A Database for Offline Arabic Handwritten Text Recognition

Sabri A. Mahmoud, Irfan Ahmad,
Mohammed Alshayeb, and Wasfi G. Al-Khatib

King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia
{smasaad,irfanics,alshayeb,wasfi}@kfupm.edu.sa

Abstract. Arabic handwritten text recognition has not received the same attention as that directed towards Latin script-based languages. In this paper, we present our efforts to develop a comprehensive Arabic Handwritten Text database (AHTD). At this stage, the database will consist of text written by 1000 writers from different countries. Currently, it has data from over 300 writers. It is composed of an images database containing images of the written text at various resolutions, and a ground truth database that contains meta-data describing the written text at the page, paragraph, and line levels. Tools to extract paragraphs from pages, segment paragraphs into lines have also been developed. Segmentation of lines into words will follow. The database will be made freely available to researchers world-wide. It is hoped that the AHTD database will stir research efforts in various handwritten-related problems such as text recognition, and writer identification and verification.

Keywords: Arabic Handwritten Text Database, Arabic OCR, Document Analysis, Form Processing.

1 Introduction

Recent advances in pattern recognition have aided the automation of many demanding tasks in our daily life. Algorithmic analysis of human handwriting has many applications such as on-line and offline handwritten text recognition, writer identification and verification, bank check processing, postal address recognition, etc. Arabic is one of the Semitic languages. According to Ethnologue [1], it is ranked as the fourth most widely spoken language in the world. It is spoken by more than 280 million people and is the official language of 22 countries.

Researchers consider the lack of freely available Arabic handwritten recognition databases as one of the reasons for the lack of research on Arabic text recognition compared with other languages [2]. There is no generally accepted database for Arabic handwritten text recognition that covers the naturalness of the Arabic language. Hence, different researchers of Arabic text recognition use different databases. Therefore, the recognition rates of the different techniques may not be comparable. This raises the need for a comprehensive database for Arabic text recognition [3]. Having such a database is crucial for Arabic handwritten text recognition and identification research. Arabic text is written from

M. Kamel and A. Campilho (Eds.): ICIAR 2011, Part II, LNCS 6754, pp. 397–406, 2011.

ا ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن هـ و ي

(a)

ء أ ب ج

(b) (c) (d) (e)

Fig. 1. (a) The 28 basic chararcters in the Arabic language, (b) Hamza, (c) Letter ALEF with Hamza, (d) Letter BAA and (e) Letter JEEM.

right to left, with Arabic having 28 basic characters as shown in Figure 1(a). Out of the 28 basic characters, 16 of them have from one to three dots. Those dots differentiate between the otherwise similar characters. Additionally, three characters can have a zigzag like stroke called (Hamza), shown in Figure 1(b). The dots and Hamza are called secondaries. They may be located above the character's primary part, as in ALEF (Figure 1(c)), or below its primary part, like in BAA (Figure 1(d)). In only one letter JEEM, the dot is located in the middle of its primary part, as shown in Figure 1(e). Written Arabic text is cursive in both machine-printed and hand-written text. Within a word, some characters connect to the preceding and/or following characters, and some do not connect. The shape of an Arabic character depends on its position in the word; a character might have up to four different shapes depending on it being isolated, connected from the right (beginning form), connected from the left (ending form), or connected from both sides (middle form).

Figure 2(a) shows the various forms of Arabic character (BAA) that appears within text. Figures 2(b), (c) and (d) show the use of diacritics, Hamza and dots, respectively. Arabic characters do not have fixed size (height and width). The character size varies according to its position in the word. Letters in a word can have short vowels (diacritics). These diacritics are written as strokes, placed either on top of the letters or below them.

ب بـبـب بـبـبـب ـَـُـِـْـّـٌـٍ أإؤئئ بـنـتـيـث

(a) (b) (c) (d)

Fig. 2. Examples of (a) Different shapes of the same character, (b) Diacritics, (c) Characters with Hamzah, and (d) Different number and locations of dots.

Besides Arabic letters, Indian numerals are commonly used in Arabic scripts, although Arabic numerals have been increasingly used in recent writings. Arabic numerals are used in Latin scripts. As in Latin, the digits in Arabic numbers are written with the most significant digit to the left. This paper is organized

as follows: Section 2 presents the published work related to developing Arabic off-line handwriting databases. Data collection phase is presented in Section 3. Then, data extraction and pre-processing phases are described in Section 4. In Section 5, the database structure along with ground-truth data is presented. Finally, we present the conclusions and future work.

2 Related Works

There has not been much effort towards developing comprehensive databases for Arabic handwriting text recognition as compared to that for other languages, for example Latin-scripts [4,5,6,7,8]. Among the earliest works in this regard, Abuhaiba et al. [9] developed a database of around 2000 samples of unconstrained Arabic handwritten characters written by four writers. The database contains the basic character shapes without dots, comprising a total of 51 shapes. This database lacks some handwritten character shapes. Al ISRA database [10] was collected by a group of researchers at the University of British Columbia in 1999. It contains 37,000 Arabic words, 10,000 digits, 2,500 signatures, and 500 free form Arabic sentences gathered from five hundred randomly selected students at Al Isra University in Amman, Jordan.

IFN/ENIT database [11,12] was developed in 2002 by the Institute of Communications Technology (IFN) at Technical University Braunschweing in Germany and The National School of Engineers of Tunis (ENIT). Version one consisted of 26,549 images of Tunisian town/village names written by 411 writers. The images are partitioned into four sets so that researchers can use and discuss training and testing data in this context. It is one of the most widely used databases. However, it lacks the naturalness of handwritten Arabic text as it essentially contains names of towns and villages of Tunisia. Khedher et al. [13] used a database of unconstrained isolated Arabic handwritten characters written by 48 writers, making it suitable for isolated character recognition research. AHDB database [14] was developed in 2002 by Almaadeed. This database includes words that are used in writing legal amounts on Arabic checks and some free handwriting pages from 100 writers.

Al-Ohali et al. [15], of the Center for Pattern Recognition and Machine Intelligence (CENPARMI), developed an Arabic check database in 2003 for research in the recognition of Arabic handwritten bank checks. The database includes Arabic legal and courtesy amounts that were extracted from 3000 bank checks of AlRajhi Bank, Saudi Arabia. The database contains 2499 legal amounts, 2499 courtesy amounts written in Indian/Arabic digits, 29498 Arabic subwords used in legal amounts and 15175 Indian/Arabic digits extracted from courtesy amounts. This database can be mainly used for Arabic handwritten number, digits and limited vocabulary word recognition.

ADBase database was presented by ElSherif et al. [16]. It consists of 70,000 digits written by 700 writers each writing every digit ten times. The database is partitioned into two sets: a training set consisting of 60,000 digits samples and a test set of 10,000 digit samples. This database can be used for research

in Arabic handwritten digits recognition. A database containing Arabic dates, isolated digits, numerical strings, letters, words and some special symbols were presented by AlAmri [17]. A database of Arabic (Indian) digits was presented by Mahmoud [18]. 44 writers each wrote 48 samples of each digit. This database is suitable for research in isolated digit recognition. The database is limited in size and naturalness.

3 Data Collection

In order to build a database for Arabic Handwritten text recognition, a data collection form is designed. Each form consists of four pages. The first page includes fields for writer information; the name, age category, upbringing country, qualification, gender, left/right-handedness, and a section for management purposes. Figure 3(a) shows the first page of a filled form. The writer name is masked for privacy reasons. The remaining three pages contain six paragraphs, two in each page. The second page, shown in Figure 3(b), consists of two paragraphs, the first paragraph is a summarized paragraph that covers all Arabic characters and forms (beginning, middle, end, and isolated). The second paragraph contains text that is randomly selected from an Arabic corpus that we collected from different published work in different topics. The third page has two paragraphs (viz. the third and fourth paragraphs shown in Figure 3(c)). The third paragraph is a randomly selected paragraph (similar to the second one) and the fourth is a repetition of the first paragraph of the second page. The second and third paragraphs are distinct over all the forms and the first and fourth paragraphs are repeated in all the forms. This was done to enable the database to be used by researchers for writer identification and verification applications in addition to Arabic handwritten text research. Paragraphs 5 and 6 are free form paragraphs that are kept in the fourth page. The writer writes two paragraphs on any topic he/she likes. The sixth paragraph has writing lines to enable researchers to analyze both lined and unlined text. Figure 3(d) shows a sample of a filled fourth page of the form. The forms are distributed in several countries and entities for data collection; mainly from native Arabic speakers. Currently, most of the participants are from Saudi Arabia.

4 Data Extraction and Pre-processing

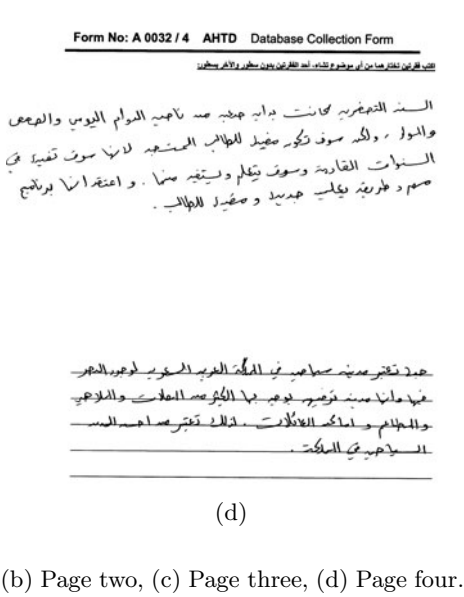
Once the forms are filled by writers and collected, they are scanned at 200 dpi, 300 dpi, and 600 dpi gray scale resolution. Tools are developed to binarize, skew correct, and de-noise the forms. Further the tool extracts the paragraphs from these forms and save them separately for researchers that need to address full paragraphs as is the case with segmentation and writer identification and verification.

The ground truth of these paragraphs is also stored. This enables researchers to compare their recognition results with the true text. Then these paragraphs are segmented into lines. These lines are separately stored for researchers who

(a)

(b)

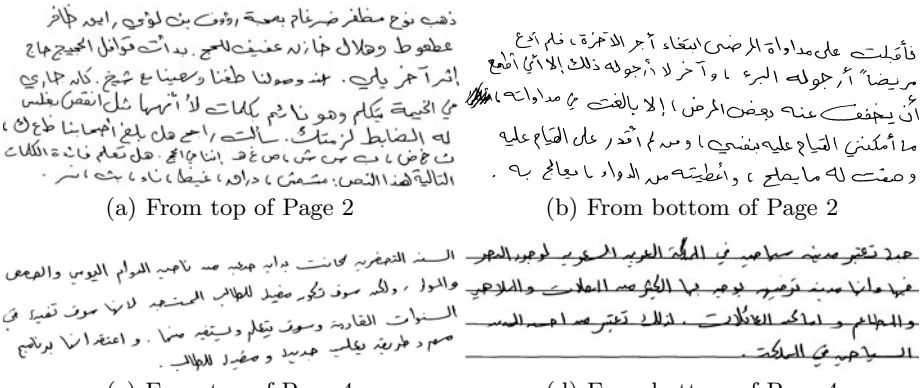
(c)



(d)

Fig. 3. A sample filled form. (a) Page one, (b) Page two, (c) Page three, (d) Page four.

need to conduct research on lines of text. For example, text recognition using HMM, segmentation of words, sub-words, characters, etc. The truth value at the line level is also stored. This can be used by researchers to generate labeled training data, in the training phase and compare with recognized text in the testing phase. These tools have the option to accept user interaction so that the ground truth may be corrected both at the paragraph and line levels. Figure 4 shows samples of extracted paragraphs from a sample form. Figure 5 shows one of the paragraphs along with the segmented lines. Figure 6 shows the segmented lines along with its associated ground truth. The tools work both in an interactive mode (enabling user intervention to correct and/or improve the outputs manually at each stage) and in batch mode (where multiple forms are processed automatically without user intervention).



(a) From top of Page 2 (b) From bottom of Page 2

(c) From top of Page 4 (d) From bottom of Page 4

Fig. 4. Samples of extracted paragraphs from the form

5 AHTD Database Overview

The AHTD database is composed of two databases: the images database and the ground truth database.

5.1 The Images Database

This database consists of gray scale images scanned at 200, 300 and 600 dpi. Thus at the highest level, the database will have three series. Each series consists of gray scale images of the form for a certain resolution i.e. 200, 300 and 600 dpi respectively. Figure 7 shows the database structure. Apart from the original scanned forms, each series is divided into datasets. There are basically three datasets (viz. datasets of pages, paragraphs and lines). The datasets will be extended to include a dataset of words. Each of these sets has associated

Fig. 5. Sample of a paragraph and its corresponding lines after segmentation

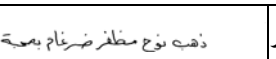
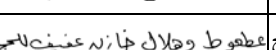
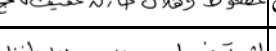
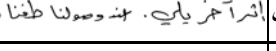
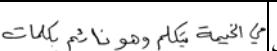
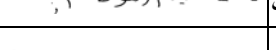
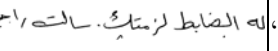
	ذهب نوح مظفرضرغام بصحبة رؤوف بن لؤي رايق ظافر
	عطعوط وهلال خازن عفيف للحج . بدأت قوافل الحجيج حاج
	إثر آخر يلبي . عند وصو لنا طفنا و سعينا مع شيخ. كان جاري
	في الخيمة يتكلم وهو نائم بكلمات لاأفهمها مثل انقض بغلس
	له الضابط لزمتك . سألت راجح هل بلغ أصحابنا ظ ع ك،
	ث خ ض،ب س ش، ص غ هـ أننا في الحج. هل تعلم فائدة الكلمات
	التالية لهذا النص: مشمش ، دراق ، غيظ ، ناء ، بث ، نسر.

Fig. 6. Segmented Lines along with ground truths

ground truth at the page, the paragraph and the line levels. At the page level, there are 3000 pages consisting of a total of 6000 paragraphs. Out of these 6000 paragraphs, 4000 paragraphs are written based on the printed samples (i.e. pages two and three) whereas the remaining 2000 paragraphs are freely written by the writers. The 4000 paragraphs include 2000 paragraphs with same truth value.

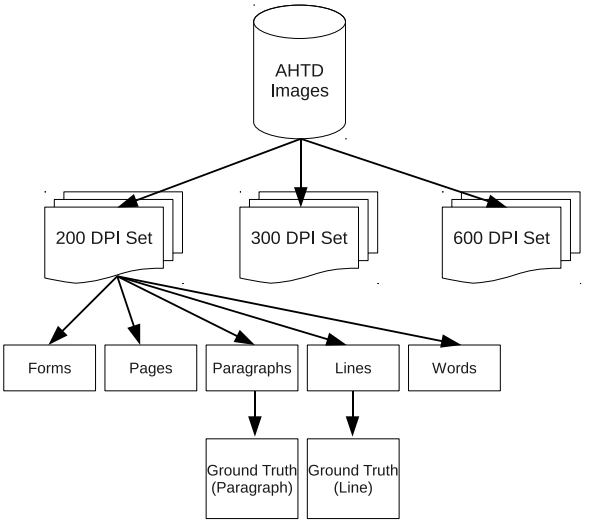


Fig. 7. The database structure

5.2 Ground Truth Database

The ground truth of the data is available at the form, at the paragraph and at the line levels. In addition, the database contains information at the form level related to the writer of the form. The information includes the writer ID, writer age, country of upbringing, educational qualification, gender and handedness (left or right handed).

6 Conclusions and Future Works

A freely available database for Arabic language text recognition that represents variety of handwriting styles is currently lacking and is crucial for Arabic handwritten text recognition and writer identification research. In this paper, we are presenting an Arabic handwritten text recognition database. To build a comprehensive database we designed a form consisting of four pages having a total of six paragraphs. Out of these six paragraphs, one paragraph is repeated twice in each form. Two other paragraphs are unique to each form and the remaining two paragraphs are open for the writers to select a topic of their choice and write. Forms were distributed to the writers in several countries. The collected written forms are scanned and stored at 200, 300 and 600 dpi levels. Each form is further processed by binarizing, correcting skew, extracting paragraphs and segmenting the paragraphs into lines. Moreover ground truth is recorded at the paragraph and line levels for each sample. Additional 1000 forms will be added in the second phase. As an extension to this, words of the segmented text lines will be segmented and stored separately and their ground truth will also be made available to interested researchers. Different experiments in different research areas

related to Arabic handwriting will be conducted on the data. We believe that the database will be of great help and value for the research community. The database may prove useful for various research applications. Arabic handwritten text recognition research at the paragraph and line levels may be conducted. Moreover researchers can propose binarization and noise removal techniques and validate them using gray/binary scale data, propose skew correction techniques at the page level and validate them. Researchers can propose line segmentation techniques and validate them at the paragraph level. Further, researchers can propose slant correction techniques and validate them at the line level. In addition, the database can be used for Arabic handwritten text writer identification and verification research.

Acknowledgement

The authors would like to acknowledge the support provided by King Abdul-Aziz City for Science and technology (KACST) through the Science & Technology Unit at King Fahd University of Petroleum and Minerals (KFUPM) for funding this work through project no. 08-INF99-4 as part of the National Science, Technology and Innovation Plan. In addition, we would like to thank all the writers for filling the forms and those who helped in this effort.

References

1. Paul Lewis, M.: Ethnologue: Languages of the World. In: SIL International, 16th edn., Dallas, Texas (2011) `http://www.ethnologue.com/` (last accessed on January 24, 2011)
2. Al-Badr, B., Mahmoud, S.A.: Survey and bibliography of Arabic optical text recognition. Signal Processing 41(1), 49–77 (1995)
3. Al-Muhtaseb, H.A., Mahmoud, S.A., Qahwaji, R.S.: Recognition of off-line printed Arabic text using hidden markov models. Signal Process. 88(12), 2902–2912 (2008)
4. Märgner, V., El Abed, H.: Databases and competitions: Strategies to improve arabic recognition systems. In: Doermann, D., Jaeger, S. (eds.) SACH 2006. LNCS, vol. 4768, pp. 82–103. Springer, Heidelberg (2008)
5. Lorigo, L.M., Govindaraju, V.: Offline Arabic handwriting recognition: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(5), 712–724 (2006)
6. Marti, U.-V., Bunke, H.: A full English sentence database for off-line handwriting recognition. In: Proceedings of the Fifth International Conference on Document Analysis and Recognition, pp. 705–708 (1999)
7. Hull, J.J.: A database for handwritten text recognition research. IEEE Transactions on Pattern Analysis and Machine Intelligence 16(5), 550–554 (1994)
8. Dimauro, G., Impedovo, S., Modugno, R., Pirlo, G.: A new database for research on bank-check processing. In: Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition, pp. 524–528 (2002)
9. Abuhaiba, I.S.I., Mahmoud, S.A., Green, R.J.: Recognition of handwritten cursive Arabic characters. IEEE Transactions on Pattern Analysis and Machine Intelligence 16(6), 664–672 (1994)

10. Kharma, N., Ahmed, M., Ward, R.: A new comprehensive database of handwritten Arabic words, numbers, and signatures used for ocr testing. In: IEEE Canadian Conference on Electrical and Computer Engineering, vol. 2, pp. 766–768 (1999)
11. Pechwitz, M., Snoussi Maddouri, S., Mrgner, V., Ellouze, N., Amiri, H.: IFN/ENIT - database of handwritten Arabic words. In: Proceedings of the 7th Colloque International Francophone sur l'Ecrit et le Document, CIFED (2002)
12. El Abed, H., Mrgner, V.: The IFN/ENIT-database - a tool to develop Arabic handwriting recognition systems. In: IEEE International Symposium on Signal Processing and its Applications, ISSPA (2007)
13. Khedher, M.Z., Abandah, G.: Arabic character recognition using approximate stroke sequence. In: Arabic Language Resources and Evaluation - Status and Prospects Workshop, Third Int'l Conf. on Language Resources and Evaluation, LREC 2002 (2002)
14. Al-Ma'adeed, S., Elliman, D., Higgins, C.A.: A data base for Arabic handwritten text recognition research. In: Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR 2002), pp. 485–489 (2002)
15. Al-Ohali, Y., Cheriet, M., Suen, C.: Databases for recognition of handwritten Arabic cheques. Pattern Recognition 36(1), 111–121 (2003)
16. El-Sherif, E.A., Abdelazeem, S.: A two-stage system for Arabic handwritten digit recognition tested on a new large database. In: International Conference on Artificial Intelligence and Pattern Recognition, pp. 237–242 (2007)
17. Alamri, H., Sadri, J., Suen, C.Y., Nobile, N.: A novel comprehensive database for Arabic off-line handwriting recognition. In: Proceedings of the 11 th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 664–669 (2008)
18. Mahmoud, S.: Recognition of writer-independent off-line handwritten Arabic (Indian) numerals using hidden markov models. Signal Processing 88(4), 844–857 (2008)