A Novel Minimal Script for Arabic Text Recognition Databases and Benchmarks

Husni A. Al-Muhtaseb, Sabri A. Mahmoud, and Rami S. Qahwahi

***Abstract*—** This paper presents a minimal Arabic text that covers the different basic shapes of Arabic alphabet (viz. standalone, initial, medial, and terminal). It is designed with minimal repetition of character shapes in the minimal text. The novelty of the suggested script could be seen from different perspectives. It enables the collection of handwritten text from different writers with minimized effort and time. It is enough for a writer to write three lines of meaningful Arabic text to cover all possible character shapes, a total of 125 shapes. The written text is designed to have even distribution of letter frequencies. This assures enough samples of all character shapes when text is collected from enough number of writers. The same is true for printed Arabic text. This is especially useful when using large number of features with classifiers that require large number of samples for each category. Hidden Markov Models and Neural networks are two examples of these classifiers. The use of the minimal text enables proper training, as all Arabic character shapes are present with adequate frequency, hence resulting in higher recognition rates. This is not the case with natural text where the frequency of some Arabic characters differ widely, where in some cases 100 folds or more. The proposed minimal text may be used to build a data base of handwritten Arabic text collected of many writers. This covers the need for a database in the research of Arabic handwritten text recognition and benchmarking.

In addition, this paper presents statistical analysis of Arabic corpora for estimating the number of occurrences of the different shapes of Arabic characters in large corpora. The frequency of Arabic characters could be used in different applications. In this research work, it was utilized in enhancing the search for the minimal Arabic text.

***Keywords*—** Arabic text recognition, Arabic OCR databases, Minimal Arabic script.

I. Introduction

THIS paper addresses a need for Arabic text recognition research as noted by several researchers [1] - [3]. Although Arabic text recognition is receiving more attention in current decade [4] – [19], there is a lack of Arabic text databases (both typed and hand-written) for use in Arabic text recognition system (ATRS) research. Such databases should include printed Arabic text with several types and sizes of fonts as well as handwritten text written by different writers. This need is addressed in this paper in two ways; by presenting a minimal text that covers all the shapes of all Arabic alphabets and by estimating the frequencies of the use of Arabic alphabet in different shapes and positions. The frequencies are estimated by statistically analyzing Arabic documents of over 4 million characters. This combination (viz. the minimal Arabic text and the frequencies of Arabic alphabets in corpora of Arabic text) may be used as alternative to the use of text databases in the research of Arabic text recognition.

Few Arabic databases with limited contents are available for research in ATRS. Some of them have been prepared for specific domains and applications such as checks, numerals contents, and postal addresses. Arabic literal amounts of 4800 words were used in [20]. A database consisting of 26,459 Arabic names, presenting 937 Tunisian town/ village names, handwritten by 411 different writers is presented in [21], [22] and used in several research work [23], [24]. A database prepared from text of 100 writers each wrote 67 literal numbers, 29 most popular words in Arabic writing, three sentences representing numbers and quantities used in checks, and a free subject chosen by the writer (around 4700 handwritten words produced by 100 writers.) is discussed in [25] - [28]. A small database for digits that involves 17 writers each wrote 10 digits 10 times each is presented in reference [29]. Arabic and Persian isolated characters database consisting of 220,000 handwritten forms filled by more than 50,000 writers with isolated characters is presented in [30]. The database in [31], [32] presents 29,498 images of sub-words, 15,175 images of Indian digits and image samples of each of legal and courtesy amounts from 3000 real-life bank checks. Another database for bank-checks includes 70 words of Arabic literal amounts extracted from 5000 checks by 100 writers was used in [33], [34]. An automatically generated printed database of 946 Tunisian town names is discussed in [35]. 360 handwritten addresses of around 4000 words were used in [36]. [37] Presents a database consisting of more than 17820 names of 198 cities of Iran. A general database with signatures has 37,000 words, 10,000 digits, 2,500 signatures, and 500 free-form Arabic sentences [38]. A small isolated character database consisting of 50 images for each character by 5 writers; each wrote the whole 28-character alphabet ten times was used in [39]. DARPA Arabic Corpus consists of

Manuscript received April 11, 2009: Revised version received May 4, 2009. This work was supported in part by King Fahd University of Petroleum & Minerals.

Husni Al-Muhtaseb is with the Department of Information & Computer Science at King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia. (Phone: +966503820471; e-mail: muhtaseb@kfupm.edu.sa).

Sabri Mahmoud is with the Department of Information & Computer Science at King Fahd University with Petroleum & Minerals, Dhahran 31261, Saudi Arabia. (e-mail: smasaad@kfupm.edu.sa).

Rami Qahwaji is with the School of Computing, Informatics and Media at Bradford University, UK. (e-mail: R.S.R.Qahwaji@brad.ac.uk).

345 scanned pages of printed text in 4 different fonts [40]. The system in [41] was based on 40 pages of DARPA database. The research presented in [42] used 700 digitized pages from 45 printed documents. The segmentation work in [43] was based on around 240 digitized pages written by 178 writers, where each writer wrote one or two pages of 10 prepared pages of 13 lines each.

By reviewing the available databases for Arabic text recognition we found out that none of the databases contain even distribution of character frequencies of the collected data. In some cases, some characters are appearing 50 times or more than other characters (see [43] as an example). Moreover, if a handwritten database is targeted, it will be very hard to require writers to write long text, say one page or more. Even if we are able to collect two handwritten pages or more per writer some of the characters may not be present in the text with adequate frequency. In a test for training of a Hidden Markov Model using an arbitrary 2500 lines of Arabic text some characters had as low as 8 occurrences while other characters had over 1000 occurrences. It was noticed that characters with low number of occurrences had low correct recognition rates as expected due to the lack of enough samples for training. These reasons are some of the motivations that encouraged us to look for a minimal Arabic script to be used for preparing databases and benchmarks for Arabic text recognition research especially for handwritten text. To our knowledge this is the first attempt at addressing this problem using minimal text and hence no references are available for citation on minimal text coverage of Arabic language.

It is clear from the above literature review that there is no adequate Arabic text databases freely available for use in the research of Arabic handwritten (and typewritten) text. The minimal text of this paper in addition to the work of the authors presented in [44], which presented the probabilities of the occurrence of the Arabic alphabets in different positions of Arabic words, is an efficient solution to the above problem.

This paper which is an extension of [45] is organised as follows: the characteristics of Arabic text are described in Section II. Section III describes briefly the used corpora. The process of finding the minimal Arabic script is discussed in Section IV. The implementation of the system and experimental results are reported in Section V. Finally, the concluding remarks are given in Section VI.

II. Some Characteristics Of Arabic Text

Arabic is a cursive language written from right to left. It has 28 basic alphabets. An Arabic letter might have up to four different shapes depending on the position of the letter in the word: whether it is a standalone letter, connected only from right (initial form), connected only from left (terminal form), or connected from both sides (medial form). Letters of a word may overlap vertically (even without touching). Arabic letters do not have fixed size (height and width). Letters in a word can have diacritics (short vowels) such as *Fat-hah, Dhammah, Shaddah, sukoon* and *Kasrah*. Moreover, *Tanween* may be formed by having double *Fat-hah*, *Dhammah*, or *Kasrah.* Fig. 1 lists these diacritics. These diacritics are written as strokes,

Fat-hah ـَ Dhammah ـُ Shaddah ـّ
Kasrah ـِ Sukoon ـْ TanweenFat-h ـً
Tanween Dhamm ـٌ Tanween Kasr ـٍ

Fig. 1 Arabic short vowels (diacritics)

placed either on top of, or below, the letters. A different diacritic on a letter may change the meaning of a word. Readers of Arabic are accustomed to reading un-vocalized text by deducing the meaning from context. Fig. 2 shows some of the characteristics of Arabic text. It shows a base line, overlapping letters, diacritics, and two shapes of *Noon* character (initial and medial).

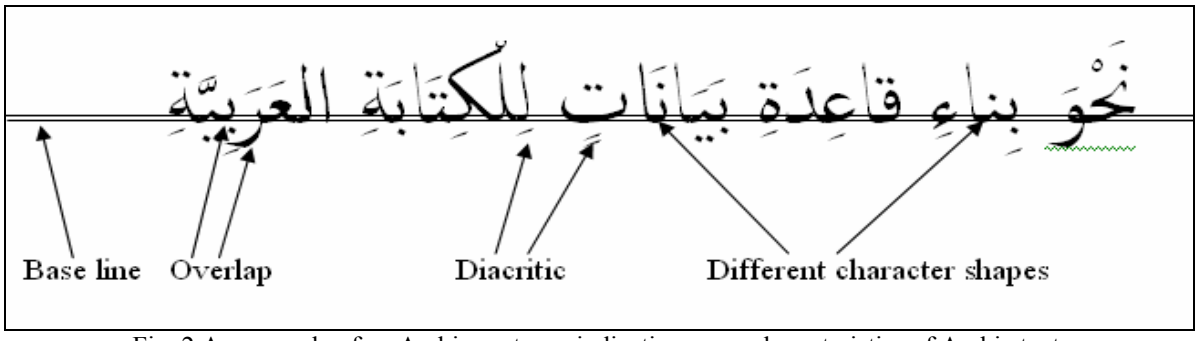

Fig. 2 An example of an Arabic sentence indicating some characteristics of Arabic text.

As Arabic numbers are not connected and are used globally, we concentrated our work on Arabic letters throughout this paper. As we stated earlier, Arabic has 28 main letters as shown in Fig. 3. When considering presenting Arabic

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 ه و ي

Fig. 3 Basic Arabic 28 letters.

characters in computers, some of the main letters had been extended into separate letters for easiness of presentations and usability by the Arab Standardization and Metrology Organization (ASMO). The standard Arabic codepages (character sets) ASMO-449, ASMO-708 and ISO 8859-6 define 36 Arabic letters (See Fig. 4). When we consider Arabic optical character recognition, we need to add *Lam-Alef* in its 4 different forms. Although *Lam-Alef* is a sequence of two alphabets, they are written as one set. We think that we need to treat this sequence as one set. So, we added four more

ء آ أ ؤ إ ئ ا ب ة ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك
ل م ن ه و ى ي

Fig. 4 Extend Arabic letters by ASMO.

sets to the alphabets; one with bare *Alef*, the second with *Alef-Maddah*, the third with *Alef-up-Hamza* and the fourth with *Alef-down-Hamza* as shown in Fig. 5. This leads us to expand the number of Arabic letters to 40. Each alphabet can take different number of shapes (from 1 to 4). Hence, the total

ء آ أ ؤ إ ئ ا ب ة ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك
ل م ن ه و لآ لأ لإ ا ى ي

Fig. 5 Expanded Arabic alphabets by adding different versions of Lam-Alef sequence.

number of shapes is 125 (one letter has only one shape, others have two, and the most have four shapes).

Table 1 shows the basic Arabic letters with their categories. We are grouping them into 3 different classes according to the number of shapes a letter takes. The first Class (class 1) consists of a single shape of the *Hamza* which comes in stand-alone state (Number 1 in Table 1). *Hamza* does not connect with any other letter. The second class (class 2) presents the letters that can come either standalone or connected only from right (medial category). This class consists of *Alef Madda, Alef up Hamza, Waw Hamza, Alef down Hamza, Alef, Tah Marboutah, Dal, Dhal, Ra, Zain, Waw, Lam Alef Madda, Lam Alef Hamza up, Lam Alef Hamza down, and Lam Alef* (numbers 2-5, 7, 9, 15-18, and 35-39 in Table 1). The third class (class 4) consists of the letters that can be connected from either side or both sides as well as they can appear as standalone. This class consists of *Hamza Kursi, Baa, Taa, Thaa, Jeem Haa, Khaa, Seen, Sheen, Sad, Dhad, Dhaa, THaa, Ain, Gain, Faa, Qaaf, Kaaf, Laam ,Meem, Noun, Haa, Yaa* (numbers 6, 8, 10-14, 19-33, and 40 in Table 1). Table 2 shows a summary of these classes.

Although an Arabic letter might have up to 4 different shapes, each letter is saved using only one code. It is the duty of a computer built-in driver to make contextual analysis to decide the right shape to display, depending on the previous and next characters if available.

When it is needed to consider different shapes of Arabic letters for a given Arabic text file, a contextual analysis algorithm is needed. Such algorithm takes the letter, its predecessor, and its successor and identifies the right shape depending on the classes of the letters. This algorithm was implemented in this work to label characters to their positional shapes.

III. The Used Corpora

The used corpora for our analysis consist of Arabic text of two Arabic lexicons [46], [47], two *HADITH* books [48], [49] and a lexicon containing the meaning of Quran tokens in Arabic [50]. The electronic versions of such books and other old Arabic classical books could be found from different sites in the internet including [51].

Table 1 Basic Shapes of Arabic alphabets.

no	S-alone	Term.	Medial	Initial	Shapes	Class
1	ء	ء	ء	ء	1	1
2	آ	ـآ	ـآ	آ	2	2
3	أ	ـأ	ـأ	أ	2	2
4	ؤ	ـؤ	ـؤ	ؤ	2	2
5	إ	ـإ	ـإ	إ	2	2
6	ئ	ـئ	ـئ	ئ	4	3
7	ا	ـا	ـا	ا	2	2
8	ب	ـب	ـبـ	بـ	4	3
9	ة	ـة	ـة	ة	2	2
10	ت	ـت	ـتـ	تـ	4	3
11	ث	ـث	ـثـ	ثـ	4	3
12	ج	ـج	ـجـ	جـ	4	3
13	ح	ـح	ـحـ	حـ	4	3
14	خ	ـخ	ـخـ	خـ	4	3
15	د	ـد	ـد	د	2	2
16	ذ	ـذ	ـذ	ذ	2	2
17	ر	ـر	ـر	ر	2	2
18	ز	ـز	ـز	ز	2	2
19	س	ـس	ـسـ	سـ	4	3
20	ش	ـش	ـشـ	شـ	4	3
21	ص	ـص	ـصـ	صـ	4	3
22	ض	ـض	ـضـ	ضـ	4	3
23	ط	ـط	ـطـ	طـ	4	3
24	ظ	ـظ	ـظـ	ظـ	4	3
25	ع	ـع	ـعـ	عـ	4	3
26	غ	ـغ	ـغـ	غـ	4	3
27	ف	ـف	ـفـ	فـ	4	3
28	ق	ـق	ـقـ	قـ	4	3
29	ك	ـك	ـكـ	كـ	4	3
30	ل	ـل	ـلـ	لـ	4	3
31	م	ـم	ـمـ	مـ	4	3
32	ن	ـن	ـنـ	نـ	4	3
33	ه	ـه	ـهـ	هـ	4	3
34	و	ـو	ـو	و	2	2
35	لآ	ـلآ	ـلآ	لآ	2	2
36	لأ	ـلأ	ـلأ	لأ	2	2
37	لإ	ـلإ	ـلإ	لإ	2	2
38	لا	ـلا	ـلا	لا	2	2
39	ى	ـى	ـى	ى	2	2
40	ي	ـي	ـيـ	يـ	4	3

IV. The Minimal Arabic Script

The novel idea, which we are introducing here, is to use a script that consists of minimum number of letters (using meaningful Arabic words) that covers all possible shapes

Table 2 Classes of Arabic alphabets depending on number of possible basic shapes.

Class	# of possible shapes	Alphabets
1	1	ء
2	2	آ أ ؤ إ ا ة د ذ ر ز و لآ لأ لإ ا ى
3	4	ئ ب ت ث ج ح خ س ش ص ض ط ظ ع غ ف ق ك ل م ن ه ي

when preparing databases and benchmarks for Arabic optical character recognition. Several utility programs, implementing different algorithms to address this issue, were written to search huge corpora of Arabic script to find the minimum meaningful words that cover all Arabic alphabet-shapes.

As we stated earlier, Arabic text is saved using a unique code for each character irrespective of its position and shape. When statistics of Arabic character shapes are required, a procedure needs to be used to figure out the shape of each character. An algorithm was implemented to decode to tag the letters with their position code in the word according to the position of the letter in the word (initial. medial, terminal, or isolated). The classes used in the algorithm are those that are presented earlier and summarized in Table 2.

Fig. 6 shows the pseudocode of an algorithm (processWord) to process words extracted from Arabic corpora to generate the minimal text. These are the main steps of the processWord algorithm

1. Initially the word is validated to see if it is already in the minimal text, if this is true then the word is not processed and the function is terminated.
2. The word is decoded to give the proper letter shapes of the word using the implemented contextual analysis algorithm.
3. The word is validated for multiple occurrences of a letter, if it has multiple occurrences of a letter then the word is not processed and the function is terminated as multiple occurrences are not allowed.
4. Each letter of the word is checked with the letters' table (holding the different shapes of Arabic alphabet). If any letter in the word is already flagged in the letters' table then the word is ignored and the function is terminated.

```
Definition of used variable/parameters:

aword: Arabic word
wordShapes: List of aword letters with specific shapes
element: a letter from wordShapes
minTextTable: A table holding minimum text
alphabetTable: Arabic alphabet table including extra column for flagging used
   letters.
characerTagged: A flag to indicate tagging of letters

function processWord(aword)
{ //checks if the word is already in minTextTable
  if(aword is in minTextTable)
        exit;

  Decode the word into letter shapes and put them in wordShapes
// Each letter of the word is given letter and shape code by the
// implemented contextual analysis algorithm.

  charTagged = 0;   // initialize charTagged flag to 0

  for(i=0;i< count(wordShapes);i++)(
    element = wordShapes[i]
       If(element is flagged in alphabetTable){ //Was this letter used?
        charTagged=1;
        break;
   }
  }

   if(not(charTagged)){
     Add word to minTextTable;  //add aword to minimal text
   for(i=0;i< count(wordShapes);i++){ //tag aword letters in alphabetTable
     element = wordShapes[i]
     flag element in alphabetTable;
   }
   charTagged = 0; // clear tagging flag
  }

}
```

Fig. 6 Pseudocode code for processing Arabic words for the generation minimal text to cover all Arabic letters in all positions of a word.

5. If a word passes the previous validations then
   a. The word is added to the minimal text, and
   b. The shapes of the all-shapes table corresponding to the letters of the word are flagged.

Several search criteria was conducted on the corpora to generate the minimal Arabic text using the *processWord* function. In one iteration the corpora were searched sequentially for words from the beginning taking each word into account. This process was continued until the whole corpora are searched. Then the resulting letters' table is checked for un-flagged letters. It was clear that this process could not flag all shapes and hence the minimum text was not covering all shapes. In other iterations, different sequence of the data in the corpora was selected (i.e. the sequence of searching the corpora was changed several times). This did not result in acquiring a minimum text to cover all Arabic alphabet shapes. In another iteration the words are randomly selected from the corpora. This showed better results, however, the minimal produced text was not including all Arabic alphabet shapes. Another search algorithm was executed which starts by selecting words having letters of minimal frequencies of usage utilizing the estimated frequencies of Arabic alphabet in Arabic script which is presented in Section IV. Hence, less frequently used letters were given higher priority. This resulted in improving the minimal produced text. In all of these experiments there was a constraint of not using a letter shape more than one time. It is to be noted that several versions of the processWord functions were tested in this work.

The table of Arabic alphabets (Table 1) was analyzed and it showed that there are 39 shapes of letters that might come at the end of a word (terminal form) and 23 shapes of the letters that might come at the beginning of a word (initial form). Hence, there should be some repetitions of the letters shapes that come at the beginning. In order to include all the shapes of Arabic letters the previous search algorithms were applied again, allowing the possibility of an initial beginning letter to have up to two occurrences. In addition to these constraints, we limited the total number of extra occurrences of these letters to 16. By using corpora of around 20 Megabytes of text and running programs for several days, we could reach to a nearly optimal script. An early minimal script has been identified as shown in Fig. 7. The script then has been

جعثق سؤق ذم ظأب رث مجس غضبى قمين شتف وس ضغط أي بحظل
حذف نسكه طخفة خصهم صئك إظ فت زؤى كنب دك آخ ال هوج ثآئي طء
لطاع يقره عزة تشدح لاغ لأص لآت لإض ش ئ مج حث جح صخ فإن
يجئ نص قش عض لظ بلغ سع

Fig. 7 An Early minimal Arabic Script.

optimized manually in several iterations until reaching the presented text in Fig. 8.

The statistics of letter shape distribution for the suggested

عزة كأب جنة طفق ميس غضبى كثف ضغط أص فظل حذف نسكه خصتهم
صئك إظ رؤى شعب دك آخ ملأ سطوع يقرء تشدح لاغ للآت لإض هج
حث جح صخ فإي يجئ نخص قش عض تحظ بلغ سع ظمآن قن طائي ثلاث
لأج لأه بؤس ذم ائت للإوز ق ط ش ل ئ

Fig. 8 The Minimal Arabic Script.

minimal Arabic script is shown in Table 3. It is clear from the table that 16 initial letter shapes have two occurrences each to compensate for the extra shapes of Arabic letter shapes that come at the end of a word (terminal). All other letter shapes are used only once. Hence, the presented text is the minimal possible text that covers all the shapes of Arabic alphabets. It is minimal in terms of the number of shapes used.

It might be clear to the reader that the minimal script is not unique. Theoretically speaking, there is infinite number of

Table 3 Minimal text usage of the different shapes of Arabic alphabets.

Letter	Stand-alone	Term-inal	Initial	Medial
ء	1			
آ	1	1		
أ	1	1		
ؤ	1	1		
إ	1	1		
ئ	1	1	2	1
ا	1	1		
ب	1	1	2	1
ة	1	1		
ت	1	1	2	1
ث	1	1	1	1
ج	1	1	2	1
ح	1	1	2	1
خ	1	1	1	1
د	1	1		
ذ	1	1		
ر	1	1		
ز	1	1		
س	1	1	2	1
ش	1	1	1	1
ص	1	1	2	1
ض	1	1	1	1
ط	1	1	2	1
ظ	1	1	1	1
ع	1	1	2	1
غ	1	1	1	1
ف	1	1	2	1
ق	1	1	2	1
ك	1	1	2	1
ل	1	1	2	1
م	1	1	2	1
ن	1	1	2	1
ه	1	1		
و	1	1		
لآ	1	1		
لأ	1	1		
لإ	1	1		
لا	1	1		
ى	1	1	1	1
ي	1	1	2	1

different scripts. A main characteristic of all these minimal scripts is that they all should have only 141 Arabic alphabet shapes that cover all Arabic letter shapes.

V. Arabic Script Statistics

Samples of classical Arabic corpora have been analyzed to find the distribution of basic Arabic letter shapes used in Arabic Script. A text of 4.25 million characters has been used. The text consists of Two *HADITH* books, *Al-Bukari* and *Muslim*. Tables 4 - 6 present the frequencies of using the different forms of Arabic letters in the books of Bukhari, Muslim, and Bukhari and Muslim, respectively. The frequencies of usage of Arabic may be biased by the type of text being searched. However, the frequency of usage of Arabic letter shapes of different books gave close results. The researchers will extend the search to more data in the future. We could have domain specific alphabet usage frequencies in addition to the general frequencies. As an alternative, previous work of the researchers [28] which addressed this issue by giving the probabilities of the usage of each letter in the different positions of Arabic words may be used.

Table 4 Alphabet shapes distribution in classic Arabic for *Al-Bukari* book.

Let.		Term.	Initial	Medial	Total
	11896	0	0	0	11896
ا	213359	296148	0	0	
إ	29670	6321	0		35991
أ	103938	22656		0	126594
آ	3551		0	0	5041
ب		9922	140434	67938	233835
	17597	37078	0	0	54675
ت	6132	27033	29353	35826	
ث	2261	4368	49989		65367
ج	2964	1496		13394	41671
ح	3258	3388	69860	31432	107938
خ	243	264	22807	7089	30403
د	18950	114451	0	0	133401
ذ	13526	15441	0	0	28967
ر	56138	115896	0	0	172034
ز	8623	12608	0	0	21231
س	5836	7233	75992	25487	114548
ش	330	1647	15469	13647	31093
ص	481	896	32115	13835	47327
ض	1562	1481	8791	6677	18511
ط	414	1170	4504	10245	16333
ظ	40	955	925	2599	4519
ع	1963	11170	136499	54625	204257
غ	217	483	5536	4608	10844
ف	2334	3655	76296	18397	100682
ق	4966	3497	64630	36482	109575
ك	3076	13896	29424	20548	66944
ل	68146	26147	242196	196946	533435
م	14831	62246	80246	80934	238257
ن	41669	130747	49601	103031	325048
ه	10781	122380	33486	37409	204056
و	111734	81885	0	0	193619
ؤ	792	2310	0	0	3102
ى	2870	62266	0	0	65136
ي	9800	72091	76211	120189	278291
ئ	184	184	6718	2852	9938
Total	789673	1274899	1274899	912939	4252410

Table 5 Alphabet shapes distribution in classic Arabic for *Muslim* book.

Let.	S-alone	Term.	Initial	Medial	Total
ء	4882	0	0	0	4882
ا	93283	130023	0	0	223306
إ	12574	3145	0	0	15719
أ	48367	9591	0	0	57958
آ	1292	672	0	0	1964
ب	6220	4792	72910	30187	114109
ة	7759	17666	0	0	25425
ت	2492	11262	10148	14771	38673
ث	1231	2747	26245	5017	35240
ج	1214	804	10618	5683	18319
ح	2434	1515	36727	16800	57476
خ	140	114	10513	2931	13698
د	8844	57376	0	0	66220
ذ	5625	7008	0	0	12633
ر	24844	55294	0	0	80138
ز	4648	5493	0	0	10141
س	2150	3312	36342	10328	52132
ش	148	801	7531	5976	14456
ص	244	372	13953	5710	20279
ض	664	556	2124	2976	6320
ط	137	441	1895	4349	6822
ظ	21	784	372	1064	2241
ع	784	5315	61392	23589	91080
غ	84	190	2415	1900	4589
ف	899	1264	31638	8346	42147
ق	2737	1366	28335	15427	47865
ك	1361	5757	13278	8805	29201
ل	30787	10964	102656	85995	230402
م	6196		34919	37982	105548
ن	17620	66300	22569	46892	153381
ه	4695	52532	15688	15670	88585
و	52251	37023	0	0	89274
ؤ	273	938	0	0	1211
ى	1199	25422	0	0	26621
ي	4442	33522	35974	56828	130766
ئ	91	90	2660	1148	3989
Total	352632	580902	580902	408374	1922810

Table 6. Alphabet shapes distribution in classic Arabic For *Al-Bukari* and *Muslim*.

Let.	S-alone	Term.	Initial	Medial	Total
ء	11896				11896
ا	213359	296148			509507
إ	29670	6321			35991
أ	103938	22656			126594
آ	3551	1490			5041
ب	15541	9922	140434	67938	233835
ة	17597	37078			54675
ت	6132	27033	29353	35826	98344
ث	2261	4368	49989	8749	65367
ج	2964	1496	23817	13394	41671
ح	3258	3388	69860	31432	107938
خ	243	264	22807	7089	30403
د	18950	114451			133401
ذ	13526	15441			28967
ر	56138	115896			172034
ز	8623	12608			21231
س	5836	7233	75992	25487	114548
ش	330	1647	15469	13647	31093
ص	481	896	32115	13835	47327
ض	1562	1481	8791	6677	18511
ط	414	1170	4504	10245	16333
ظ	40	955	925	2599	4519
ع	1963	11170	136499	54625	204257
غ	217	483	5536	4608	10844
ف	2334	3655	76296	18397	100682
ق	4966	3497	64630	36482	109575
ك	3076	13896	29424	20548	66944
ل	68146	26147	242196	196946	533435
م	14831	62246	80246	80934	238257
ن	41669	130747	49601	103031	325048
ه	10781	122380	33486	37409	204056
و	111734	81885			193619
ؤ	792	2310			3102
ى	2870	62266			65136
ي	9800	72091	76211	120189	278291
ئ	184	184	6718	2852	9938
Total	789673	1274899	1274899	912939	4252410

VI. Conclusion And Future Work

This paper presented minimal Arabic text that covers all of the Arabic letters with all possible shapes for each letter. We showed that the text is minimal and covers all the Arabic letter shapes. This presents one of the minimal solutions as this solution is not unique. The number of used letter shapes is the minimum possible to cover all the shapes of Arabic alphabet (i.e. 141 letters). However, different words may be generated as a minimal text depending on the algorithm of searching the corpora.

Any new analysis may produce new words in the new minimal text. However, the number of letters in the minimal text that covers all Arabic alphabet cannot be less than the size of our minimal text (i.e. 141). Using this text we can build Arabic handwritten or typewritten Arabic databases, for use in Arabic text recognition systems, with minimal effort.

In addition, we presented the frequencies of using the different shapes of Arabic alphabets in a data consisting of over 4 million characters. In our future work we will build databases using this script and the probabilities of each Arabic letter in practical applications using our previous work related to the probabilities of Arabic alphabet usage. As we intend to use machine learning algorithms to recognise Arabic alphabet in the future, we believe that including the probabilities of each Arabic letter shape as one of the input features for the training stage will prove to be valuable. This database is expected to address one of the problems of conducting Arabic text recognition research (i.e. the lack of suitable Arabic text databases) as reported by many researchers.

Acknowledgment

The first two authors wish to acknowledge KFUPM for support.

References

[1] Badr Al-Badr, Sabri A. Mahmoud, "Survey and Bibliography of Arabic Optical Text Recognition," *J. of Signal Processing*, Vol. 41, No.1, pp.49-77 (Jan. 1995).

[2] A.Amin,"Offline Arabic Character Recognition: The State of the Art," *Pattern Recognition*, vol.31, pp.517-530,1998.

[3] M. Khorsheed, "Off-line Arabic Character Recognition – A Review," *Pattern Analysis & Applications*, 5:31-45, 2002.

[4] L. Lorigo and V. Govindaraju, "Off-line Arabic Handwriting Recognition: A survey," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 712-724, 2006.

[5] T. Sari and M. Sellami "Cursive Arabic Script Segmentation and Recognition System," *International Journal of Computers and Applications*, vol. 27, no. 3, pp. 161-168, 2005.

[6] A.A. ABURAS and S.M. REHIEL "Off-line Omni-style Handwriting Arabic Character Recognition System Based on Wavelet Compression," *Arab Research Institute For Science & Engineering*, vol. 3, no. 4, pp. 123-135, 2007.

[7] Shirali-Shahreza, M. H. and Shirali-Shahreza, S. 2005. A robust page segmentation method for Persian/Arabic documents. In *Proceedings of the 5th WSEAS international Conference on Signal Processing, Computational Geometry & Artificial Vision* (Malta, September 15 - 17, 2005). B. Castagnolo, Ed. World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, 163-169.

[8] H. Bentouns and M. Batouche "Incremental support vector machines for handwritten Arabic character recognition," *Information and Communication Technologies: From Theory to Applications*, pp. 477-478, 2004.

[9] N. Ben Amor and N.E.B. Amara "Multifont Arabic Character Recognition Using Hough Transform and Hidden Markov Models," *4th International Symposium on Image and Signal Processing and Analysis (2005)*, pp. 285-288, 2005.

[10] Amor, N. B. and Amara, N. E. 2006. A hybrid approach for multifont Arabic characters recognition. In *Proceedings of the 5th WSEAS international Conference on Artificial intelligence, Knowledge Engineering and Data Bases* (Madrid, Spain,

[11] M. Shirali-Shahreza and S. Shirali-Shahreza "Persian/Arabic Text Font Estimation using Dots," *Sixth IEEE International Symposium on Signal Processing and Information Technology*, pp. 420-425, 2006.

[12] M. Khorsheed "Offline recognition of omnifont Arabic text using the HMM ToolKit (HTK)," *Pattern Recognition Letters*, vol. 28, no. 12, pp. 1563-1571, 2007.

[13] Al-Jarrah, O., Al-Kiswany, S., Al-Gharaibeh, B., Fraiwan, M., and Khasawneh, H. 2006. A new algorithm for Arabic optical character recognition. In *Proceedings of the 5th WSEAS international Conference on Artificial intelligence, Knowledge Engineering and Data Bases* (Madrid, Spain, February 15 - 17, 2006). P. L. Espi, J. M. Giron-Sierra, and A. S. Drigas, Eds. World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, 211-224.

[14] L. Lorigo and V. Govindaraju "Segmentation and Pre-Recognition of Arabic Handwriting," *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pp. 605-609, 2005.

[15] A. Broumandnia, M. Shanbehzadeh and M. Nourani "Segmentation of Printed Farsi/Arabic Words," *IEEE/ACS International Conference on Computer Systems and Applications, 2007. (AICCSA '07)*, pp. 761-766, 2007.

[16] M. Syiam, T.M. Nazmy, A.E. Fahmy, H. Fathi and K. Ali "Histogram clustering and hybrid classifier for handwritten Arabic characters recognition," *Proceedings of the 24th IASTED international conference on Signal processing, pattern recognition, and applications*, pp. 44-49, 2006.

[17] Y.S. Elarian "A Lexicon of Connected Components for Arabic Optical Text Recognition," *Computer Engineering*, -122, 2006.

[18] S. Touj, N.B. Amara and H. Amiri "Arabic Handwritten Words Recognition Based on a Planar Hidden Markov Model," *Int. Arab J. Inf. Technol*, vol. 2, no. 4, pp. 318-325, 2005.

[19] S. Mahmoud, "Recognition of writer-independent off-line handwritten Arabic (Indian) numerals using hidden Markov models," *Signal Processing*, vol. 88, no. 4, pp. 844-857, 2008.

[20] N. Farah, L. Souici and M. Sellami "Classifiers combination and syntax analysis for Arabic literal amount recognition," *Engineering Applications of Artificial Intelligence*, vol. 2006, no. 19, pp. 29-39, 2006.

[21] http://www.ifnenit.com/ "IFN/ENIT-database – Database Of Handwritten Arabic Words," 2006.

[22] M. Pechwitz and Maergner,Volker "HMM Based Approach for Handwritten Arabic Word Recognition Using the IFN/ENIT-Database," The Seventh *International Conference on Document Analysis and Recognition (ICDAR 2003)*, pp. 890-894, 2003.

[23] M. Pechwitz, S.S. Maddouri, V. Märgner, N. Ellouze and H. Amiri "Ifn/Enit - Database Of Handwritten Arabic Words," *CIFED*, 2002.

[24] V. Margner, M. Pechwitz and H. El Abed "ICDAR 2005 Arabic Handwriting Recognition Competition," *International Conference on Document Analysis and Recognition*, pp. 70-74, 2005.

[25] C. Higgins and D.G. Elliman "Off-line recognition of handwritten Arabic words using multiple hidden Markov models," *Knowledge-Based Systems*, vol. 17, no. 2-4, pp. 75-79, 2004.

[26] S. Al-Ma'adeed, D. Elliman and C. Higgins "A Database for Arabic Handwritten Text Recognition Research," *International Arab Journal on Information Technology*, vol. 1, no. 1, 2004.

[27] S. Alma'adeed, C. Higgens and D. Elliman "Off-line recognition of handwritten Arabic words using multiple hidden Markov models," *Knowledge-Based Systems*, vol. 2004, no. 17, pp. 75-79, 2004.

[28] S. Al-Ma'adeed, D. Elliman and C.A. Higgins "A Database for Arabic Handwritten Text Recognition Research," *The Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR'02)*, 2002.

[29] Y.A. Alotaibi "High Performance Arabic Digits Recognizer Using Neural Networks," *The International Joint Conference On Neural Networks 2003*, pp. 670-674, 2003.

[30] M. soleymani, F. Razzazi "An Efficient Front-End system for Isolated Persian/Arabic Character Recognition of Handwritten Data-Entry Forms," pp. 6 2003.

[31] Y. Al-Ohali, M. Cheriet and C. Suen "Databases for recognition of handwritten Arabic cheques," Pattern Recognition, vol. 2003, no. 36, pp. 111-121, 2003.

[32] Y. Al-Ohali, M. Cheriet and C. Suen "Databases For Recognition Of Handwritten Arabic Cheques," *Seventh International Workshop on Frontiers in Handwriting Recognition*, pp. 601-606, 2000.

[33] S.S. Maddouri, H. Amiri, A. Belaïd and C. Choisy "Combination of Local and Global Vision Modeling for Arabic Handwritten Words Recognition," *Eighth International Workshop on Frontiers in Handwriting Recognition* (IWFHR'02), pp. 128-135, 2002.

[34] S.S. Maddouri, H. Amiri, A. Belaid and C. Choisy "Combination of local and global vision modelling for arabic handwritten words recognition," *Frontiers in Handwriting Recognition*, pp. 128-135, 2002.

[35] V. Margner and M. Pechwitz "Synthetic Data for Arabic OCR System Development," *The 6th International Conference on Document Analysis and Recognition, ICDAR'01*, pp. 1159-1163, 2001.

[36] Hamid and R. Haraty "A Neuro-Heuristic Approach for Segmenting Handwritten Arabic Text," *ACS/IEEE International Conference on Computer Systems and Applications* (AICCSA'01), pp. 110-113, 2001.

[37] M. Dehghan, K. Faez, M. Ahmadi and M. Shridhar "Handwritten Farsi (Arabic) word recognition: a holistic approach using discrete HMM," *Pattern Recognition*, vol. 2001, no. 34, pp. 1057-1065, 2001.

[38] N. Kharma, M. Ahmed and R. Ward "A New Comprehensive Database of Hadritten Arabic Words, Numbers, and Signatures used for OCR Testing," *1999 IEEE Canadian Conference on Electrical and Computer Engineering*, pp. 766-768, 1999.

[39] N. Wanas, M.S. Kamel, G. Auda and F. Karray "Feature-based decision aggregation in modular neural network classifiers," *Pattern Recognition Letters*, vol. 1999, no. 20, pp. 1353-1359, 1999.

[40] J. Makhoul , R. Schwartz, C. Lapre and I. Bazzi "A Script-Independent Methodology For Optical Character Recognition," *Pattern Recognition*, vol. 31, no. 9, pp. 1285-1294, 1998.

[41] Bazzi, C. LaPre, J. Makhoul, C. Raphael and R. Schwartz "Omnifont and Unlimited-Vocabulary OCR for English and Arabic," *4th International Conference Document Analysis and Recognition (ICDAR '97)*, pp. 842-846, 1997.

[42] J. Trenkle, E. Erlandson, A. Gillies and S. Schlosser "Arabic Character Recognition",*Symposium on Document Image*, pp. 191-195, 1995.

[43] M. Melhi "Off-line Arabic Cursive Handwriting Recognition Using Artificial Neural Networks," Department of Cybernetics, Internet and Virtual Systems, 2001.

[44] Husni A. Al-Muhtaseb, Sabri Mahmoud, and Rami S. Qahwaji, "Statistical Analysis for the support of Arabic Text Recognition", International Symposium on Computer and Arabic Language, Riyadh, Saudi Arabia, November 2007 (in Arabic)

[45] Husni A. Al-Muhtaseb, Sabri Mahmoud, and Rami S. Qahwaji, "A Novel Minimal Arabic Script for Preparing Databases and Benchmarks for Arabic Text Recognition Research", *8th WSEAS International Conference on SIGNAL PROCESSING (SIP '09)*, Istanbul, Turkey, June 2009.
[46] Al-Fayrouzabadi, *"Al Qamoos Al Muheet*", A dictionary, Darul Kutubul Ilmiyyah, Beirut, 1996 (in Arabic).
[47] Ibn Zakariyya, Abi Al-Hussein Ahmad Bin Faris, "*Mu'jam Maqayees Al-Lughah*", Dar Al-Jeel, Beirut, 1999 (in Arabic).
[48] Al-Bukhari, Mohammad, "*Al-Jame' Al-Saheeh (Sahih Al-Bukhari)*", Dar Al-Jeel, Beirut, 2005 (in Arabic).
[49] Al-Naysabouri, Muslem, "*Al-Jame' Al-Saheeh (Sahih Muslim)*", Dar Al-Jeel, Beirut, 2006 (in Arabic).
[50] Al-Asfahani, Al-Raghib, "*Mu'jam Alfath Al-Qurann*", Darul Kutubul Ilmiyyah, Beirut, 1997 (in Arabic).
[51] http://www.muhaddith.org/

Husni A. Al-Muhtaseb was born in 1961. He received his BSc in electrical engineering, computer option from Yarmouk University, Irbid, Jordan in 1984 and his MSc in computer science and engineering from King Fahd University of Petroleum & Minerals (KFUPM), Dhahran, Saudi Arabia in 1988.

He is currently an Instructor with the Department of Information & Computer Science of KFUPM. He worked as a technical consultant for the dean of admissions and registration at KFUPM for 10 years. His research interests include software development, Arabic Computing, computer Arabization, Arabic OCR, e-learning & online tutoring and natural Arabic understanding.

Mr. Al-Muhtaseb has participated in several industrial projects and worked as a consultant with different institutes/ organizations. Mr. Al-Muhtaseb has more than 60 research publications.

Sabri A. Mahmoud received his B.S. in electrical engineering from Sind University, Pakistan in 1972, received his M.S. in Computer Sciences from Stevens Institute of Technology, U.S.A., in 1980 and his Ph.D. degree in Information Systems Engineering from the University of Bradford, U.K., in 1987.

He joined the Computer Engineering Department, College of Computers and Information Systems, King Saud University, from 1988 until 1995. He left KSU as an Associate Professor and worked in industry in software development and IT consultations from 1995 until 2003. He joined the Information Technology Department, Arab Open University from 2003 until 2005 as Staff Tutor. He joined the Information and Computer Science Department, King Fahd University of Petroleum and Minerals in September, 2005 as an Associate Professor. His research interests include Pattern Recognition, Arabic Document Analysis and Recognition, Image Analysis (including Time Varying Image Analysis and Computer Vision), and Arabic Computing.

Dr. Mahmoud is a senior member of IEEE. He published over 40 papers in refereed journals and conference proceedings.

Rami Qahwaji was born in 1972. He got his BSc in electrical engineering from the University of Mustansiriyah (Baghdad) in 1994 and his MSc in control and computer engineering from the same university in 1997. He got his PhD in computer vision systems in 2002 from Bradford University (UK).

He is currently a Reader in Visual Computing at the School of Computing, Informatics and Media at Bradford University (UK). He is the principal investigator for two 3-year EPSRC grants, total value of around 425K GBP. He is also involved in 3 small research grants. His publications include around 80 refereed papers and 6 PhD completions. He is the conference Co-Chair for the International Conference on Cyberworlds 2009 (UK) and the 1st International Conference on Computer Science from Algorithms to Applications (CSAA09) (Egypt) and had been a member of the technical committees for more than 14 international conferences.

Dr Qahwaji is a Charted Engineer (CEng– IET), a Fellow of the Higher Education academy (HEA) and a member of AGU, IET, IEEE, IEEE CS, IASTED, and others. Dr Qahwaji has refereed research proposals for the Royal Society, KFUPM, PPARC and the British Council and he is also a reviewer for the following Journals: Solar Physics, IET Microwaves, Antennas & Propagation, IET Image Processing, IET Radar, Sonar & Navigation, Pattern Recognition Letters, Neural Computing and Applications, Signal Processing and others.