A Database for Arabic Handwritten Text Image Recognition and Writer Identification

Anis Mezghani, Slim Kanoun, Maher Khemakhem

University of Sfax
MIRACL Lab, ISIMS
Sfax, Tunisia
{anis.mezghani, slim.kanoun}@gmail.com,
maher.khemakhem@fsegs.rnu.tn

Haikal El Abed

Braunschweig Technical University
Institute for Communications Technology (IfN)
Braunschweig, Germany
elabed@tu-bs.de

***Abstract*—Standard databases play essential roles for evaluating and comparing results obtained by different groups of researchers. In this paper, an Arabic Handwritten Text Images Database written by Multiple Writers (AHTID/MW) is introduced. This database can be used for research in the recognition of Arabic handwritten text with open vocabulary, word segmentation and writer identification. The AHTID/MW contains 3710 text lines and 22896 words written by 53 native writers of Arabic. In addition, ground truth annotation is provided for each text image. The database is freely available for worldwide researchers.**

Keywords-Arabic Handwritten text image; AHTID/MW Database; open vocabulary; Ground truth

I. INTRODUCTION

In the last score of years, most of the efforts in text recognition have been focused on Latin script recognition. This is due to the availability of several databases of printed and handwritten Latin text [1, 2]. Similar databases also exist for a few other languages such as Chinese and Indian [3, 4, 5].

The existing research on recognizing Arab text is still limited. The lack of freely available Arabic databases is considered as one of the reasons for the lack of research on Arabic text recognition compared with other languages. Each of research groups implemented their system on set of data gathered by them and different recognition rates were reported. Therefore, the comparison of systems is rather difficult.

In the field of Arabic handwritten text recognition, having a standard database is crucial for text image recognition and writer identification. Researchers have prepared some databases for handwritten texts [6], handwritten words [7] and bank checks [8]. To the best of our knowledge, only one database is available for Arabic handwritten text-lines with open vocabulary [9]. However, it doesn't include a dataset of words and it seems difficult to have access to this database.

Considering these issues, we propose to develop an Arabic Handwritten Text Images Database (AHTID/MW) that covers all Arabic characters and forms (beginning, middle, end, and isolated). The proposed database contains Arabic words and text-lines written by 53 different writers. Two types of ground truths based on content information (text-line image and word image) are generated. This database will be made available for the scientific community and may be used as a benchmark database where researchers can evaluate and compare their algorithms and results with other published works.

In Section 2, we will outline the published works related to developing off-line Arabic handwritten databases. In Section 3, we briefly illustrate characteristics of Arabic scripts. Section 4 describes the proposed AHTID/MW database. In Section 5, the database structure along with ground truth data is presented. Finally, concluding remarks are given in Section 6.

II. CURRENT ARABIC HANDWRITTEN DATABASES

The last years showed an increasing interest in Arabic handwritten text recognition solutions. Starting with small private data sets to evaluate their systems, more and more researchers begin to bring more attention to the dataset standardization. For example, a database of unconstrained isolated Arabic handwritten characters written by 48 writers was used by Khedher et al. [10] for isolated character recognition research. IFHCDB (Isolated Farsi Handwritten Character Database) database was presented by Mozaffari et al. [11] for the use in optical character recognition research. It consists of 52,380 gray scale images of handwritten characters and 17,740 numerals.

Al-ISRA database [12] was collected in Al-Isra University in Amman, Jordan. It contains 37,000 Arabic words, 10,000 digits, 2,500 signatures, and 500 free-form Arabic sentences gathered from five hundred students. Alma'adeed et al. presented in 2002 the AHDB database [6]. It includes images of words that are used to describe numbers, images of the most frequent words used in Arabic writing, and images of sentences used in writing legal amount on Arabic checks. Another Arabic database for research in recognition of Arabic handwritten checks (CENPARMI) was developed by Al-Ohali et al. [8]. This database can be mainly used for Arabic handwritten number, digits and limited vocabulary word recognition. A database including Arabic dates, isolated digits, numerical strings,

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.155

letters, words and some special symbols was released in 2008 by AlAmri et al. [13].

For Arabic handwritten word recognition, IFN/ENIT database was developed in 2002, by the Institute of Communications Technology (IFN) at Technical University Braunschweig in Germany and the National School of Engineers of Tunis (ENIT). It consists of 26,549 images of the 937 names of cities and towns in Tunisia, written by 411 writers [7]. The images are partitioned into four sets so that researchers can use them for training and testing. Open competitions are even regularly organized using this database [14]. In 2008, a database with the same characteristics of the IFN/ENIT database was created by Mozaffari et al. [15]. The database has been named IfN/Farsi for handwritten Farsi words. It consists of 7,271 binary images of 1,080 Iranian province/city names, collected from 600 writers.

Recently, a comprehensive Arabic Handwritten Text database (AHTD) written by 300 writers has been developed [9]. 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. This database can be used for text recognition, and writer identification.

III. Brief Description of Arabic Script Characteristics

Arabic is used in more than 20 countries by more than 300 million people. Arabic text is inherently cursive both in handwritten and printed forms and is written horizontally from right to left. The Arabic alphabet consists of 28 basic letters. Some of these letters change their shapes according to their position in the word. Several of them have four shapes: isolated, initial, medial and final. An example is shown in Figure 1. Thus, an Arabic word may be decomposed into more than one sub-word called PAW (Piece of Arabic Word), each represents one or more connected letters.

More than half of Arabic letters include in their shape dots which can be one, two or three dots. The presence of these dots in their positions allows us to differentiate between letters that belong to the same family shape. Moreover, Some Arabic letters can be written in different styles, therefore, collecting samples from those different styles is significantly important. As a result, the data entry form includes one sample of the 36 Arabic isolated letters as shown in Table 1.

غ	غـ	ـغـ	ـغ
Isolated form	Beginning form	Middle form	Ending form

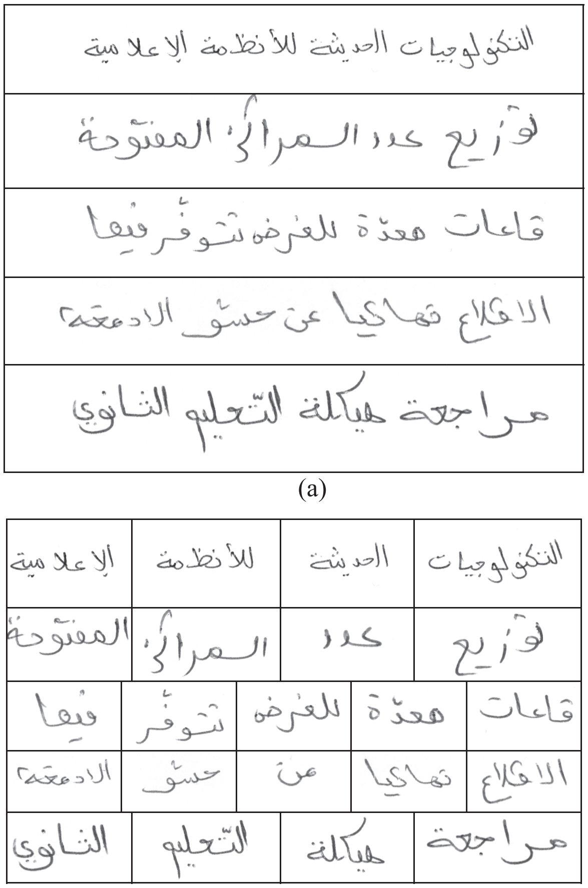
Figure 1. An example of different shapes of an Arabic letter.

IV. Overview of the AHTID/MW

For data collection, the texts are given to 53 individuals with different ages and educational backgrounds and are asked to write text-lines with no restrictions for choosing the type of pen. The handwritten texts are then scanned in gray-scales with the resolution of 300 dpi. The images are stored in "PNG" format. Scanned images of handwritten Arabic text-lines are shown in Figure 2(a).

Since we considered handwritten texts of 53 individuals, 3710 text-line images have been gathered. These images are divided into 4 sets so that researchers can use and discuss training and testing data to evaluate the performance of their approaches. The database was then extended to include a dataset of word images (Figure 2(b)). Each of these sets has associated ground truth at the text-line and the word levels.

AHTID/MW includes 126,511 characters in total, with different forms (beginning, middle, end, and isolated) as shown in Table 1, and these characters lead to 22,896 word images. By simple computation, we get the following statistics: Each text-line has 6.17 words; each word has 5.53 characters and each text-line includes 34.1 characters.


التكنولوجيات الحديثة للأنظمة الإعلامية
توزيع عدد المراكز المفتوحة
قاعات معدة للعرض تتوفر فيها
الاقلاع نهائيا عن حشو الادمغة
مراجعة هيكلة التعليم الثانوي

(a)

(b)

Figure 2. Sample of text-lines and its corresponding words.

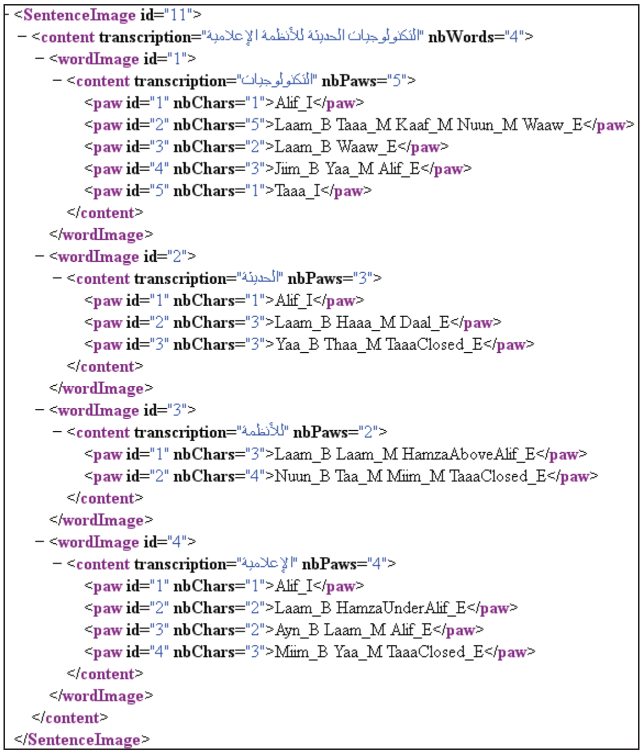
In most of the text-lines of the AHTID/MW database, either an overlapping or touching words exist. The AHTID/MW can be used for Arabic sentence recognition, word recognition, word spotting, and writer identification.

V. GROUND TRUTH DESCRIPTION

In the final structure of our Arabic database, each folder that contains handwritten samples is also provided with the ground truth data file for the samples. This ground truth is useful to evaluate the recognition results. The ground truth of the data is described at the text-line and the word levels using an XML file.

At the text-line level, the ground truth file includes the following information: sequence of words, sequence of PAWs for each word and sequence of characters for each PAW. An example of such XML file is given in Figure 3. Each word image in AHTID/MW database is also described using an XML file containing information about the sequence of PAWs as well as information about the sequence of characters for each PAW.

Inspired by the published work related to developing Arabic Printed Text Image Database [16], we put forward a Latin string for each character. The different character labels can be observed in Table 1 showing statistics. In Arabic handwriting, the shape of each character varies according to its position in the word. To make the label more significant, we added an additional Latin character as a suffix to specify the character position: ’B’ stand for beginning, ’M’ for middle, ’E’ for end and ’I’ for isolated character shapes.



```
- <SentenceImage id="11">
  - <content transcription="التكنولوجيات الحديثة للأنظمة الإعلامية" nbWords="4">
    - <wordImage id="1">
      - <content transcription="التكنولوجيات" nbPaws="5">
          <paw id="1" nbChars="1">Alif_I</paw>
          <paw id="2" nbChars="5">Laam_B Taaa_M Kaaf_M Nuun_M Waaw_E</paw>
          <paw id="3" nbChars="2">Laam_B Waaw_E</paw>
          <paw id="4" nbChars="3">Jiim_B Yaa_M Alif_E</paw>
          <paw id="5" nbChars="1">Taaa_I</paw>
        </content>
      </wordImage>
    - <wordImage id="2">
      - <content transcription="الحديثة" nbPaws="3">
          <paw id="1" nbChars="1">Alif_I</paw>
          <paw id="2" nbChars="3">Laam_B Haaa_M Daal_E</paw>
          <paw id="3" nbChars="3">Yaa_B Thaa_M TaaaClosed_E</paw>
        </content>
      </wordImage>
    - <wordImage id="3">
      - <content transcription="للأنظمة" nbPaws="2">
          <paw id="1" nbChars="3">Laam_B Laam_M HamzaAboveAlif_E</paw>
          <paw id="2" nbChars="4">Nuun_B Taa_M Miim_M TaaaClosed_E</paw>
        </content>
      </wordImage>
    - <wordImage id="4">
      - <content transcription="الإعلامية" nbPaws="4">
          <paw id="1" nbChars="1">Alif_I</paw>
          <paw id="2" nbChars="2">Laam_B HamzaUnderAlif_E</paw>
          <paw id="3" nbChars="2">Ayn_B Laam_M Alif_E</paw>
          <paw id="4" nbChars="3">Miim_B Yaa_M TaaaClosed_E</paw>
        </content>
      </wordImage>
    </content>
</SentenceImage>
```

Figure 3. An example of a ground truth data files at the text-line level.

TABLE I. ARABIC LETTERS WITH USED LABELS AND OCCURRENCE IN AHTID/MW DATABASE

Character label	Isolate	Begin	Middle	End
Alif	11395(ا)		9328(ـا)	
Baa	159(ب)	2385(بـ)	1696(ـبـ)	265(ـب)
Taaa	1378(ت)	2120(تـ)	4293(ـتـ)	212(ـت)
Thaa	53(ث)	212(ثـ)	371(ـثـ)	53(ـث)
Jiim	53(ج)	954(جـ)	1007(ـجـ)	106(ـج)
Haaa	53(ح)	742(حـ)	1325(ـحـ)	159(ـح)
Xaa	53(خ)	583(خـ)	265(ـخـ)	53(ـخ)
Daal	1113(د)		2438(ـد)	
Thaal	265(ذ)		636(ـذ)	
Raa	2067(ر)		3816(ـر)	
Zaay	265(ز)		424(ـز)	
Siin	106(س)	1113(سـ)	1325(ـسـ)	212(ـس)
Shiin	53(ش)	530(شـ)	424(ـشـ)	53(ـش)
Saad	53(ص)	583(صـ)	636(ـصـ)	159(ـص)
Daad	212(ض)	106(ضـ)	318(ـضـ)	53(ـض)
Thaaa	53(ط)	424(طـ)	636(ـطـ)	106(ـط)
Taa	53(ظ)	265(ظـ)	530(ـظـ)	53(ـظ)
Ayn	318(ع)	1696(عـ)	2120(ـعـ)	212(ـع)
Ghayn	106(غ)	212(غـ)	424(ـغـ)	53(ـغ)
Faa	159(ف)	1060(فـ)	1007(ـفـ)	53(ـف)
Gaaf	265(ق)	689(قـ)	1431(ـقـ)	371(ـق)
Kaaf	53(ك)	795(كـ)	1007(ـكـ)	212(ـك)
Laam	530(ل)	11607(لـ)	2544(ـلـ)	636(ـل)
Miim	318(م)	3021(مـ)	3816(ـمـ)	901(ـم)
Nuun	689(ن)	1802(نـ)	1802(ـنـ)	848(ـن)
Haa	159(ه)	636(هـ)	1060(ـهـ)	1113(ـه)
Waaw	1060(و)		4081(ـو)	
Yaa	265(ي)	2809(يـ)	4664(ـيـ)	1060(ـي)
Hamza	212(ء)			
HamzaAboveAlif	1007(أ)		1113(ـأ)	
HamzaUnderAlif	795(إ)		689(ـإ)	
TildAboveAlif	212(آ)		53(ـآ)	
TaaaClosed	1060(ة)			4240(ـة)
AlifBroken	106(ى)			318(ـى)
HamzaAboveAlif Broken	212(ئ)	53(ئـ)	159(ـئـ)	53(ـئ)
HamzaAboveWaaw	53(ؤ)		318(ـؤ)	

VI. CONCLUSIONS

With the increasing interest in Arabic handwriting recognition, the need for a freely standard Arabic handwriting database that represents variety of handwriting styles is highly required. In this paper, we presented an Arabic Handwritten Text Images Database with open vocabulary naturally written by Multiple Writers. The AHTID/MW database contains 3710 text lines and 22896 word images. For each piece of the database, a corresponding ground truth file is available. The database may prove useful for various research applications such as Arabic handwritten text recognition and writer identification systems. We would note that the AHTID/MW database will be made freely available to interested researchers. We believe that the database will be of great help and value for the research community.

ACKNOWLEDGMENT

The authors would like to thank all writers who contributed to AHTID/MW database generation.

REFERENCES

[1] J. J. Hull, “A database for Handwritten Text Recognition Research,” IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 16, pp. 550–554, 1994.

[2] Y. LeCun, L. Bottou, Y. Bengio and P. Haffiner, “Gradient based learning applied to document recognition,” Proceedings of IEEE, vol. 86(11), pp. 2278–2324, 1998.

[3] T. Saito, H. Yamada and K. Yamomoto, “One database ELT9 of handprinted characters in JIS Chinese characters and its analysis (in Japanese),” Transaction of IECEJ, vol. J.68-D(4), pp. 757–764, 1985.

[4] U. Bhattacharya and B. B. Chaudhuri, “Databases for Research on Recognition of Handwritten Characters of Indian Scripts,” International Conference of Document Analysis and Recognition, pp. 789–793, 2005.

[5] S. Mihov, K. U. Schulz, C. Ringlsteller, V. Dojchinova, V. Nakaova, K. Kalpakchieva, O. Gerasimov, A. Gotsharek and C. Gercke, “A Corpus for Comparative Evaluation of OCR Software and Postcorrection Techniques,” Proceeding of International Conference of Document Analysis and Recognition, pp. 162–166, 2005.

[6] S. Al-Ma’adeed, D. Ellimam and C. A. Higgins, “A data Base for Arabic Handwritten Text Recognition Research,” Proceeding of Eighth International Workshop on Frontiers in Handwriting Recognition, pp. 485–489, 2002.

[7] M. Pechwitz, S. S. Maddouri, V. Maergner, N. Ellouze and H. Amiri, “IFN/ENIT - database of handwritten Arabic words,” Proceeding of Colloque International Francophone sur l’Écrit et le Document, pp. 129–136, 2002.

[8] Y. Al-Ohali, M. Cheriet and C. Suen, “Databases for recognition of handwritten Arabic checks,” Pattern Recognition, vol. 36, pp. 111–121, 2003.

[9] S. A. Mahmoud, I. Ahmad, M. Alshayeb and W. G. Al-Khatib, “A Database for Offline Arabic Handwritten Text Recognition,” International Conference on Image Analysis and Recognition, part. II, pp. 397–406, 2011.

[10] M. Z. Khedher and G. Abandah, “Arabic character recognition using approximate stroke sequence,” Arabic Language Resources and Evaluation - Status and Prospects Workshop, Third International Conference on Language Resources and Evaluation, 2002.

[11] S. Mozaffari, K. Faez, F. Faradji, M. Ziaratban and S. M. Golzan, “A Comprehensive Isolated Farsi/Arabic Character Database for Handwritten OCR Research,” International Workshop on Frontiers in Handwriting Recognition, 2006.

[12] N. Kharma, M. Ahmed and R. Ward, “A New Comperehensive Database of Hand-written Arabic Words, Numbers, and Signatures used for OCR Testing,” IEEE Canadian Conference on Electrical and Computer Engineering, vol. 2, pp. 766–768, 1999.

[13] H. Alamri, J. Sadri, C. Y. Suen and N. Nobile, “A novel comprehensive database for Arabic off-line handwriting recognition,” Proceedings of the 11 th International Conference on Frontiers in Handwriting Recognition, pp. 664–669, 2008.

[14] H. El Abed and V. Märgner, “Base de Données et Compétitions - Outils de Développement et d’Évaluation de Systèmes de Reconnaissance de Mots Manuscrits Arabes,” Actes du dixième Colloque International Francophone sur l’Écrit et le Document, 2008.

[15] S. Mozaffari, H. El Abed, V. Märgner, K. Faez and A. Amirshahi, “IfN/Farsi-Database: A Database of Farsi Handwritten City Names,” Proceedings of the 11 th International Conference on Frontiers in Handwriting Recognition, 2008.

[16] F. Slimane. R. Ingold, S. Kanoun, A. M. Alimi and J. Hennebert, “A New Arabic Printed Text Image Database and Evaluation Protocols,” International Conference on Document Analysis and Recognition, pp. 946–950, 2009.