UNIVERSITÉ DE FRIBOURG SUISSE
UNIVERSITÄT FREIBURG SCHWEIZ

Database and Evaluation Protocols for Arabic Printed Text Recognition

Fouad Slimane[1,2] Rolf Ingold[1] Slim Kanoun[3] Adel M. Alimi[2] Jean Hennebert[1,4]

{Fouad.Slimane, Jean.Hennebert, Rolf.Ingold}@unifr.ch,
{ Slim.Kanoun, Adel.Alimi}@enis.rnu.tn

February 6, 2009

DEPARTMENT OF INFORMATICS RESEARCH REPORT

Département d'Informatique – Department für Informatik • Université de Fribourg – Universität Freiburg • Boulevard de Pérolles 90 • 1700 Fribourg • Switzerland

Phone +41 (26) 300 84 65 fax +41 (26) 300 97 26 Diuf-secr-pe@unifr.ch http://diuf.unifr.ch

---

[1] DIVA-DIUF, University of Fribourg, Switzerland
[2] REGIM, University of Sfax, Tunisia
[3] Research Unit of Technologies of Information and Communication (UTIC), Tunis, Tunisia
[4] HES-SO // Wallis, University of Applied Sciences Western Switzerland, Switzerland

Abstract

We report on the creation of a database composed of images of Arabic Printed Text. The purpose of this database is the large-scale benchmarking of open-vocabulary, multi-font, multi-size and multi-style text recognition systems in Arabic. Such systems take as input a text image and compute as output a character string corresponding to the text included in the image. The database is called APTI for Arabic Printed Text Image. The challenges that are addressed by the database are in the variability of the sizes, fonts and style used to generate the images. A focus is also given on low-resolution images where anti-aliasing is generating noise on the characters to recognize. The database is synthetically generated using a lexicon of 113'284 words, 10 Arabic fonts, 10 font sizes and 4 font styles. The database contains 45'313'600 single word images totaling to more than 250 million characters. Ground truth annotation is provided for each image thanks to a XML file. The annotation includes the number of characters, the number of PAWs (Pieces of Arabic Word), the sequence of characters, the size, the style, the font used to generate each image, etc. The database is called APTI for Arabic Printed Text Images.

Keywords: Arabic Text Recognition system, benchmarking, text image databases, OCR

1 Introduction and motivations

With a quite large user base of about 300 million people worldwide, Arabic is important in the culture of many people. In the last fifteen years, most of the efforts in Arabic text recognition have been put for the recognition of scanned off-line printed documents [Khorsheed 07] [Husni 08] [Shaaban 08] [Slimane 08]. Most of these developments have been benchmarked on private databases and therefore, the comparison of systems is rather difficult.

To our knowledge, there are currently few large-scale image databases of Arabic printed text available for the scientific community. One of the only references we have found is about the ERIM database containing 750 scanned pages collected from Arabic books and magazines [Schlosser 95]. However, it seems difficult to have access to this database. In the field of Arabic handwriting recognition, public databases do exist such as the freely available IFN/ENIT-database [Pechwitz 02] Open competitions are even regularly organized using this database [Margner 05] [Margner 07].

On the other hand, text corpus or lexical databases in Arabic are available from different associations or institutes [Graff 06] [Abbes 04] [AbdelRaouf 08]. However, such text corpora are not directly usable for the benchmarking of recognition systems that take images as input.

Considering this, we have initiated the development of a large database of images of printed Arabic words. This database will be used for our own research and will be made available for the scientific community to evaluate their recognition systems. The database has been named APTI for Arabic Printed Text Image.

The purpose of the APTI database is the large-scale benchmarking of open-vocabulary, multi-font, multi-size and multi-style text recognition systems in Arabic. The images in the database are synthetically generated from a large corpus using automated procedures. The challenges that are addressed by the database are in the variability of the sizes, fonts and style used to generate the images. A focus is also given on low-resolution images where anti-aliasing is generating noise on the characters to recognize. By nature, APTI is well suited for the evaluation of screen-based OCR systems that take as input images extracted from screen captures or pdf documents. Performances of classical scanned-based OCR or camera-based

OCR systems could also be measured using APTI. However, such evaluations should take into account the absence of typical artefacts present in scanned or camera documents.

While synthetically generated, the challenges of the database remain multiple:

- Large-scale evaluation with a realistic sampling of most of the Arabic character shapes and their accompanying variations due to ligatures and overlaps;
- Availability of multiple fonts, styles and sizes that must be nowadays treated by recognition systems;
- Emphasis on low resolution images that are nowadays frequently present on computer screens;
- Isolated word images where inter-word language models cannot be used;
- Semi-blind evaluation protocols with decoupled development/evaluation sets.

The objective of this paper is to describe the APTI database and the evaluation protocols defined on the database. In section 2, we present details about lexicon, fonts, font-sizes, rendering procedure, Sources of variability and ground truth description. In section 3, statistical information about the content of the database are given. The evaluation protocols are showed in section 4. Finally, some conclusions are presented in section 5.

2 Specifications of APTI-Database

The APTI database is synthetic and images are generated using automated procedures. In this section, we present the specification of this database.

2.1 Lexicon

The APTI database contains a mix of decomposable and non-decomposable words images. Decomposable words are generated from root Arabic verbs using Arabic schemes [Kanoun 2005] and non-decomposable words are formed by Arabic proper names, general names, country/town/village names, Arabic prepositions, etc.

To generate the lexicon, we have parsed different Arabic books such as *The Muqaddimah - An introduction to history of Ibn Khaldun*[5] *and Al-bukhala of Gahiz*[6] as well as Arabic articles taken from the Internet. This parsing procedure totalled 113'284 single different Arabic words, leading to a pretty good coverage of the Arabic words mostly used in texts. The language used in our sources is exclusively in standard Arabic with no dialect.

2.2 Fonts, styles and sizes

Taking as input the words in the lexicon, the images of APTI are generated using 10 different fonts presented in Fig. 1: Andalus, Arabic Transparent, AdvertisingBold, Diwani Letter, DecoType Thuluth, Simplified Arabic, Tahoma, Traditional Arabic, DecoType Naskh, M Unicode Sara. These fonts have been selected to cover different complexity of shapes of Arabic printed characters, going from simple fonts with no or few overlaps and ligatures (AdvertisingBold) to more complex fonts rich in overlaps, ligatures and flourishes (Diwani Letter or Thuluth).

[5] Ibn Khaldoun, (May 27,1332 – March 19, 1406) was a famous historien, scholar, theologian, and statesman born in North Africa in presentday Tunisia. (http://en.wikipedia.org/wiki/Ibn_Khaldoun)

[6] Al-Jahiz, (born in Basra, c. 781 – December 868 or January 869) was a famous Arab scholar, believed to have been an Afro-Arab of East African descent.(http://en.wikipedia.org/wiki/Al-Jahiz)

Different sizes are also used in APTI: **6** points**, 7** points, **8** points, **9** points, **10** points, **12** points, **14** points, **16** points, **18** points and **24** points. We also used 4 different styles namely plain, italic, bold and combination of italic and bold.

These sizes, fonts and styles are widely used on computer screen, Arabic newspapers, books and many other documents. The combination of fonts, styles and sizes guaranties a wide variability of images in the database.

Overall, the APTI database contains 45’313’600 single words images, taking into account the full lexicon where the different combinations of fonts, style and sizes are applied.

Fig. 1: Fonts used to generate the APTI database: (A) *Andalus*, (B) *Arabic Transparent*, (C) *AdvertisingBold*, (D) *Diwani Letter*, (E) *DecoType Thuluth*, (F) *Simplified Arabic*, (G) *Tahoma*, (H) *Traditional Aatbic*, (I) *DecoType Naskh*, (J) *M Unicode Sara*

2.4 Rendering procedure

The text images are generated using automated procedures. As a consequence, artefacts or noise usually present for scanned or camera-based documents are not present in the images. Such degradations could actually be artificially added, if needed [Baird 08], but it is currently out of the scope of APTI.

Image generation of text, for example on screen, can be done in many different ways. They are usually all leading to slight variations of the target image. We have opted for a rendering procedure that allows us to include effects of downsampling and antialiasing. These effects are interesting in terms of variability of the images, especially in low-resolution.

The procedure involves the downsampling of a high resolution source image into a low resolution image using antialiasing filtering. We also use different grid alignments to introduce variability in the application of the antialiasing filter. The details of the procedure are the following:

1. A gray-scale source image is generated in high resolution (360 pixels/inch) from the current word in the lexicon, using the selected font, size and style (Example in Fig. 2, height of image = 119, width of image =247).
2. Columns and rows of white pixels are added to the right hand side and to the top of the image. The number of columns and rows is chosen to have a height and width multiple of the downsampling factor (for example image in Fig. 3, we add 3

white columns and 1 white row). This effect allows to have the same deformation in all images and artificially moving the downsampling grid.

3. Downsampling and antialiasing filtering are applied to obtain the target image in lower resolution (72 pixels/inch) (Example in Fig. 3, height of image =24, width of image = 50). The target image is in grey level. The downsampling and antialiasing algorithms are the one implemented in the Java abstract class Image. In our implementation, we used the SCALE_SMOOTH option of the Java method which optimize the downsampling algorithm selection according to quality and speed.

Fig. 2: Example of Arabic image word source

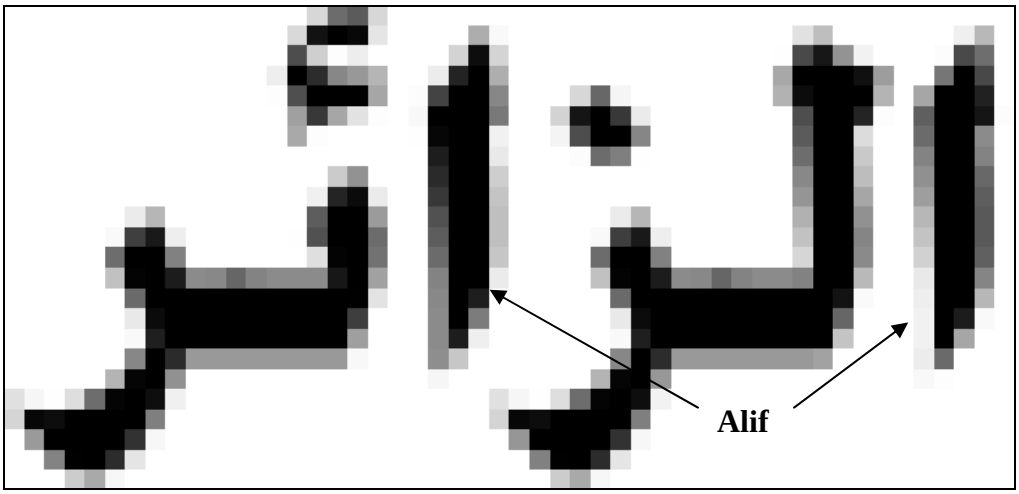

Fig. 3: Example of anti-aliasing effect and down sampling result approach

In Fig. 3, Character “Alif” is presented in two different forms (different presentation of anti-aliasing effect) in the same word image although it has the same characteristics (Font, Font Size, Style,...).

2.5 Sources of variability

The sources of variability in the generation procedure of text images in APTI are the following:

1. 10 different fonts: Andalus, Arabic Transparent, AdvertisingBold, Diwani Letter, DecoType Thuluth, Simplified Arabic, Tahoma, Traditional Arabic, DecoType Naskh, M Unicode Sara;
2. 10 different sizes: 6, 7, 8, 9, 10, 12, 14, 16, 18 and 24 points;
3. 4 different styles: plain, bold, italic, italic and bold;
4. Various forms of ligatures and overlaps of characters thanks to the large combination of characters in the lexicon and thanks to the used fonts;
5. Very large vocabulary that allows to test systems on unseen data;
6. Various artefacts of the downsampling and antialiasing filters due to the random insertion of columns of white pixels at the beginning of image words;
7. Variability of the height of each word image.

The last point of the previous list is actually intrinsic to the sequence of characters appearing in the word. In APTI, there is actually no a priori knowledge of the position of the baseline and it is up to the recognition algorithm to compute the baseline, if needed.

2.6 Ground truth description

Each word image in APTI database is fully described using an XML file containing ground truth information about the sequence of characters as well as information about the generation. An example of such XML file is given in Fig. 4.

```
  <?xml version="1.0" encoding="UTF-8" ?>
- <wordImage id="78">
  - <content transcription="الآلاف" nPaws="4">
      <paw id="1" nbChars="1">Alif_I</paw>
      <paw id="2" nbChars="2">Laam_B TildAboveAlif_E</paw>
      <paw id="3" nbChars="2">Laam_B Alif_E</paw>
      <paw id="4" nbChars="1">Faa_I</paw>
    </content>
    <font name="Arabic Transparent" fontStyle="Plain" size="24" />
    <specs encoding="png" width="96" height="36" effect="none" />
    <generation type="downsampling5" renderer="java" filtering="antialiasing" />
  </wordImage>
```

Fig. 4: Example of XML file including ground truth information about a given word image

The XML file is composed by four principal markups sections:

- *Content*: in this element, we have the transcription of Arabic word, the number of Piece of Arabic Word (nPaws) and sub-elements for each PAW with the sequence of characters. In our representation, characters are identified using plain English labels as described below.
- *Font*: in this element, we specify the font name, font style and size used to generate the image word.
- *Specs*: in this element, we present the encoding of image, width, height and eventual addition effect. In the current version of APTI, there is actually no added effects but we have planned to use this attribute for later versions of image rendering where effects could be present.
- *Generation*: in this element, we indicate the type of generation, the tool used for generation and the used filter in generation. In the current version of APTI, this element is constant as the same generation procedure has been applied. The type ‘downsampling5’ is here indicating that the generation procedure correspond to a

downsampling, using factor 5, from high resolution source images as explained in Section 2.4.

Letter label	Number of Occurence	Isolate	Begin	Middle	End
Alif	90353	ا		ـا	
Baa	28119	ب	بـ	ـبـ	ـب
Taaa	59343	ت	تـ	ـتـ	ـت
Thaa	3803	ث	ثـ	ـثـ	ـث
Jiim	11455	ج	جـ	ـجـ	ـج
Haaa	17866	ح	حـ	ـحـ	ـح
Xaa	8492	خ	خـ	ـخـ	ـخ
Daal	18399	د		ـد	
Thaal	3100	ذ		ـذ	
Raa	37571	ر		ـر	
Zaay	6325	ز		ـز	
Siin	21648	س	سـ	ـسـ	ـس
Shiin	8668	ش	شـ	ـشـ	ـش
Saad	8310	ص	صـ	ـصـ	ـص
Daad	5548	ض	ضـ	ـضـ	ـض
Thaaa	8610	ط	طـ	ـطـ	ـط
Taa	1438	ظ	ظـ	ـظـ	ـظ
Ayn	16552	ع	عـ	ـعـ	ـع
Ghayn	5912	غ	غـ	ـغـ	ـغ
Faa	13749	ف	فـ	ـفـ	ـف
Gaaf	16819	ق	قـ	ـقـ	ـق
Kaaf	12711	ك	كـ	ـكـ	ـك
Laam	41159	ل	لـ	ـلـ	ـل
Miim	47084	م	مـ	ـمـ	ـم
Nuun	44186	ن	نـ	ـنـ	ـن
NuunChadda	1343	نّ	نّـ	ـنّـ	ـنّ
Haa	16094	ه	هـ	ـهـ	ـه
Waaw	26008	و		ـو	
Yaa	40215	ي	يـ	ـيـ	ـي
YaaChadda	4348	يّ	يّـ	ـيّـ	ـيّ
Hamza	1142	ء			
HamzaAboveAlif	8770	أ		ـأ	
TaaaClosed	8376	ة			ـة
HamzaUnderAlif	1501	إ		ـإ	
AlifBroken	972	ى			ـى
TildAboveAlif	500	آ		ـآ	
HamzaAboveAlifBroken	1253	ئ	ئـ	ـئـ	ـئ
HamzaAboveWaaw	538	ؤ		ـؤ	
Quantity of Characters	***648'280***				
Quantity of PAWs	***274833***				
Quantity of words	***113'284***				

Table 1: Arabic letters with used labels and occurrence in APTI database

The different character labels are summarized in Table 1. As the shape of characters are varying according to their position in the word, the character labels also include a suffix to specify the position of the character in the word: “B” standing for beginning, “M” for Middle, “E” for end and “I” for isolated. The character “Hamza” being always isolated, we don’t use the position suffix for this character. We also artificially inserted characters labels such as “NuunChadda” or “YaaChadda” to represent the character shape issued from the combination of “Nuun” and “Chadda” or “Yaa” and “Chadda”.

3 Database statistics

The APTI database consists of 113’284 different single words presented in 10 fonts, 10 font-sizes and 4 font-styles. Table 2 shows the total quantity of word images, PAWs (Piece of Arabic Words), and characters in APTI database.

	Number of Words	*Number of PAWs*	*Number of characters*
	113’284	274’833	648’280
Number of Font	10	10	10
Number of Font Size	10	10	10
Number of Font Styles	4	4	4
Total	***45’313’600***	***109’933’200***	***259’312’000***

Table 2: Quantity of words, PAWs, characters in database

3.1 Division into sets

We have divided the database into six equilibrated sets to allow for flexibility in the composition of development and evaluation partitions. The words in each set are different but the distribution of all used letters is nearly the same in the various sets (see Table 3). The five first sets are available for the scientific community and the sixth set is kept internal for potential future evaluation of systems in blind mode.

The algorithm for the distribution of words in the different sets has been designed to have similar allocations of letters and words in all sets. The algorithm is presented in details in Fig. 5. The steps of the algorithms are the following. First (step 1 in Fig. 5), we read all the words from the database and we accumulate the number of occurrence of each used letters. The letters are then sorted according to their number of occurrence, from the smallest number of occurrence to the largest. Second, (step 2 in Fig. 5), bins (vectors) are created for each letters and they are ordered according to the occurrences computed in step 1. For each word of the database, we go through the bins and we look if the word contains the character associated to the bin. If yes, then the word is associated to the bin and we go to the next word. Doing this, we actually build sets of words having letters with low occurrences. Third (step 3 in Fig. 5), we go through each bin and distribute the word sequentially in our final 6 sets, emptying each bin one after the other.

In short, this procedure is simply stressing a fair distribution of words that include characters with few occurrences. Such a distribution is important to avoid that a given character is under-represented in a given set and therefore to avoid potential problems of during training or testing time.

```
# Inputs: list of Arabic Words
# Output: six Sets of Arabic words with similar distribution of words and characters
Begin

1. for all words w_i, i ∈ {1…113'285} in APTI
for all used letters l_i, i ∈ {1…38}
integer tab[]=findNbOccureneOfLetters(l_i);
endfor
endfor
increasingSort(tab);
2. for all words w_i, i ∈ {1…113'285} in APTI
   for all used letters l_j, j ∈ {1…38} sorted by NbOccurence
        if (l_j ⊂ w_i)
                add w_i in vector V_j, j ∈ {1…38}
                go to 2
        endif
    endfor
    endfor
3. for all V_s, s ∈ {1…38}
        read w_i, i ∈ {1…NbWordInV_s} from V_s
        if i mod 6=0
                add w_i in S_1
        if i mod 6=1
                add w_i in S_2
        if i mod 6=2
                add w_i in S_3
        if i mod 6=3
                add w_i in S_4
        if i mod 6=4
                add w_i in S_5
        if i mod 6=5
                add w_i in S_6
    endfor
end
```

Fig. 5: Algorithm used for distribution

Letter label	**Set 1**	**Set 2**	**Set 3**	**Set 4**	**Set 5**	**Set 6**
Alif	15078	14925	15165	15120	15046	15019
Baa	4513	4763	4692	4704	4730	4717
Taaa	9926	9884	9897	9797	9942	9897
Thaa	634	633	631	634	643	628
Jiim	1893	1897	1887	1924	1915	1939
Haaa	2953	2963	3017	2933	3000	3000
Xaa	1407	1435	1439	1401	1403	1407
Daal	3187	3033	3075	2990	3028	3086
Thaal	514	520	528	504	516	518
Raa	6304	6243	6169	6335	6253	6267
Zaay	1064	1054	1054	1066	1042	1045
Siin	3674	3556	3674	3512	3629	3603
Shiin	1457	1446	1418	1434	1455	1458
Saad	1374	1377	1388	1411	1371	1389
Daad	922	943	936	906	921	920
Thaaa	1419	1426	1431	1426	1446	1462
Taa	242	238	240	238	239	241
Ayn	2764	2823	2769	2718	2755	2723
Ghayn	981	970	983	984	990	1004
Faa	2305	2256	2221	2313	2339	2315
Gaaf	2784	2734	2853	2883	2762	2803
Kaaf	2101	2090	2099	2145	2136	2140
Laam	6745	6926	6972	7002	6790	6724
Miim	7871	7836	7957	7806	7797	7817
Nuun	7484	7433	7289	7316	7400	7264
NuunChadda	225	224	224	223	224	223
Haa	2670	2687	2590	2718	2705	2724
Waaw	4421	4313	4325	4333	4264	4352
Yaa	6641	6630	6876	6685	6648	6735
YaaChadda	725	727	709	719	735	733
Hamza	192	187	190	193	192	188
HamzaAboveAlif	1437	1483	1455	1512	1456	1427
TaaaClosed	1417	1407	1394	1364	1409	1385
HamzaUnderAlif	253	250	256	247	248	247
AlifBroken	162	161	164	163	161	161
TildAboveAlif	84	84	83	83	83	83
HamzaAboveAlifBroken	210	208	208	209	208	210
HamzaAboveWaaw	89	90	89	91	89	90
Quantity of Characters	***108'122***	***107'855***	***108'347***	***108'042***	***107'970***	***107'944***
Quantity of PAWs	***45'982***	***45'740***	***45'792***	***45'884***	***45'630***	***45'805***
Quantity of words	***18897***	***18892***	***18886***	***18875***	***18868***	***18866***

Table 3: Distribution of characters in the different sets of APTI

3.2 Distribution of letters in sets

Tables 4 to 9 are presenting the distribution of each shape of characters in their respective sets.

Letter label	Nb. Occ	Isolate	Begin	Middle	End
Alif	15078	ا 5823		ـا 9255	
Baa	4513	ب 128	بـ 1978	ـبـ 2226	ـب 181
Taaa	9926	ت 587	تـ 3626	ـتـ 5332	ـت 381
Thaa	634	ث 12	ثـ 261	ـثـ 341	ـث 20
Jiim	1893	ج 60	جـ 781	ـجـ 1016	ـج 36
Haaa	2953	ح 69	حـ 1135	ـحـ 1648	ـح 101
Xaa	1407	خ 16	خـ 587	ـخـ 782	ـخ 22
Daal	3187	د 988		ـد 2199	
Thaal	514	ذ 167		ـذ 347	
Raa	6304	ر 1813		ـر 4491	
Zaay	1064	ز 389		ـز 675	
Siin	3674	س 68	سـ 1434	ـسـ 2083	ـس 89
Shiin	1457	ش 18	شـ 580	ـشـ 831	ـش 28
Saad	1374	ص 14	صـ 439	ـصـ 882	ـص 39
Daad	922	ض 41	ضـ 358	ـضـ 497	ـض 26
Thaaa	1419	ط 42	طـ 392	ـطـ 920	ـط 65
Taa	242	ظ 6	ظـ 58	ـظـ 163	ـظ 15
Ayn	2764	ع 67	عـ 1003	ـعـ 1575	ـع 119
Ghayn	981	غ 12	غـ 413	ـغـ 543	ـغ 13
Faa	2305	ف 87	فـ 1213	ـفـ 923	ـف 82
Gaaf	2784	ق 97	قـ 937	ـقـ 1614	ـق 136
Kaaf	2101	ك 69	كـ 914	ـكـ 988	ـك 130
Laam	6745	ل 175	لـ 3546	ـلـ 2206	ـل 818
Miim	7871	م 177	مـ 4043	ـمـ 2844	ـم 807
Nuun	7484	ن 2437	نـ 1264	ـنـ 1905	ـن 1878
NuunChadda	225	ّن 0	ّنـ 0	ـّنـ 225	ـّن 0
Haa	2670	ه 223	هـ 704	ـهـ 1196	ـه 548
Waaw	4421	و 1621		ـو 2800	
Yaa	6641	ي 317	يـ 2516	ـيـ 2640	ـي 1168
YaaChadda	725	ّي 0	ّيـ 192	ـّيـ 533	ـّي 0
Hamza	192	ء 192			
HamzaAboveAlif	1437	أ 1102		ـأ 335	
TaaaClosed	1417	ة 441			ـة 976
HamzaUnderAlif	253	إ 182		ـإ 71	
AlifBroken	162	ى 53			ـى 109
TildAboveAlif	84	آ 32		ـآ 52	
HamzaAboveAlifBroken	210	ئ 3	ئـ 167	ـئـ 39	ـئ 1
HamzaAboveWaaw	89	ؤ 30		ـؤ 59	

Table 4: Distribution of letters in set 1

Letter label	Nb. Occ	Isolate	Begin	Middle	End
Alif	14925	ا 5777		ـا 9148	
Baa	4763	ب 150	بـ 2039	ـبـ 2344	ـب 230
Taaa	9884	ت 642	تـ 3551	ـتـ 5347	ـت 344
Thaa	633	ث 19	ثـ 230	ـثـ 349	ـث 35
Jiim	1897	ج 54	جـ 756	ـجـ 1034	ـج 53
Haaa	2963	ح 93	حـ 1159	ـحـ 1619	ـح 92
Xaa	1435	خ 18	خـ 622	ـخـ 777	ـخ 18
Daal	3033	د 963		ـد 2070	
Thaal	520	ذ 166		ـذ 354	
Raa	6243	ر 1823		ـر 4420	
Zaay	1054	ز 379		ـز 675	
Siin	3556	س 77	سـ 1338	ـسـ 2041	ـس 100
Shiin	1446	ش 22	شـ 558	ـشـ 838	ـش 28
Saad	1377	ص 22	صـ 420	ـصـ 906	ـص 29
Daad	943	ض 42	ضـ 374	ـضـ 492	ـض 35
Thaaa	1426	ط 38	طـ 401	ـطـ 925	ـط 62
Taa	238	ظ 7	ظـ 66	ـظـ 149	ـظ 16
Ayn	2823	ع 85	عـ 1074	ـعـ 1543	ـع 121
Ghayn	970	غ 15	غـ 444	ـغـ 495	ـغ 16
Faa	2256	ف 62	فـ 1184	ـفـ 937	ـف 73
Gaaf	2734	ق 104	قـ 872	ـقـ 1632	ـق 126
Kaaf	2090	ك 63	كـ 891	ـكـ 1002	ـك 134
Laam	6926	ل 193	لـ 3513	ـلـ 2334	ـل 886
Miim	7836	م 162	مـ 4152	ـمـ 2704	ـم 818
Nuun	7433	ن 2391	نـ 1262	ـنـ 1848	ـن 1932
NuunChadda	224	ّن 0	ّنـ 0	ـّنـ 224	ـّن 0
Haa	2687	ه 224	هـ 705	ـهـ 1201	ـه 559
Waaw	4313	و 1480		ـو 2833	
Yaa	6630	ي 317	يـ 2432	ـيـ 2701	ـي 1183
YaaChadda	727	ّي0	ّيـ 210	ـّيـ 517	ـّي0
Hamza	187	ء 187			
HamzaAboveAlif	1483	أ 1156		ـأ 327	
TaaaClosed	1407	ة 429			ـة 978
HamzaUnderAlif	250	إ 160		ـإ 90	
AlifBroken	161	ى 47			ـى 114
TildAboveAlif	84	آ 40		ـآ 44	
HamzaAboveAlifBroken	208	ئ 0	ئـ 166	ـئـ 34	ـئ 8
HamzaAboveWaaw	90	ؤ 32		ـؤ 58	

Table 5: Distribution of letters in set 2

Letter label	Nb. Occ	Isolate	Begin	Middle	End
Alif	15165	ا 5988		ـا 9177	
Baa	4692	ب 156	بـ 1955	ـبـ 2343	ـب 238
Taaa	9897	ت 617	تـ 3546	ـتـ 5380	ـت 354
Thaa	631	ث 16	ثـ 245	ـثـ 335	ـث 35
Jiim	1887	ج 53	جـ 784	ـجـ 998	ـج 52
Haaa	3017	ح 63	حـ 1194	ـحـ 1659	ـح 101
Xaa	1439	خ 11	خـ 643	ـخـ 765	ـخ 20
Daal	3075	د 947		ـد 2128	
Thaal	528	ذ 185		ـذ 343	
Raa	6169	ر 1746		ـر 4423	
Zaay	1054	ز 362		ـز 692	
Siin	3674	س 75	سـ 1411	ـسـ 2085	ـس 103
Shiin	1418	ش 18	شـ 545	ـشـ 827	ـش 28
Saad	1388	ص 17	صـ 390	ـصـ 948	ـص 33
Daad	936	ض 50	ضـ 346	ـضـ 511	ـض 29
Thaaa	1431	ط 39	طـ 393	ـطـ 937	ـط 62
Taa	240	ظ 1	ظـ 46	ـظـ 176	ـظ 17
Ayn	2769	ع 64	عـ 1015	ـعـ 1560	ـع 130
Ghayn	983	غ 12	غـ 423	ـغـ 530	ـغ 18
Faa	2221	ف 54	فـ 1178	ـفـ 910	ـف 79
Gaaf	2853	ق 107	قـ 984	ـقـ 1640	ـق 122
Kaaf	2099	ك 76	كـ 904	ـكـ 996	ـك 123
Laam	6972	ل 183	لـ 3606	ـلـ 2259	ـل 924
Miim	7957	م 190	مـ 4066	ـمـ 2899	ـم 802
Nuun	7289	ن 2319	نـ 1293	ـنـ 1811	ـن 1866
NuunChadda	224	نّ 0	نّـ 0	ـنّـ 224	ـنّ 0
Haa	2590	ه 192	هـ 631	ـهـ 1222	ـه 546
Waaw	4325	و 1507		ـو 2818	
Yaa	6876	ي 318	يـ 2527	ـيـ 2764	ـي 1270
YaaChadda	709	يّ 0	يّـ 198	ـيّـ 511	ـيّ 0
Hamza	190	ء 190			
HamzaAboveAlif	1455	أ 1133		ـأ 322	
TaaaClosed	1394	ة 435			ـة 959
HamzaUnderAlif	256	إ 169		ـإ 87	
AlifBroken	164	ى 58			ـى 106
TildAboveAlif	83	آ 39		ـآ 44	
HamzaAboveAlifBroken	208	ئ 4	ئـ 170	ـئـ 27	ـئ 7
HamzaAboveWaaw	89	ؤ 21		ـؤ 68	

Table 6: Distribution of letters in set 3

Letter label	Nb. Occ	Isolate	Begin	Middle	End
Alif	15120	ا 5866		ـا 9254	
Baa	4704	ب 132	بـ 1979	ـبـ 2362	ـب 231
Taaa	9797	ت 633	تـ 3625	ـتـ 5208	ـت 331
Thaa	634	ث 29	ثـ 219	ـثـ 360	ـث 26
Jiim	1924	ج 61	جـ 808	ـجـ 1016	ـج 39
Haaa	2933	ح 68	حـ 1205	ـحـ 1552	ـح 108
Xaa	1401	خ 16	خـ 615	ـخـ 749	ـخ 21
Daal	2990	د 909		ـد 2081	
Thaal	504	ذ 144		ـذ 360	
Raa	6335	ر 1833		ـر 4502	
Zaay	1066	ز 400		ـز 666	
Siin	3512	س 63	سـ 1349	ـسـ 2006	ـس 94
Shiin	1434	ش 17	شـ 596	ـشـ 796	ـش 25
Saad	1411	ص 19	صـ 422	ـصـ 937	ـص 33
Daad	906	ض 34	ضـ 381	ـضـ 457	ـض 34
Thaaa	1426	ط 34	طـ 399	ـطـ 929	ـط 64
Taa	238	ظ 0	ظـ 64	ـظـ 159	ـظ 15
Ayn	2718	ع 72	عـ 1016	ـعـ 1518	ـع 112
Ghayn	984	غ 12	غـ 399	ـغـ 566	ـغ 7
Faa	2313	ف 73	فـ 1264	ـفـ 894	ـف 82
Gaaf	2883	ق 106	قـ 999	ـقـ 1639	ـق 139
Kaaf	2145	ك 86	كـ 935	ـكـ 978	ـك 146
Laam	7002	ل 207	لـ 3656	ـلـ 2247	ـل 892
Miim	7806	م 157	مـ 3963	ـمـ 2848	ـم 838
Nuun	7316	ن 2341	نـ 1239	ـنـ 1860	ـن 1876
NuunChadda	223	نّ 0	نّـ 0	ـنّـ 223	ـنّ 0
Haa	2718	ه 201	هـ 681	ـهـ 1252	ـه 585
Waaw	4333	و 1494		ـو 2839	
Yaa	6685	ي 322	يـ 2443	ـيـ 2699	ـي 1221
YaaChadda	719	يّ 0	يّـ 215	ـيّـ 504	ـيّ 0
Hamza	193	ء 193			
HamzaAboveAlif	1512	أ 1164		ـأ 348	
TaaaClosed	1364	ة 398			ـة 966
HamzaUnderAlif	247	إ 171		ـإ 76	
AlifBroken	163	ى 42			ـى 121
TildAboveAlif	83	آ 38		ـآ 45	
HamzaAboveAlifBroken	209	ئ 5	ئـ 161	ـئـ 35	ـئ 8
HamzaAboveWaaw	91	ؤ 24		ـؤ 67	

Table 7: Distribution of letters in set 4

Letter label	Nb. Occ	Isolate	Begin	Middle	End
Alif	15046	ا 5689		ـا 9357	
Baa	4730	ب 161	بـ 1991	ـبـ 2341	ـب 237
Taaa	9942	ت 580	تـ 3629	ـتـ 5389	ـت 344
Thaa	643	ث 26	ثـ 242	ـثـ 347	ـث 28
Jiim	1915	ج 60	جـ 809	ـجـ 990	ـج 56
Haaa	3000	ح 83	حـ 1134	ـحـ 1680	ـح 103
Xaa	1403	خ 15	خـ 611	ـخـ 754	ـخ 23
Daal	3028	د 901		ـد 2127	
Thaal	516	ذ 159		ـذ 357	
Raa	6253	ر 1824		ـر 4429	
Zaay	1042	ز 386		ـز 656	
Siin	3629	س 59	سـ 1401	ـسـ 2091	ـس 78
Shiin	1455	ش 25	شـ 566	ـشـ 838	ـش 26
Saad	1371	ص 14	صـ 413	ـصـ 896	ـص 48
Daad	921	ض 41	ضـ 369	ـضـ 470	ـض 41
Thaaa	1446	ط 33	طـ 412	ـطـ 934	ـط 67
Taa	239	ظ 5	ظـ 52	ـظـ 169	ـظ 13
Ayn	2755	ع 68	عـ 1017	ـعـ 1552	ـع 118
Ghayn	990	غ 15	غـ 422	ـغـ 534	ـغ 19
Faa	2339	ف 73	فـ 1257	ـفـ 920	ـف 89
Gaaf	2762	ق 103	قـ 959	ـقـ 1574	ـق 126
Kaaf	2136	ك 84	كـ 914	ـكـ 980	ـك 158
Laam	6790	ل 188	لـ 3433	ـلـ 2288	ـل 881
Miim	7797	م 175	مـ 4067	ـمـ 2732	ـم 823
Nuun	7400	ن 2435	نـ 1273	ـنـ 1825	ـن 1867
NuunChadda	224	ںّ 0	ںّـ 0	ـںّـ 224	ـںّ 0
Haa	2705	ه 178	هـ 699	ـهـ 1297	ـه 531
Waaw	4264	و 1466		ـو 2798	
Yaa	6648	ي 327	يـ 2507	ـيـ 2656	ـي 1160
YaaChadda	735	يّ 0	يّـ 168	ـيّـ 567	ـيّ 0
Hamza	192	ء 192			
HamzaAboveAlif	1456	أ 1158		ـأ 298	
TaaaClosed	1409	ة 433			ـة 976
HamzaUnderAlif	248	إ 171		ـإ 77	
AlifBroken	161	ى 55			ـى 106
TildAboveAlif	83	آ 46		ـآ 37	
HamzaAboveAlifBroken	208	ئ 2	ئـ 167	ـئـ 32	ـئ 7
HamzaAboveWaaw	89	ؤ 28		ـؤ 61	

Table 8: Distribution of letters in set 5

Letter label	Nb. Occ	Isolate	Begin	Middle	End
Alif	15019	ا 5797		ـا 9222	
Baa	4717	ب 146	بـ 1998	ـبـ 2354	ـب 219
Taaa	9897	ت 641	تـ 3612	ـتـ 5304	ـت 340
Thaa	628	ث 22	ثـ 227	ـثـ 353	ـث 26
Jiim	1939	ج 49	جـ 803	ـجـ 1048	ـج 39
Haaa	3000	ح 83	حـ 1180	ـحـ 1655	ـح 82
Xaa	1407	خ 7	خـ 618	ـخـ 751	ـخ 31
Daal	3086	د 939		ـد 2147	
Thaal	518	ذ 164		ـذ 354	
Raa	6267	ر 1864		ـر 4403	
Zaay	1045	ز 377		ـز 668	
Siin	3603	س 73	سـ 1359	ـسـ 2062	ـس 109
Shiin	1458	ش 26	شـ 582	ـشـ 817	ـش 33
Saad	1389	ص 21	صـ 415	ـصـ 921	ـص 32
Daad	920	ض 43	ضـ 335	ـضـ 503	ـض 39
Thaaa	1462	ط 24	طـ 428	ـطـ 937	ـط 73
Taa	241	ظ 4	ظـ 65	ـظـ 158	ـظ 14
Ayn	2723	ع 80	عـ 1007	ـعـ 1510	ـع 126
Ghayn	1004	غ 15	غـ 425	ـغـ 540	ـغ 24
Faa	2315	ف 62	فـ 1226	ـفـ 928	ـف 99
Gaaf	2803	ق 99	قـ 974	ـقـ 1584	ـق 146
Kaaf	2140	ك 85	كـ 913	ـكـ 1004	ـك 138
Laam	6724	ل 174	لـ 3466	ـلـ 2203	ـل 881
Miim	7817	م 166	مـ 4038	ـمـ 2779	ـم 834
Nuun	7264	ن 2411	نـ 1231	ـنـ 1835	ـن 1787
NuunChadda	223	ںّ 0	ںّـ 0	ـںّـ 223	ـںّ 0
Haa	2724	ه 230	هـ 695	ـهـ 1236	ـه 565
Waaw	4352	و 1514		ـو 2838	
Yaa	6735	ي 301	يـ 2535	ـيـ 2652	ـي 1250
YaaChadda	733	يّ	يّـ 199	ـيّـ 534	ـيّ
Hamza	188	ء 188			
HamzaAboveAlif	1427	أ 1113		ـأ 314	
TaaaClosed	1385	ة 430			ـة 955
HamzaUnderAlif	247	إ 179		ـإ 68	
AlifBroken	161	ى 43			ـى 118
TildAboveAlif	83	آ 37		ـآ 46	
HamzaAboveAlifBroken	210	ئ 6	ئـ 164	ـئـ 39	ـئ 1
HamzaAboveWaaw	90	ؤ 23		ـؤ 67	

Table 9: Distribution of letters in set 6

4 Evaluation Protocols

In this section, we propose the definition of a set of robust benchmarking protocols on top of the APTI database. Preliminary experiments with a baseline recognition system have helped in calibrating and validating these protocols.. From the obtained results, we believe that the large number of data available in APTI and the different source of variability (cf Section 2.5) make it well suited for significant and challenging evaluation of systems.

4.1 Error estimation

The objective of any benchmarking of recognition systems is to estimate, as reliably as possible, the classification error rate $\hat{P}_e$. It is important to keep in mind that, whatever the task and data used, $\hat{P}_e$ is a function of the split of the data into training and test sets. Different splits will result in different error estimates. Hopefully, APTI is composed of quite large sets of data, which is helping in reaching stable estimates of $\hat{P}_e$.

Our objective is then to obtain a reliable estimate of $\hat{P}_e$ while keeping the computation load tractable. Therefore, we have opted for a *rotation method*, as described in [Jain 00, Section 7]. The idea is to reach a trade-off between the *holdout method* which leads to pessimistic and biased values of the error rate and the *leave-one-out method* that gives a better estimate but at the cost of larger computational requirements. The rotation method we are proposing is illustrated in Fig. 6. The procedure is to perform independent runs on 5 different partitions between training and testing data.

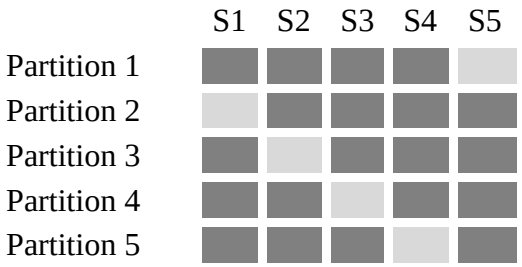

Fig. 6: Illustration of the rotation method. For a given partition, the training sets are depicted in dark grey and the testing sets in light grey.

The final error estimate is taken as the average of the error rates obtained on the different partitions.

$$\hat{P}_e = \frac{1}{5}\sum_{i=1}^{5} \hat{P}_{e,i}$$

In the previous formula, $\hat{P}_{e,i}$ is the error rate obtained independently on a system trained and tested using the sets defined in partition *i*. The procedure actually corresponds to computing the average of performance of 5 independent systems.

4.2 Train and test conditions

Using the procedure described in section 4.1, we can define different combinations of train and test conditions. The objectives are to measure the impact of some of the variability of the data. We therefore propose 20 protocols as summarized in Table 3.

The notations Tr(font, style, size) and Te(font, style, size) define the training and testing conditions with:

1. the font label as indicated in Fig. 1
2. the style where *p*, *i*, *b* and *bi* are for plain, italic, bold and bold+italic
3. the size in points

We suggest researchers willing to define new protocols to use this notation to specify the conditions of their training and testing.

Protocol name	Train choice Tr(font, Style, Size)	Test choice Te (font, Style, Size)
APTI 1	Tr(B, p, 10)	Te(B, p, 10)
APTI 2	Tr(B, p, 10)	Te(B, i, 10)
APTI 3	Tr(B, p, 10)	Te(B, b, 10)
APTI 4	Tr(B, p, 10)	Te(B, bi, 10)
APTI 5	Tr(B, p, [6, 10, 14, 18])	Te(B, p, [6, 10, 14, 18])
APTI 6	Tr(B,[p,i,b], [6, 10, 14, 18])	Te(B,[p,i,b], [6, 10, 14, 18])
APTI 7	Tr([A,B,C,F,H], p, 10)	Te([A,B,C,F,H], p, 10)
APTI 8	Tr([D,E,G,I,J], p, 10)	Te([D,E,G,I,J], p, 10)
APTI 9	Tr([A,B,C,F,H], [p,i,b], 10)	Te([A,B,C,F,H], [p,i,b], 10)
APTI 10	Tr([D,E,G,I,J], [p,i,b], 10)	Te([D,E,G,I,J], [p,i,b], 10)
APTI 11	Tr([A,B,C], p, 10)	Te([F,H], p, 10)
APTI 12	Tr([D,E,G], p, 10)	Te([I,J],i, 10)
APTI 13	Tr([A,B,C], p,[6,10,14,18])	Te([F,H], p, [6,10,14,18])
APTI 14	Tr([D,E,G], p,[6,10,14,18])	Te([I,J], p, [6,10,14,18])
APTI 15	Tr(B, p, 6)	Te(B, p, 6)
APTI 16	Tr(B, p, 8)	Te(B, p, 8)
APTI 17	Tr(B, p, 10)	Te(B, p, 6)
APTI 18	Tr(B, p, 6)	Te(B, p, 10)
APTI 19	Tr(B, p, [6, 10, 14, 18])	Te(B, p, [7,9,12,24])
APTI 20	Tr(all, all, all)	Te(all, all, all)

Table 3: APTI protocols

The objectives behind the protocols of Table 3 can be explained as follows:

- **APTI 1**: This is the baseline protocol where performances should be the highest as there are no mismatched between training and testing conditions.
- **APTI 2,3,4**: We measure here the capability of systems trained using plain style to generalize on italic, bold and bold+italic.
- **APTI 5,6**: While using the same font, we measure the capability of the system to treat different sizes.
- **APTI 7,8,9,10**: These experiments measure the capability of systems to recognize muti-font text.
- **APTI 11,12,13,14**: We measure the capability of systems to recognize unseen fonts text.
- **APTI 1,15,16,17,18,19**: Firstly, we measure the potential degradation of performance using smaller sizes. Secondly, we measure the capability to recognize unseen sizes.
- **APTI 20**: This is the global experiment where all available data is used for training and testing.

5 Conclusions

APTI, a new large Arabic printed text images database is presented together with evaluation protocols. APTI aims at the large-scale benchmarking of open-vocabulary text recognition systems. While it can be used for the evaluation of any OCR systems, APTI is, by nature, well suited for the evaluation of screen-based OCR systems. The challenges addressed by the database are in the variability of the sizes, fonts and style and the protocols that are defined are crafted to put into evidence the impact of such variability. APTI will be made publicly available for the purpose of research.

6 References

[Pechwitz 02] M. Pechwitz, S. S. Maddouri, V. Maergner, N. Ellouze, and H. Amiri. IFN/ENIT - database of handwritten Arabic words. In Proc. of CIFED 2002, pages 129–136, Hammamet, Tunisia, October 21-23 2002

[Slimane 08] F. Slimane, R. Ingold, M. A. Alimi and J. Hennebert, Duration Models for Arabic Text Recognition using Hidden Markov Models. CIMCA 2008, Vienne, Austria, December 10-12 2008

[Jain 00] A. K. Jain, R. Duin and J. Mao, Statistical Pattern Recognition: A Review, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 1, January 2000

[Khorsheed 07] M. S. Khorsheed, Offline recognition of omnifont Arabic text using the HMM ToolKit (HTK). Pattern Recognition Letters 28(12): 1563-1571, 2007

[Schlosser 95] S. Schlosser, “ERIM Arabic Database”, Document Processing Research Program, Information and Materials Applications Laboratory, Environmental Research Institute of Michigan, 1995

[Margner 05] V. Margner, M. Pechwitz, H. El Abed, “Arabic Handwriting Recognition Competition”, In ICDAR, 2005, pp.70 – 74

[Margner 07] V. Margner and H. E. Abed. “ICDAR 2007 Arabic handwriting recognition competition”. In ICDAR, Sept. 2007 vol. 2, pp. 1274–1278.

[Graff 06] D. Graff, K. Chen, J. Kong, and K. Maeda, “Arabic Gigaword Second Edition”, Linguistic Data Consortium, Philadelphia, 2006

[Abbes 04] R. Abbes, J.D. Hassoun, “The Architecture of a Standard Arabic Lexical Database, Some Figures, Ratios and Categories from the DIINAR.1 Source Program”, Workshop of Computational Approaches to Arabic Script-Based Languages, Geneva, 2004

[Husni 08] Husni A. Al-Muhtaseb, Sabri A. Mahmoud, Rami S. Qahwaji, Recognition of off-line printed Arabic text using Hidden Markov Models. European Signal Processing Conference. Vol. 88, Issue 12, Pages 2902-2912, Lausanne, Switzerland, August 25-29, 2008

[Shaaban 08] Z. Shaaban, A New Recognition Scheme for Machine-Printed Arabic Texts based on Neural Networks. Proceedings of World Academy of Science, Engineering and Technology, Vol. 31, Vienna, Austria, July 25-27 2008

[AbdelRaouf 08] A. AbdelRaouf, C. A Higgins, and M. Khalil, A Database for Arabic Printed Character Recognition. ICIAR 2008, LNCS 5112, pages 567–578, 2008.

[Kanoun 2005] S. Kanoun, A. M. Alimi, Y. Lecourtier, "Affixal approach for Arabic decomposable vocabulary recognition a validation on printed word in only one font", In ICDAR, Sept. 2005, vol. 2 pp.1025 - 1029

[Baird 08] H. S. Baird. "State of the Art of Document Image Degradation Modeling". Proceedings of the 4th IAPR Workshop on Document Analysis Systems, DAS 2000.