. First Personal Information Page.

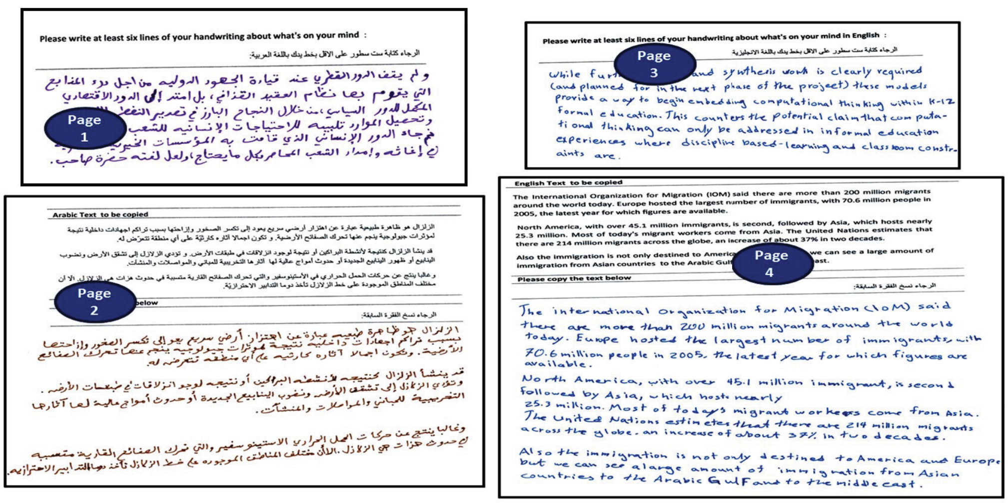


Figure 2. Format of documents per writer.

identification of the gender and handiness of a specific writer, as well as his or her age and nationality.

5. Conclusion

A dataset that contains handwritten Arabic and English handwritings has been described in this paper. This dataset is unique due to the variety of handwritings, the languages used and the backgrounds of the writers in terms of genders, nationalities, ages and educational levels.

This dataset will serve as a benchmarking dataset for the development and evaluation of systems in writer identification, as well as the identification of the gender, age range, handiness and nationality of different writers.

It is planned to organize several contests for those tasks. This dataset will be made public progressively through the participation in these evaluations campaigns.

The dataset is also being annotated at the word-level in order to make it useful for handwriting recognition purposes.

Acknowledgment

This work is supported by Qatar National Research Fund (QNRF) grant through National Priority Research Program (NPRP) No. 09 - 864 - 1 – 128.
We would also like to thank all the volunteers who dedicated their valuable time and effort in the data collection process.

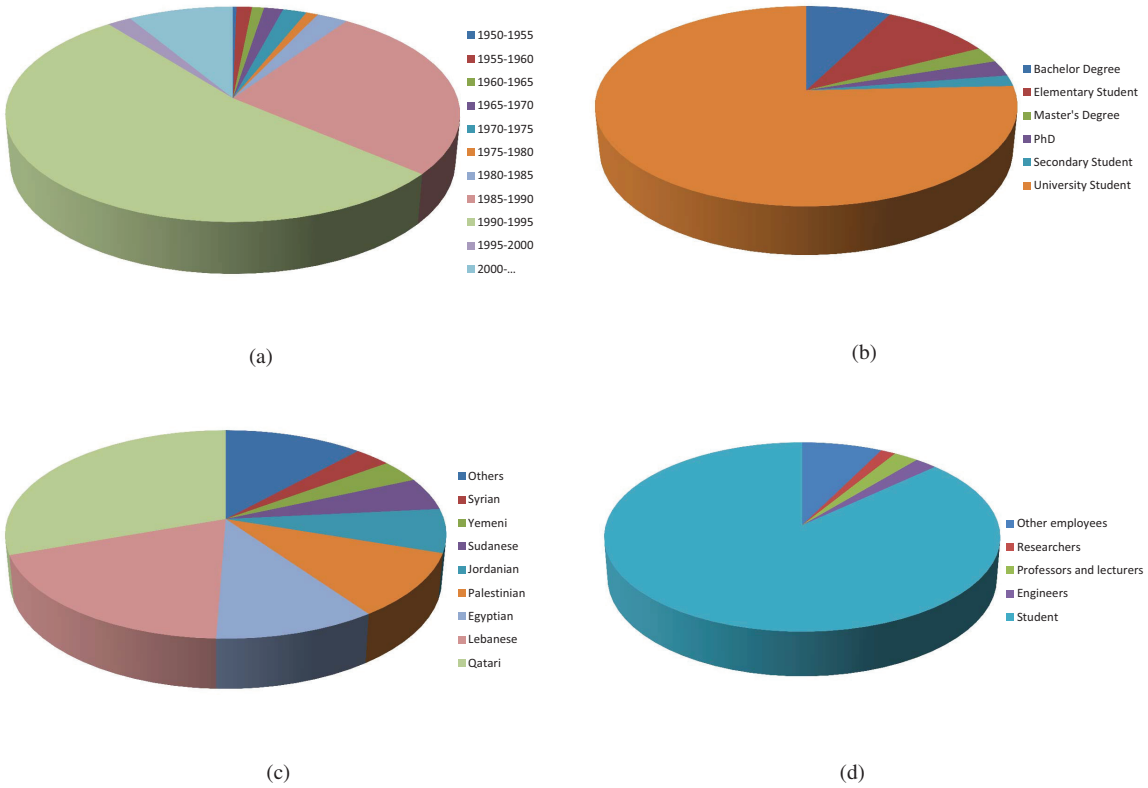


Figure 3. Distribution of writers with respect to (a) Ages, (b) Educational levels, (c) Nationalities and (d) Positions.

References

[1] S. Al-Maadeed, D. Elliman, and C. Higgins. A Dataset for Arabic Handwritten Text Recognition Research. *The International Arab Journal of Information Technology*, 2004.

[2] H. Cao, R. Prasad, and P. Natarajan. Improvements in hmm adaptation for handwriting recognition using writer identification and duration adaptation. In *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*, pages 154 –159, nov. 2010.

[3] L. Cheng-Lin, Y. Fei, W. Da-Han, and W. Qiu-Feng. CASIA Online and Offline Chinese Handwriting Databases. In *Document Analysis and Recognition, International Conference on*, pages 37–41, Los Alamitos, CA, USA, 2011. IEEE Computer Society.

[4] H. El Abed and V. Margner. The IFN/ENIT-database - a tool to develop Arabic handwriting recognition systems. In *Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on*, pages 1 –4, feb. 2007.

[5] R. Fernandez-de Sevilla, F. Alonso-Fernandez, J. Fierrez, and J. Ortega-Garcia. Forensic Writer Identification Using Allographic Features. In *Proceedings of the 2010 12th International Conference on Frontiers in Handwriting Recognition*, ICFHR '10, pages 308–313, Washington, DC, USA, 2010. IEEE Computer Society.

[6] A. Fornés, A. Dutta, A. Gordo, and J. Lladós. Cvc-muscima: a ground truth of handwritten music score images for writer identification and staff removal. *International Journal on Document Analysis and Recognition*, pages 1–9, 2011. 10.1007/s10032-011-0168-2.

[7] A. Fornés, A. Dutta, A. Gordo, and J. Lladós. The ICDAR 2011 Music Scores Competition: Staff Removal and Writer Identification. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 1511 –1515, sept. 2011.

[8] U. Garain and T. Paquet. Off-Line Multi-Script Writer Identification Using AR Coefficients. In *Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*, ICDAR '09, pages 991–995, Washington, DC, USA, 2009. IEEE Computer Society.

[9] G. Ghiasi and R. Safabakhsh. An Efficient Method for Offline Text Independent Writer Identification. In *Pattern Recognition (ICPR), 2010 20th International Conference on, title=An Efficient Method for Offline Text Independent Writer Identification*, pages 1245 –1248, aug. 2010.

[10] E. Grosicki, M. Carr, J. Brodin, and E. Geoffrois. RIMES evaluation campaign for handwritten mail processing. In *Proc. of the Int. Conf. on Frontiers in Handwriting Recognition*, 2008.

[11] A. Hassaïne, S. Al-Maadeed, J. Alja'am, A. Jaoua, and A. Bouridane. The ICDAR2011 Arabic Writer Identification Contest. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 1470 –1474, sept. 2011.

[12] M. Liwicki and H. Bunke. IAM-OnDB - an On-Line English Sentence Database Acquired from Handwritten Text on a Whiteboard. In *Proceedings of the Eighth International Conference on Document Analysis and Recognition*, ICDAR '05, pages 956–961, Washington, DC, USA, 2005. IEEE Computer Society.

[13] G. Louloudis, N. Stamatopoulos, and B. Gatos. ICDAR 2011 Writer Identification Contest. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 1475 –1479, sept. 2011.

[14] U.-V. Marti and H. Bunke. A Full English Sentence Database for Off-Line Handwriting Recognition. In *Proceedings of the Fifth International Conference on Document Analysis and Recognition*, ICDAR '99, pages 705–, Washington, DC, USA, 1999. IEEE Computer Society.

[15] F. Nejad and M. Rahmati. A New Method for Writer Identification and Verification Based on Farsi/Arabic Handwritten Texts. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02*, ICDAR '07, pages 829–833, Washington, DC, USA, 2007. IEEE Computer Society.

[16] P. Purkait, R. Kumar, and B. Chanda. Writer Identification for Handwritten Telugu Documents Using Directional Morphological Features. In *Proceedings of the 2010 12th International Conference on Frontiers in Handwriting Recognition*, ICFHR '10, pages 658–663, Washington, DC, USA, 2010. IEEE Computer Society.

[17] S. Srihari, S.-H. Cha, H. Arora, and S. Lee. Individuality of handwriting: a validation study. In *Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on*, pages 106 –109, 2001.

[18] S. Strassel. Linguistic resources for Arabic handwriting recognition. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, April 2009.

[19] C. Viard-Gaudin, P. Lallican, P. Binter, and S. Knerr. The IRESTE On/Off (IRONOFF) Dual Handwriting Database. In *Proceedings of the Fifth International Conference on Document Analysis and Recognition*, ICDAR '99, pages 455–, Washington, DC, USA, 1999. IEEE Computer Society.