Database for Arabic Printed Text Recognition Research

Faten Kallel Jaiem[1], Slim Kanoun[1], Maher Khemakhem[1],
Haikal El Abed[2], and Jihain Kardoun[3]

[1] MIRACL laboratory, ISIMS, University of Sfax, Tunisia
{kallelfaten,slim.kanoun}@gmail.com,
maher.khemakhem@fsegs.rnu.tn
[2] Institute for Communications Technology, Braunschweig University, Germany
elabed@tu-bs.de
[3] Department of Computer Engineering, ENIS, University of Sfax, Tunisia
jihen.kardoun@gmail.com

Abstract. This paper presents a real database for the Arabic printed text recognition, APTID / MF (Arabic Printed Text Image Database / Multi-Font).This database can be used to evaluate the system that recognizes Arabic printed texts with an open vocabulary. APTID / MF may be also used for research in word segmentation and font identification. APTID / MF is obtained from 387 pages of Arabic printed documents scanned with grayscale format and 300 dpi resolutions. From this documents, 1,845 text-blocks have been extracted. In addition ground truth file is provided for each texts-block. APTID / MF also includes an Arabic printed character image dataset made up of 27,402 samples. The database is freely available to interested researchers.

Keywords: Arabic printed text, APTID / MF database, Open vocabulary, Ground truth.

1 Introduction

A deep observation of the printed Arabic documents reveals that the number of font styles and sizes used is important. So far, there has been no standard corpus which includes multi-font, multi-style and multi-size printed Arabic writing that can be used to evaluate the score (recognition rate) of any given Arabic OCR system.

Indeed, intensive experiments achieved on some existing Arabic OCR systems reveal that their recognition rates is sometimes very sensitive to the variability of the font style and/or the font size.

Consequently, this paper presented a detailed description of Arabic multi-font, multi-style and multi-size printed text database. The first section presents the Arabic script complexity and the diversity of the fonts and styles of the Arabic printed script. The second section sheds some light on the existing databases for the printed Arabic script recognition. The third section gives a detailed description of the published databases. In the last section of this paper, we talk about the various prospects and future work for the database extension.

A. Petrosino (Ed.): ICIAP 2013, Part I, LNCS 8156, pp. 251–259, 2013.

2 Brief Description of Arabic Script Characteristics

The Arabic script has certain characteristics which cause the complexity of its recognition. These characteristics come from its cursive nature and the variation of its character forms according to their position in the word [1]. The Arabic characters change forms not only according to their position in the word but also according to the used calligraphic style. Calligraphy is very developed in the Arab-Muslim world. In fact, the Arabic writing can appear in different calligraphic styles such as Neskhi, Thoulthi, Diwani [2]. The analysis of the Arabic documents shows the existence of a diversity of writing fonts for the printed paper resulting from the Arabic calligraphic styles.

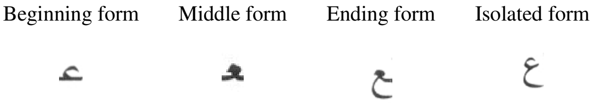


Fig. 1. An example of different shapes of an Arabic letter

3 Current Arabic Printed Database

In literature, the database for text recognition research is often organized in two classes. The first class represents the database of the printed documents; the second includes the handwritten documents [3] [4] [5]. This section presents an overview of the current Arabic Printed Database.

3.1 APTI

The APTI database [6] is made up of 45,313,600 Arabic Printed word images. These word images cover approximately 250,000,000 characters. These images result from 113,284 various word writings with 10 fonts, 10 sizes and 4 styles. The Xml files associated with each element of the database present the ground truth data.

The APTI database is a synthetic and it is designed for the evaluation of screen-based OCR systems . The images are generated at 72 dpi by an automatic program. This presents a disadvantage when evaluating the systems for recognizing scanned Arabic printed documents.

3.2 DARPA

The DARPA (Defense Advanced Research Projects Agency) Arabic corpus was created by Scientific Application International Company for the US Department of Defense [7]. DARPA Corpus data were collected from the books, the Magazines, the newspapers and the computer generated documents covering only 4 fonts. The DARPA corpus includes 345 Arabic printed pages with ground truth data. These pages were scanned with a 600 dpi resolution.

3.3 PATDB

The PATDB (Printed Arabic Text DataBase) was published by Al Hashim in 2010 [8]. This database was issued from the printed Arabic pages of texts, resulting from books, advertisements, magazines, newspapers and reports. These pages were scanned with 3 resolution levels 200, 300 and 600 dpi. The database is made up of 6,954 pages of Arabic printed texts. Ground truth data files were attributed to each page of the database.

4 Overview of the APTID / MF

In this section we present our database of Arabic printed text called APTID / MF (Arabic Printed Text Image Database / Multi-Font). In our work we initially aimed to create an Arabic multi-font and multi-size printed text database. The APTID / MF included an Arabic printed text image dataset, and an Arabic printed character image dataset.

4.1 Arabic Printed Text Image Dataset

The Arabic printed text image dataset was selected from the official site of the tunisien newspaper “El-chourouk”. The set of document pages, included in the APTID / MF, is written in two writing styles (normal and bold).

The document pages are organized in 10 sets and we attributed a writing font for each set, The figure below present this fonts. The set of documents is written with 4 sizes (12, 14, 16, and 18 pts).

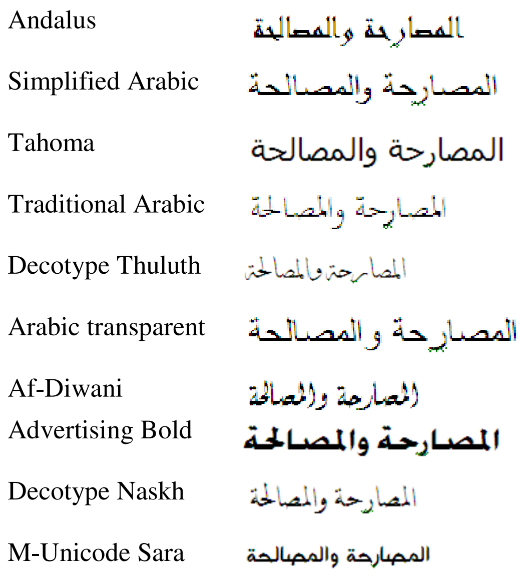
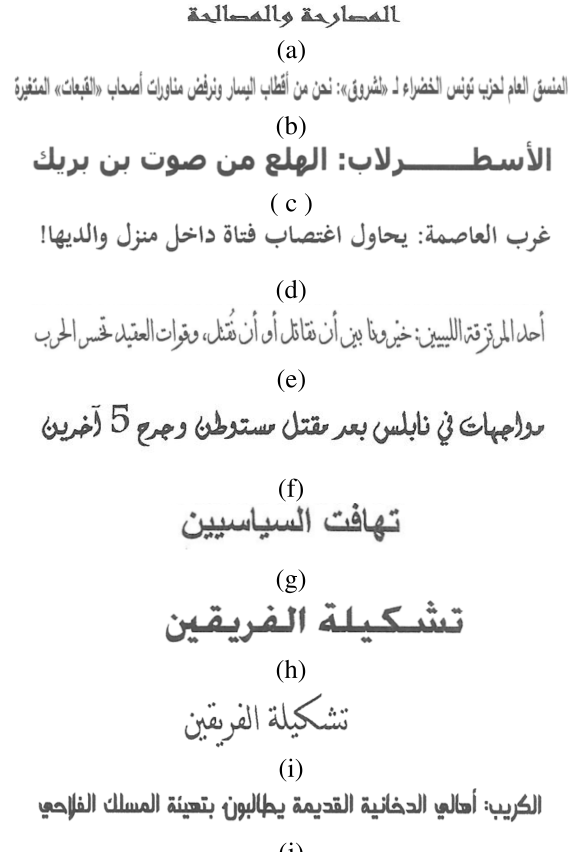

Andalus المصارحة والمصالحة
Simplified Arabic المصارحة والمصالحة
Tahoma المصارحة والمصالحة
Traditional Arabic المصارحة والمصالحة
Decotype Thuluth المصارحة والمصالحة
Arabic transparent المصارحة و المصالحة
Af-Diwani المصارحة والمصالحة
Advertising Bold المصارحة والمصالحة
Decotype Naskh المصارحة والمصالحة
M-Unicode Sara المصارحة والمصالحة

Fig. 2. An exemple of Arabic text written with the used fonts

Table 1. The distribution of document page

	Size 12	*Size 14*	*Size 16*	*Size 18*
The number of document pages	40	54	45	53

This set of document pages are printed with a laser printer and an inkjet printer. The set is stored in tow groups: the first includes the pages printed using a laser printer and the second includes those printed using an inkjet printer. Then, the pages are scanned. The first set of printed pages is scanned with an HP scanner while, the pages obtained by the second printer are scanned with an Epson one. All the page images are scanned with 300 dpi resolution in Grayscale format. The images are stored in "PNG" format. The APTID / MF contains 386 page images of Arabic printed texts. These page images are divided into text-blocks with a manual segmentation. This phase gave us 1,845 Arabic printed image text-blocks made up of 126,792 Arabic words, resulting from 6,989 distinct words.



المصارحة والمصالحة

(a)

المنسق العام لحزب تونس الخضراء لـ «الشروق»: نحن من أقطاب اليسار ونرفض مناورات أصحاب «القبعات» المتغيرة

(b)

الأسطـــــــرلاب: الهلع من صوت بن بريك

(c)

غرب العاصمة: يحاول اغتصاب فتاة داخل منزل والديها!

(d)

أحد المرتزقة الليبيين: خيرونا بين أن نقاتل أو أن نُقتل، وقوات العقيد تخسر الحرب

(e)

مواجهات في نابلس بعد مقتل مستوطن وجرح 5 آخرين

(f)

تهافت السياسيين

(g)

تشكيلة الفريقين

(h)

تشكيلة الفريقين

(i)

الكريب: أهالي الدخانية القديمة يطالبون بتهيئة المسلك الفلاحي

(j)

Fig. 3. Examples of image texts-blocks : (a)Andalus, (b) Simplified Arabic, (c) Tahoma, (d) Traditional Arabic, (e) Decotype Thuluth, (f) Af-Diwani, (g) Arabic transparent, (h)Advertising Bold, (i) Decotype Naskh and (j) M Unicode Sara

Statistics. The database APTID / MF included 1845 text-blocks images. The different text-block images in the APTID / MF are stored in different sets. TABLE 2 presents the distribution of APTID / MF text-blocks.

Table 2. The distribution of APTID / MF text-blocks

A laser printer and an HP scanner					
Font	***size 12***	***size 14***	***size 16***	***size 18***	***Total***
Andalus	26	26	24	21	97
ArabicTransparent	19	19	17	21	76
AdvertisingBold	24	24	23	26	97
Diwani Letter	21	21	20	21	83
DecoTypeThuluth	24	23	23	26	96
Simplified Arabic	20	20	20	20	80
Tahoma	19	19	19	21	78
Traditional Arabic	27	27	27	26	107
DecoType Naskh	20	20	21	17	78
M Unicode Sara	30	30	30	30	**120**
An inkjet printer and an Epson scanner					
Font	***size 12***	***size 14***	***size 16***	***size 18***	***Total***
Andalus	26	26	26	29	107
ArabicTransparent	19	19	17	25	80
AdvertisingBold	24	24	23	28	99
Diwani Letter	21	21	20	23	85
DecoTypeThuluth	24	23	23	28	98
Simplified Arabic	20	20	20	19	79
Tahoma	19	19	19	19	76
Traditional Arabic	27	27	27	26	101
DecoType Naskh	20	20	21	20	81
M Unicode Sara	30	30	30	31	121
TOTAL	**460**	**458**	**450**	**477**	**1845**

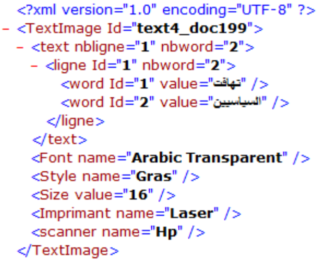
Ground Truth File Description. The database APTID / MF included 1,845 text-block images with their associated metadata files (XML file). These files present the ground-truth value of each sample of text image dataset, These files described at the text-block and line levels using XML file.

At the text-block level, these XML files include the following information: the text-block name (<TextImage Id = …>) and the number of lines and words in text-block (<text nbligne=… nbword = …>)

At the line level, these XML files include the following information: the row of line and the number of words on the line (<ligne Id= … nbword=…> and the row and electronics of the word in the line (<word Id =… value=... />).

In addition these Xml files present the font (<Font name=… />), the style (<Style name=… />) and the size (<Size value=… />) of the text-block, the type of printer used(<Imprimant name=… />) and the name of the scanner used (<scanner name=…/>).

In the final structure of our text dataset, each folder that contains printed text-block samples is also provided with the ground truth data file for the sample. This ground truth is useful to evaluate the recognition results.



```
  <?xml version="1.0" encoding="UTF-8" ?>
- <TextImage Id="text4_doc199">
  - <text nbligne="1" nbword="2">
    - <ligne Id="1" nbword="2">
        <word Id="1" value="تهافت" />
        <word Id="2" value="السياسيين" />
      </ligne>
    </text>
    <Font name="Arabic Transparent" />
    <Style name="Gras" />
    <Size value="16" />
    <Imprimant name="Laser" />
    <scanner name="Hp" />
  </TextImage>
```

Fig. 4. Example of XML files including ground truth

4.2 Arabic Printed Character Image dataset

The Arabic printed character image dataset was selected from various printing forms, magazines, Book chapters and newspapers… The construction process starts by scanning all the pages with a 300 dpi resolution in grayscale format scanner. After that, a binary copy of these pages is segmented in character images. The character dataset contains 27,402 binary images of characters. The different forms of the characters are used to store the character images in 32 classes.

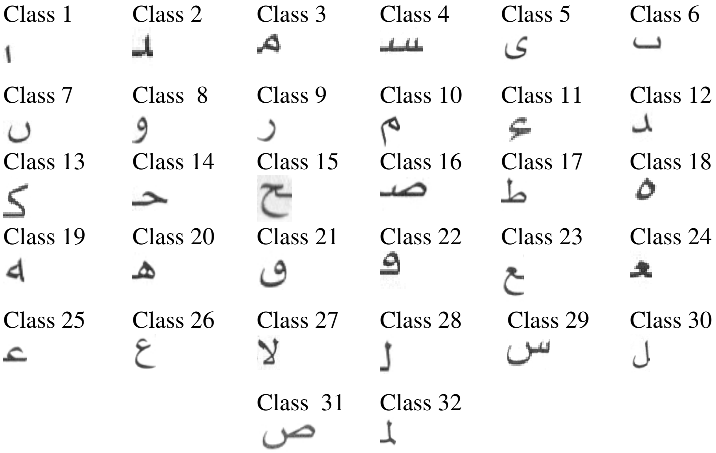


Fig. 5. Character class

The dataset of character images is divided into a training set that contains 18,404 samples and a test set which contains 8,998 samples.

Table 3. The distribution characters of training set

Class	Number image	Class	Number image
1	733	17	543
2	752	18	523
3	650	19	615
4	604	20	564
5	578	21	483
6	609	22	695
7	572	23	488
8	622	24	618
9	640	25	536
10	570	26	449
11	510	27	590
12	625	28	643
13	607	29	497
14	636	30	459
15	537	31	354
16	556	32	546

Table 4. The character distribution of the test set

Class	Number image	Class	Number image
1	361	17	267
2	371	18	258
3	316	19	301
4	296	20	277
5	282	21	238
6	295	22	339
7	279	23	242
8	301	24	304
9	311	25	263
10	278	26	218
11	248	27	285
12	306	28	313
13	297	29	245
14	310	30	223
15	263	31	171
16	273	32	267

Recognition Results. A character recognition system is developed . In our work we used the Hu's invariant moments [9], Affine invariant moments [10], Zernike's moments [11], Tsirikolias-Mertzios Moments [12], Fourier Mellin Transform [13], and Fourier Descriptor [14],which all represent the statistical features. In addition we referred to the Freemanchain codes [15] for the structural feature. Finally we chose the works of Heutte [16] for the topological feature to analyze the statistical and structural features existing in the literature. In this paper , ower aim is to present an Arabic characters dataset. So, as a first experiment of this dataset, we have chosen the K-nearest neighbor as classifier with K=1 and the training set and test set of the character image dataset. The table below presents the results.

Table 4. Recognitions Rate

Features	Recognitions Rate
Affine invariants Moment	65.40%
Hu's invariants Moments	84.07%
Zernike's Moments	76.77%
Tsirikolias–Mertzios Moments	78.04%
Fourier Mellin Transform	76.76%
Fourier Descriptors	67.70%
Topological features	96.81%
Freeman code chain	96.97%

5 Conclusions

In this paper, an Arabic printed text database is presented. This database may be used by the Arabic printed text recognition and font identification research community. The APTID / MF contains 1,845 image text-blocks that are scanned at 300 dpi resolution in grayscale format. The character dataset includes 27,402 image characters. For each piece of the text-block dataset, a corresponding ground truth file is available. APTID / MF was then extended to include a dataset of printed multi-font multi-style and multi-size Arabic image words with large vocabulary resulting from these text-block dataset. The APTID / MF is prepared to organize a competition for the large vocabulary Arabic printed text recognition.

References

1. Amara, N.B.: On the Problematic and Orientations in Recognition of the Arabic Writing. In: CIFED 2002, pp. 1–10 (2002)
2. Kanoun, S., Alimi, A.M., Lecourtier, Y.: Affixal Approach for Arabic Decom-posable Vocabulary Recognition: A Validation on Printed Word in Only One Font. In: ICDAR 2005, pp. 1025–1029 (2005)
3. Pechwitz, M., Maddouri, S., Margner, V., Ellouze, N., Amiri, H.: IFN/ENIT-Database of Handwritten Arabic Words. In: CIFED 2002, pp. 127–136 (2002)
4. Mozaffari, S., Faez, K., Faradji, F., Ziaratban, M., Golzan, M.: Isolated Far-si/Arabic character database for handwritten OCR research. In: International Work-shop on Frontiers of Handwriting Recognition, pp. 385–389 (2006)
5. Mozaffari, S., El Abed, H., Margner, V., Faez, K., Amirshahi, A.: IfN/Farsi-Database: A Database of Farsi Handwritten City Names. ICFHR (2008)
6. Slimane, F., Ingold, R., Kanoun, S., Alimi, A., Hennebert, J.: A New Arabic Printed Text Image Database and Evaluation Protocols. In: proc. of 10th IEEE International Conference on Document Analysis and Recognition, ICDAR 2009, pp. 946–950 (2009)
7. Davidson, R., Hopely, R.: Arabic and Persian OCR Training and Test Data Sets. In: Proceedings of Symposium. On Document Image Understanding Technology (1997)
8. AL-hashim, A.G., Mahmoud, S.A.: Benchmark Database and GUI Environment for Printed Arabic Text Recognition Research. Wseas Transactions Information Science and Applications 7(4), 10 (2010)

9. Hu, M.: Visual pattern recognition by moment invariants. IRE Trans. Information Theory, IT 8, 179–187 (1962)
10. Flusser, J., Suk, T.: Pattern recognition by affine moment invariants. Pattern Recognition 26(1), 167–174 (1993)
11. Zernike, F.: Diffraction theory of the cut procedure and its improved form, the phase contrast method. Physica 1, 689–704 (1934)
12. Tsirikolias, K., Mertzios, B.G.: Statistical pattern recognition using efficient two dimensional moments with applications to character recognition. Pattern Recognition 26, 877–882 (1993)
13. Derrode, S., Ghorbel, F.: Digital Fourier Mellin Transform- Reconstruction and es-timate of objects movement on levels of gray. In: Proc. of GRETSI conference, Grenoble, France, pp. 566–658 (1997)
14. Davis, C.B., Beecher, R., Beecher, M.: The statistical use of Fourier descriptors. Original Research Article Mathematical and Computer Modeling 11, 419–424 (1988)
15. Freeman, H.: On the encoding of arbitrary geometric configurations. IEEE Trans. Electronic Comp. EC-10, 260–268 (1968)
16. Heutte, L.: Reconnaissance de caractères manuscrits: Application a la lecture au-tomatique des chèques et des enveloppes postales. Doctorat Thesis, University of Rouen (1994)