BENTHAM OPEN

CrossMark

The Open Cybernetics & Systemics Journal

Content list available at: www.benthamopen.com/TOCSJ/

DOI: 10.2174/1874110X01812010042

RESEARCH ARTICLE

Online-KHATT: An Open-Vocabulary Database for Arabic Online-Text Processing

Sabri A. Mahmoud[1,*], Hamzah Luqman[1], Baligh M. Al-Helali[1], Galal BinMakhashen[1] and Mohammad Tanvir Parvez[2]

[1]*King Fahd University of Petroleum & Minerals, Dhahran31261, Saudi Arabia*
[2]*Qassim University, Qassim 51477, Saudi Arabia*

Received: October 30, 2017 Revised: February 20, 2018 Accepted: February 28, 2018

Abstract:

Background:

An Arabic online text database called Online-KHATT is presented, which addresses the lack of a free benchmarking database of natural Arabic online text. This database consists of natural Arabic online text written without any constraints using digital pen.

Objective:

The main objective of this work is to build a comprehensive benchmarking database of online Arabic text. Part of this objective is the development of tools, techniques and procedures for online text collection, verification and transliteration. Additionally, we built a dataset for segmented online Arabic characters and ligatures with ground truth labeling and present classification results of online Arabic characters using DBN-based HMM.

Method:

The source text of Online-KHATT is the same source text of the unique paragraphs of the KHATT database, along with additional resources to increase the coverage of the database. A 3-level verification procedure aligns the online text with its ground truth. The verified ground-truth database contains meta-data that describes the online Arabic text at the line level using text, InkML and XML formats.

Results:

The database consists of 10,040 lines of Arabic text written by 623 writers using Android- and Windows-based devices. The text lines of Online-KHATT database are randomly distributed into training, testing, and verification sets that contain 70%, 15% and 15% of the text lines of the database, respectively. We have segmented part of the collected data into characters along with their ground truths. We have developed tools for the collection of data (for devices with electronic pen), verification and correction of ground truths, transliteration, and semi-automated segmentation of characters. In addition, we also present the experimental results of Arabic online character recognition using the Online-KHATT database.

Conclusion:

Online-KHATT database can be used for Arabic online text recognition, writer identification and verification, pre-processing and segmentation, etc. In addition, researchers may use the segmented characters to test their segmentation algorithms for use in online text recognition or to train online text classifiers. This database will be made freely available for interested researchers at (http://onlinekhatt.ideas2serve.net/).

* Address correspondence to this author at the King Fahd University of Petroleum & Minerals, Sabri A. Mahmoud, Dhahran 31261, Saudi Arabia; Tel: 966554430980; E-mails: smasaad@kfupm.edu.sa , smasaad@gmail.com

Keywords: Arabic online text database, Arabic online text recognition, Segmentation, Handwriting recognition, Online character recognition, HMM.

1. INTRODUCTION

This study presents a comprehensive Arabic online text database called "Online-KFUPM Handwritten Arabic TexT (KHATT)" for possible use as a benchmarking database for Arabic online text recognition research. The database consists of natural Arabic online text that was written without any constraints. This database is a continuation of the work performed for the Arabic offline handwritten text database called "KHATT" [1]. The need for the Arabic online text database "Online-KHATT" is based on the lack of a freely available benchmarking natural Arabic online text database. During the data acquisition process of Online-KHATT, we have used the same input texts that were used to collect the KHATT database of offline Arabic text. In addition, we have used additional sources of text (described in [2]) to increase the coverage of Online-KHATT. The source text of the unique paragraphs of the KHATT database had spelling mistakes, digits, rarely used diacritics, non-Arabic characters and rare symbols. In order to make the online text ready for Arabic online text recognition and enable normal natural language processing, we enhanced the source texts before collecting the online data. We corrected spelling mistakes, replaced digits with their equivalent words, removed rarely used diacritics, removed non-Arabic characters, and rare symbols. We also ensured that the resulting sentences are complete in meaning.

The Online-KHATT database was written by 623 writers using Android- and Windows-based devices that support electronic pens. Because we have used the unique paragraphs from KHATT, each writer wrote unique paragraphs in Online-KHATT. We then added samples of the fixed paragraph (*i.e.*, a paragraph that contains all Arabic characters' shapes) of KHATT database to increase the frequency of less frequent characters. This approach was proven to be useful in improving recognition accuracy when used with training data [3].

We implemented a 3-level verification procedure to align the online text with its ground truth. The verified ground-truth database contains meta-data that describe the online text at the line level using text, InkML and XML formats. Errors that may occur during writing, such as missing/adding/touching dots and broken characters, were tracked and coded. We then generated a transliteration of the ground-truth data. A portion of the collected data was semi-automatically segmented into characters, and their ground truths were documented. These segmented characters allow researchers to test their own online text-segmentation algorithms or can be used as seed data for training online text classifiers. In addition, we present our experimental results on Online-KHATT database. We used the segmented characters to recognize Arabic online characters using selected features and classifiers.

In preparing Online-KHATT database, we had to address several challenges. Due to the lack of open source tools that could be used for collecting data from touch-based systems or for online text segmentation; we had to develop those tools in addition to other related tools. The limited number of devices resulted in making the online data collection process slower compared to offline data collection. Writers take time to write a form of Arabic online text and some of the writers required training. This resulted in less volunteers to participate in the data collection. In addition, during the verification phase, we found that some writers had poor handwriting, and some started by writing eligible text in the first few lines, and then wrote unreadable text, resulting in discarding of many collected forms. A number of errors in the collected text were common like missing or incorrectly written dots, broken letters, touching letters that should be separated, added symbols that are not in the source data, missing letters or symbols, substituted letters or symbols, and missing or added words. Addressing these issues took enough time from the research team.

The Online-KHATT database will be made freely available to researchers (http://onlinekhatt.ideas2serve.net/) for various Arabic online text-related applications, such as Arabic online text recognition, writer identification and verification, preprocessing, and segmentation. The main contributions of this paper thus can be summarized as follows:

- An open vocabulary benchmarking database of online Arabic text of over 10,000 lines of Arabic online text written by 623 writers.
- Development of tools, techniques and procedures for online text collection, verification and transliteration.
- A dataset for segmented online Arabic characters and ligatures with ground truth labeling.
- A technique for classification of Online Arabic characters using DBN-based HMM.

The remainder of the paper is organized as follows. Section 2 presents a literature review of Arabic online text databases. In Section 3, we present the data collection and pre-processing steps used in this study. Section 4 details the

data verification process and Section 5 describes the dataset of segmented characters. We then discuss the transliteration process of Online-KHATT in Section 6 and present the experimental results of the Arabic online character recognition algorithm using the Online-KHATT database in Section 7. Lastly, we present the conclusions of this study in Section 8.

2. RELATED WORK

In this section, we discuss previous studies related to Arabic online text databases. Several researchers have built Arabic online text databases at the text, word, character, and digit levels. Most of these databases are at the level of words, characters, and digits; only a few of the databases are at the Arabic online text level. In general, the reported databases have limitations regarding the text level, lexicon coverage, data size, and number of writers. In the following paragraphs, we mainly consider the databases that were introduced at the word and sentence levels and summarize those that were introduced at the character and digit levels.

2.1. Databases of Words and Sentences

Several studies have presented datasets of Arabic online words that were carefully selected to meet certain research goals. In [4], a set of 800 Arabic words were selected such that the different shapes of Arabic letters were included with a nearly uniform distribution. Four writers wrote samples for this set of 800 words, while six additional writers wrote samples for 280 other words. Certain rules constraining the number, order, and location of the input strokes were followed. This dataset was used in [5]. In [6], samples of Arabic words and isolated characters were collected from 40 writers. The dataset contains 1,578 samples of a list of 66 Arabic words selected to span all Arabic letter shapes. Half of the word samples (*i.e.*, 839 words) and 27 of the writers of the isolated single characters were used for training; the remaining samples were used for testing. LMCA ("Lettres, Mots et Chiffres Arabe" in French) is an On/Off dual Arabic handwriting database presented in [7] that consists of on- and off-line samples of 500 Arabic words, 100,000 Arabic letters and 30,000 digits collected from 55 writers. The digit samples of the LMCA were used in [8], while the words subset was used in [9]. In [10], a database of online samples of Qur'anic Handwritten Words (QHW) was presented. The database was collected from 200 writers with a total of 12000 samples of the 120 most common words in the noble Qur'an.

The most popular database used for Arabic offline handwritten text recognition is the IfN/ENIT dataset [11], which contains samples of 937 Tunisian city names. The institution that introduced the IfN/ENIT dataset (Institut für Nachrichtentechnik (IfN) in cooperation with the Research group on intelligent Machines (REGIM)) presented an online dataset for the same vocabulary called ADAB (Arabic Database). This database contains 33,164 Arabic words (174,690 characters) that were written by 166 different writers. ADAB database has been used in handwriting-recognition competitions [12 - 13] to evaluate the participating systems. Thus, ADAB is probably the most widely used database to date in Arabic online text recognition research [14 - 27]. However, that database suffers from the same limitations of IfN/ENIT (*i.e.*, limited vocabulary, use of city names, and not being constructed with natural Arabic text).

Another database developed in the REGIM laboratory is called the "MAYASTROUN-database" [28]. This database is multilingual and is used for both on- and off-line, unconstrained, handwritten, cursive Latin and Arabic texts, words, characters, digits, signatures and mathematical expressions. The final version of this database contains 6,500 digits, 5,600 characters and 1,500 words written by 355 writers. The Arabic text lexicon is limited in this dataset.

In [29], an Online Handwritten Arabic Sentence Database (OHASD) was presented. Samples of paragraphs with complete sentences ranging from 15 to 46 words were collected from different writers. The collected data were filtered by excluding erratic/illegible handwritings, resulting in 154 paragraphs written by 48 writers containing 3,825 words and 19,467 characters. Although this database contains higher-level text (*i.e.*, sentences), it is limited in terms of lexicon, collected data size, and the number of writers. This database was used in [30, 31].

The Large-Vocabulary Arabic Online Handwriting database (AltecOnDB) was built by the Arabic Language Technology Center (ALTEC) [32]. It was collected from approximately 1000 writers with 152,680 samples of 39,945 different words, including 325,477 samples of 14,740 PAWs (Parts of Arabic Words). The database contains samples of characters, digits and punctuation marks and has a subset called Set-H that is more suitable for writer-dependent research. This database was collected using a device that captures handwritten ink on ordinary paper. Natural writing on paper is much more controllable than online writing on touch screen devices, and the variability of touch-screen-pen-based writing is much greater than of writing on paper. This database is not freely available and was used in [25]. Systems built using this database may have lower accuracies when used with touch-screen-pen-based devices.

Due to the difficulty of constructing comprehensive online databases, methods for generating synthetic databases of Arabic online script were presented in [27], [33]. These methods are used to synthesize large sets of shapes for each part of a word in a given lexicon. This lexicon contains the desired words and a set of handwriting prototypes that could be extracted automatically from a given small dataset of word shapes or generated manually by human writers. This approach was used in [5], [27], [33], [34] to generate synthetic databases. The datasets generated using such methods are not like natural online writing. Also, the errors during the generation of synthetic data may affect the subsequent phases in recognition. It is expected that systems trained with such data will have lower accuracies when used with real online devices.

2.2. Databases of Characters and Digits

Several databases were introduced at Arabic online character level. A database of the basic shapes of isolated Arabic characters was introduced in [35]. Approximately, 7,400 character samples were collected from 17 writers, where each writer wrote 24 samples for each character with no constraints on their handwriting style, leading to a wide variety of sizes and orientations. This dataset was also used in [36 - 40] and was expanded in [41] by adding characters written by five more writers. A total of 28 letters representing the isolated forms of the standard Arabic alphabet were considered with their diacritical marks in a few studies. In [42], Arabic letters were written five times by 10 writers, producing a dataset of 1400 samples; this dataset was used in [43]. Some other Arabic online character databases are described in [44 - 49]. Refer to [50 - 53] for databases of Arabic online digits.

Table **1** summarizes the existing Arabic online text databases. As shown in Table **1**, we are not aware of any comprehensive and open-vocabulary Arabic online text database of adequate size that reflects the naturalness of Arabic text. Thus, the Online-KHATT database is built to address these needs. It contains 10,040 lines of Arabic online text written by 623 writers using Android- and Windows-based devices. The Online-KHATT database will be made freely available to interested researchers (http://onlinekhatt.ideas2serve.net/).

Table 1. Statistics of some reported Arabic online text databases.

Dataset	Year	Digits	Characters	Words	Writers
LMCA [7]	2008	30,000	100,000	500	55
OHASD [29]	2010	-	19,467	3,825	48
ADAB [12]	2011		174,690	33,164	166
AOD [52]	2012	30,000	-	-	100
MAYASTROUN [28]	2012	6,500	5,600	1,500	355
ALTECOnDb [32]	2014	-	106,433	152,680	1000
QHW [10]	2014	-	42,800	12,000	200

3. DATA COLLECTION AND PREPROCESSING

This section describes several tasks, including the selection of data sources and data preparation and the collection and pre-verification phases of Arabic online text.

The Online-KHATT database uses the same raw data of unique paragraphs (viz. paragraphs two and three) of the KHATT database as the input text for data collection, along with some additional sources of text. The source texts of the KHATT database were taken from 40 books [1]. For ease of reference, Table **2** summarizes the sources of the collected KHATT data. The additional resources of the Online-KHATT database were taken from [2] that contains articles from Arabic journals. We added these new source-texts to improve the coverage of the Online-KHATT database. We used the unique paragraphs of the KHATT database to generate the forms of Online-KHATT. We also added samples of the fixed paragraphs of KHATT database, which contains all Arabic characters in all of their possible shapes, to increase the frequency of less frequent characters.

Table 2. Statistics of the sources of text used in KHATT [1].

Category	References	Paragraphs	Pages
Art	3	399	182
Economy	4	78	59
Education	1	46	4
Health	3	105	72

(Table 2) contd.....

Category	References	Paragraphs	Pages
Literature	5	635	319
Management	5	86	138
Nature	7	136	70
Social	5	130	720
Technology	4	189	103
World	3	22	14
Total	**40**	**1826**	**1788**

The online text of paragraphs is split into lines and each line is saved in a separate file with proper naming. Hence, Online-KHATT consists of online text lines as a basic unit (A subset is available at the character level). At the post-processing phase, the classified text of several lines may be combined into paragraphs for Natural Language Processing (NLP) processing to improve the classification accuracy.

3.1. Data Preparation

In this phase, we improved the KHATT source data in three stages of verification to ensure the consistency, correctness and suitability of the source data. In each stage, a reviewer checked the source data and removed undesired content (*e.g.* , general objectionable content, less frequent diacritical marks like Shaddah (ـّـ) and Dhammah (ـُـ), and non-Arabic characters along with special symbols). Reviewers also converted the numbers found in the raw source data into corresponding words to preserve the natural meaning of the text. The raw text sources were then stored in both text and spreadsheet formats.

To prepare the paragraphs of text for Arabic online text collection, we limited the number of words in a line of text to eight words to ensure a comfortable writing experience on tablets. Then, lines of text were grouped into paragraphs. Each paragraph contained 12 lines of text on average, and each user-written form had two paragraphs. Then, we organized the data into properly named folders. The naming scheme used for the lines of text was similar to that of the KHATT database ('*Axxxx_PrgphNum_LineNum.txt*', where *Axxxx* is the form number, *PrgphNum* is the paragraph number, and *LineNum* is the line sequence number).

3.2. Arabic Online Text Collection

We used eight devices with various screen sizes during the data collection process: seven Android-based systems (Samsung SM-P601 and SM-P605) and one Windows-based system (a tablet from Sony). This variety of devices helped provide different effects on the handwriting styles of the writers. Naturally, a volunteer writes better on larger tablet screens than smaller ones. In addition, a faster processing unit within a device could make a volunteer feel more comfortable while writing (*e.g.* , no time lag while writing).

We have developed both Windows- and Android-based versions of the data collection tool. Fig. (**1**) shows the interfaces of the software (Android version) for data collection. The software consists of writer registration and data entry interfaces. In the registration phase, a writer enters his/her information (*e.g.* , name, age, country of origin, gender, and handedness), as shown in Fig. (**1a**). After entering the writer's information, the system displays the data entry interface, as shown in Fig. (**1b-c**). An example of a paragraph of Arabic text written using the software is shown in Fig. (**1b**). The data entry interface allows entering, editing or modifying the written text.

The Windows-based version of the software has a layout that is similar to the Android-based version. To familiarize the data collectors with the process, we conducted an initial data collection process. On average, the writing of each form took approximately 15 minutes.

The collected data is saved in the Ink Markup Language (InkML) format (http://www.w3.org/TR/InkML/), while the ground truth is saved in the form of text files. The InkML data format from the World Wide Web Consortium (W3C) is used to represent data written using an electronic pen and supports the exchange of handwritten data between ink-aware modules, such as signature verifiers, handwriting and gesture recognizers. The naming scheme for an InkML file has the same naming format of the source-text files described earlier, except that an '*.inkml'* extension is used.

Fig. (**2**) shows the content of an InkML file. As shown in Fig. (**2**), all content appears within an <ink> element. The <trace> element is the basic data element of a InkML document and represents a sequence of contiguous ink points. Each data point is represented by four values (*e.g.* , *X-coordinate, Y-coordinate, Pressure, Time*), where *X* and *Y* are the coordinate location of the data point, *Pressure* describes the pen tilt and pen tip force, and *Time* records the timing

information of each data point in the trace. The writer information is included in the <Annotation> element, where detailed writer information can be found.

Fig. (1). Interfaces of the data collection tool: (a) user information screen, and (b-c) data entry screens at the paragraph and line levels, respectively.

```
<?xml version="1.0" encoding="ASCII"?>
<ink xmlns:inkml="http://www.w3.org/2003/InkML">
  <context inkSourceRef="Tablet PC" />
  <definitions>
    <traceFormat>
      <channel name="X" type="integer" />
      <channel name="Y" type="integer" />
      <channel name="Pressure" type="integer" />
      <channel name="TimerTick" type="integer" />
    </traceFormat>
  </definitions>
  <annotation type="WriterName">maha</annotation>
  <annotation type="WriterAge">18</annotation>
  <annotation type="WriterCountry">Yemen</annotation>
  <annotation type="WriterQualification">Secondary</annotation>
  <annotation type="WriterGender">Female</annotation>
  <annotation type="WriterHandedness">Right</annotation>
  <trace type="penDown" duration="0" id="id001">
    834 147 49 0,834 145 81 16,835 145 114 16,836 144 145 31
  </trace>
  <trace type="penDown" duration="0" id="id002">
    861 142 19 0,860 141 49 15,859 141 81 15,859 141 112 31
    </trace>
</ink>
```

Fig. (2). Example of online handwritten strokes as stored in InkML format.

In the Online-KHATT database, we have collected 881 verified forms written by 623 writers. Humans without context rejected other collected forms because they were not readable. Certain writers wrote more than one form to have more text from the same writers for classifiers that require more data for writer adaptation, writer identification, etc. Online-KHATT database incorporates writers with different age groups that can be divided generally into two which are less than or equal to 25 and greater than 25. Most of the writers are young people with higher education forming 77% of the database. In fact, nowadays most smart devices' users are young people. Also, it is noted that right-handed writers are forming 98% of the writers. Other Online-KHATT database statistics are summarized in Table **3**.

Table 3. Online-KHATT database writers' summary.

Category	Class	Percentage of Writers	Number
Gender	Male	79%	495
	Female	21%	128
Age	Age ≤ 25	53%	333
	Age > 25	47%	290

(Table 3) contd.....

Category	Class	Percentage of Writers	Number
Qualification	Elementary	7%	43
	High School	16%	100
	University	77%	480
Handedness	Right hand	98%	612
	Left hand	2%	11
Nationality	Yemeni	88%	547
	Saudi	10%	62
	Arabian	2%	14

Table **4** lists the word, character and line counts and the word uni-, bi-, and tri-grams of the training, validation and testing sets and of the full database. Table **5** shows the Out-Of-Vocabulary (OOV) statistics of the validation and test datasets compared to the training dataset in Online-KHATT.

Table 4. Character, word, line counts and word N-grams of Online-KHATT database.

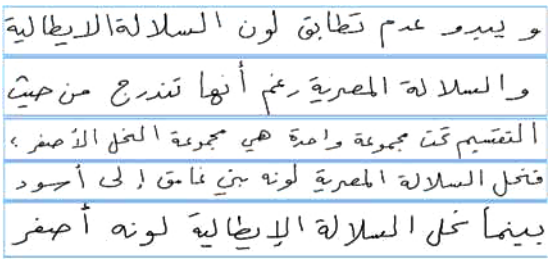
Set	Word counts	Character counts	Line counts	Unigrams	Bigrams	Trigrams
Training	56547	561754	6974	19437	45181	52117
Validation	12100	119386	1533	5887	10580	11451
Testing	12284	120281	1533	5641	10263	11098
Entire Database	**80931**	**801421**	**10040**	**30965**	**66024**	**74666**

Table 5. Out-of-vocabulary statistics of the Online-KHATT database.

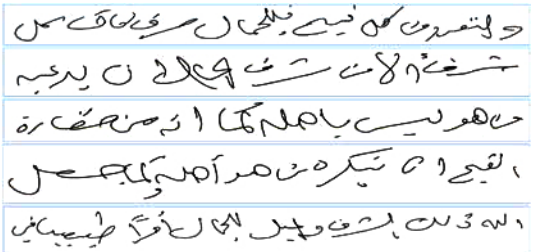
Out of Vocabulary	Tokens	Percentage
Validation	10594	43.45
Testing	10778	44.2

4. VERIFICATION OF ONLINE TEXT

Expert reviewers checked the collected unconstrained Arabic online text for writing quality. The objective of this phase was to identify text lines that were unreadable by humans without context and then divide the readable text lines into training, testing, and validation sets. Fig. (**3a**) shows samples of the accepted (*i.e.*, readable) lines that were written in unconstrained mode, and Fig. (**3b**) shows text lines that were rejected because the text was unreadable by humans without context.



(a)



(b)

Fig. (3). Samples of text from the Online-KHATT database: (a) accepted text and (b) rejected text.

We reviewed each form using a three-step process, in which different reviewers handled each step. The reviewers checked each form to verify whether the written online text matched the ground truth. Because a writer might have deviated from the text that he was supposed to write, the task of the reviewers was to modify the ground truth to match the written online text. Fig. (**4**) illustrates the flowchart of the verification phase for the Online-KHATT database. The same process illustrated in Fig. (**4**) is carried out by three reviewers independently to ensure correctness and consistency.

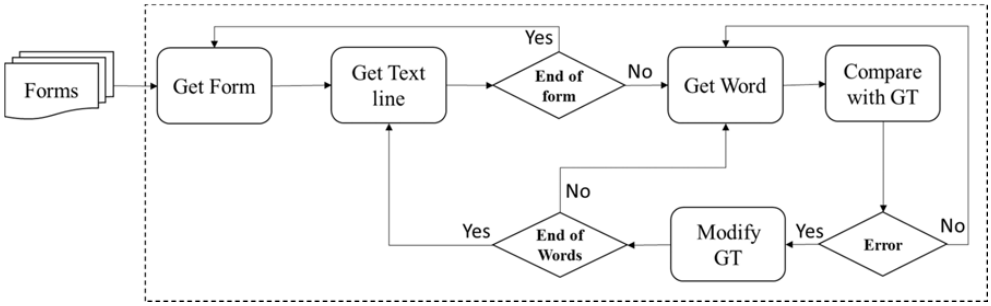
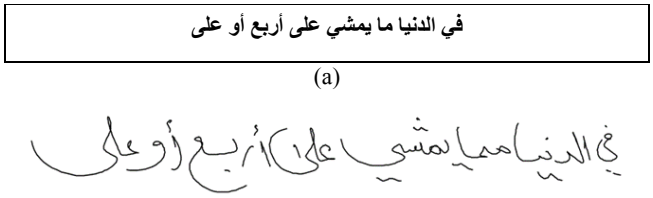


Fig. (4). General verification flowchart for the Online-KHATT database.

When there is a difference between a word in the ground truth and the corresponding word in the online text, the word in the ground truth is changed to match the word in the online text. Fig. (**5a**) shows the ground truth of a line of Arabic text, and Fig. (**5b**) shows the corresponding online text. As shown in Fig. (**5**), the word "ما" in the ground truth was written by the writer as "مما". If there is a word/letter in the ground truth that was missed by the writer, that word/letter is deleted from the ground truth. Similarly, insertion of a new word/letter by the writer will result in adding that word/letter to the ground truth. A snapshot of the verification tool is given in Fig. (**6**).



(a)

(b)

Fig. (5). Ground-truth matching in Online-KHATT database: (a) a sample of original ground truth and (b) image of the written text.

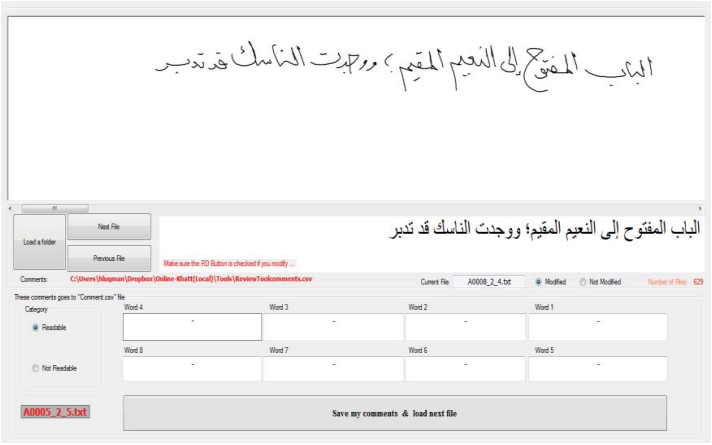


Fig. (6). Interface of the verification tool developed for Online-KHATT database.

Through the verification process, we found that the primary errors in the written Arabic online text include: missing dots or incorrectly written dots, broken letters, touching letters that should be separated, added symbols that are not in the source data, missing letters or symbols, substituted letters or symbols, and missing or added words. The positions of these errors in the written text were tracked and coded in separate files that were attached to each online text line. This will be discussed in more detail in Section **6**.

5. DATASET OF SEGMENTED CHARACTERS

The Online-KHATT database is available as Arabic online lines of text. Subsets of the lines in the Online-KHATT are segmented into characters. Some lines' were fully segmented into characters. However, some lines are not fully segmented into characters because some of their words are either unreadable or difficult to correctly segment.

The dataset of segmented Arabic characters may be used for Arabic online character recognition or for modeling Arabic characters for text recognition [54]. These segmented characters may also be used as seed characters to train classifiers or test online text-segmentation algorithms.

The segmentation process used in this study had two phases: segmentation and verification. The first phase is called semi-automatic segmentation, as the characters are extracted using a tool that needs human assistance. The tool specifies the proposed segmentation points of the text line. If the segmentation points are correct the user accepts them, and the tool saves the segmented characters. However, if some segmentation points are not proper then the user can modify them using a pen and then accept them and the tool saves them. Hence, the term semi-automatic segmentation. In this phase, began with semi-automatic segmentation of the words into characters or a set of characters (called ligatures). During this segmentation phase, certain information regarding each segmented character is recorded, including its position within the PAW (*i.e.*, initial, middle, end or isolated), the sequence number of the character in the line and the transliteration of the character. Details of this transliteration step are addressed in Section 6. Fig. (**7**) shows the segmentation tool that was developed for this purpose. Fig. (**8**) shows a handwritten line of Arabic text with some of its words being segmented into characters. Fig. (**9**) shows samples of overlapped characters that were segmented into ligatures due to difficulties in segmenting such characters.

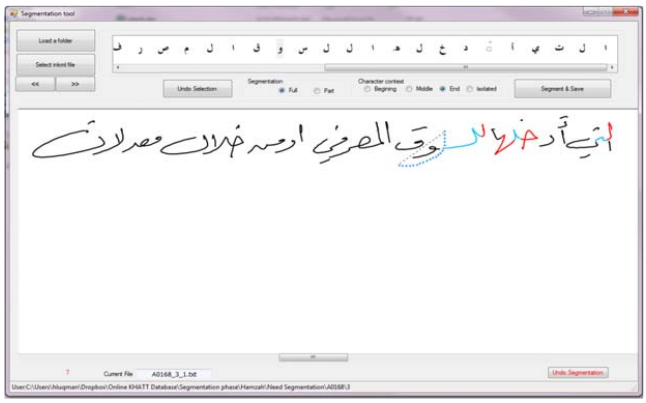


Fig. (7). Interface of the segmentation tool developed for semi-automatic segmentation of Arabic online text.

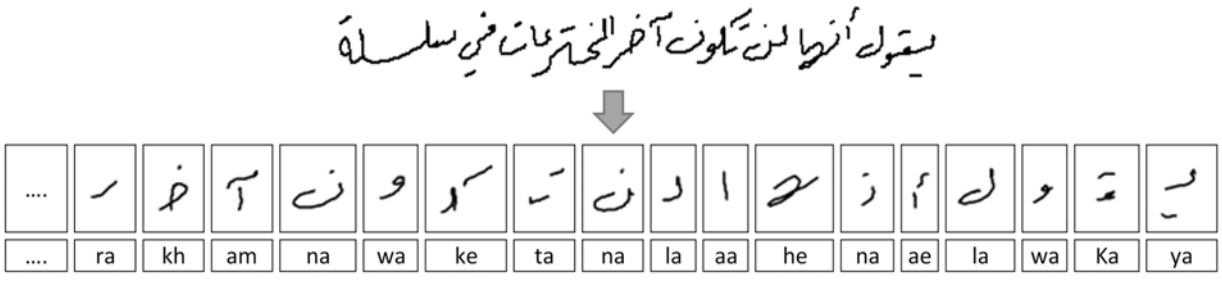
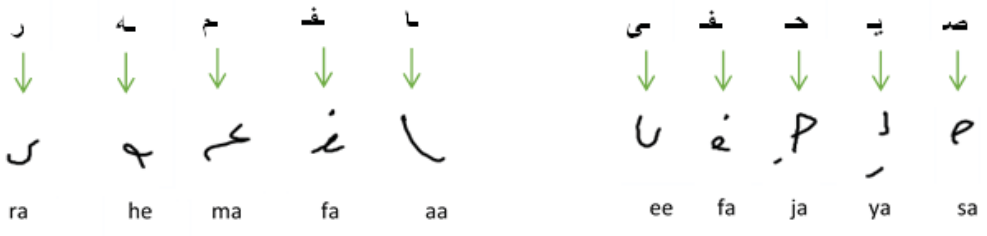


Fig. (8). Illustration of a handwritten line of Arabic text with segmentation of certain words along with their Latin transliterations.

Fig. (9). Illustration of certain ligatures segmented from the Online-KHATT database.

After segmenting the Arabic online text lines into characters, the verification phase began with the ground truth of each character in the Arabic online text being matched that of the corresponding segmented character. The readability and the stroke boundary of each segmented character were evaluated to discard unreadable and incorrectly segmented characters. Fig. (**10a**) shows samples of unreadable segmented characters, while Fig. (**10b**) shows samples of incorrectly segmented characters. Each of the segmented characters that pass the verification phase is stored in InkML format. The file name for each segmented character contains the information recorded for the character during the segmentation process. The transliteration code of a ligature is formed by concatenating the transliteration codes of the constituent characters within that ligature.



(a) (b)

Fig. (10). Examples of segmented Arabic characters from the Online-KHATT database: (a) characters that are unreadable, and (b) characters that have segmentation errors.

Table **6** shows the statistics of the segmented characters in the Online-KHATT database. As shown in the table, only a minor fraction of the total segmented characters is discarded due to readability and segmentation errors. Fig. (**11**) illustrates the distribution of the segmented character-samples. Note that in (Fig. **11**), the letters ا (Alif) and ل (Lam) have higher a number of samples compared with other Arabic characters because these characters occur more frequently.

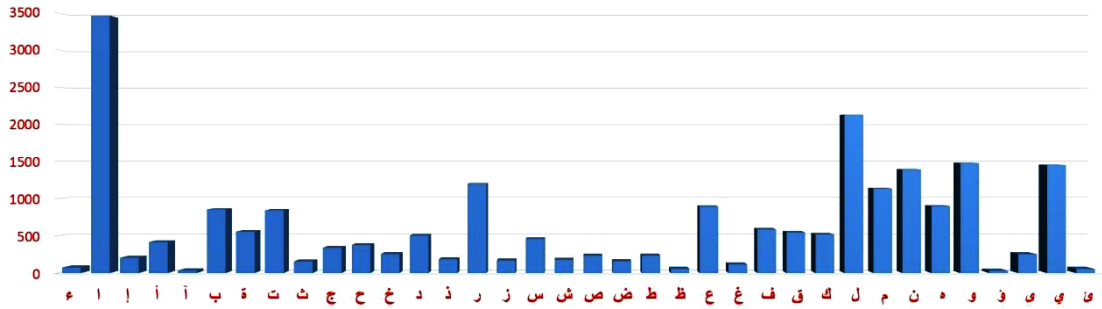


Fig. (11). Distribution of the number of samples per character in the segmented Arabic characters dataset.

Table 6. Statistics of the segmented characters in the Online-KHATT database.

Number of segmented lines	767
Number of segmented characters	31,747
Discarded characters	2995
Number of ligatures	739

6. TRANSLITERATION OF ARABIC ONLINE TEXT

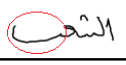
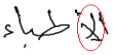
In this section, we describe the transliteration process of online Arabic text using Latin symbols for the Online-KHATT database. For this process, we have used the transliteration codes from the benchmarking KHATT database. However, we have modified the transliteration coding of the KHATT database to accommodate the transliteration of Arabic diacritics. During the transliteration process, we represent each character using a maximum of three Latin letters. We have extended the KHATT transliteration of an Arabic character by one more Latin symbol to accommodate any diacritics in the character. For example, if the Arabic letter ر has a diacritic such as the Arabic fatha, (ـَ), the transliteration of the letter will be "ra" concatenated with "f" (*i.e.*, "raf"). The transliteration codes that are used for the Arabic characters are shown in Table **7**.

Handwritten text may contain writing errors (*e.g.* , touching characters, missing dots) that can affect text-recognition performance. Keeping track of such errors can help in the analyses of the recognition output. In the Online-KHATT database, we identified various writing errors and assigned a code to each type of error, as shown in Table **8**. During the transliteration process of a line of text, if an error was found in a written character, the corresponding error code from Table **8** was inserted after the character code from Table **7**. Table **8** lists all the writing errors that were tracked in the transliteration process along with their codes, descriptions and examples from the database.

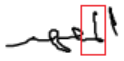
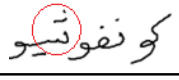
Table 7. Transliteration codes for the Arabic characters and other symbols as used in Online-KHATT database.

Arabic Character	ء	آ	أ	إ	ا	ب	ت	ة	ث	ج	ح	خ	د
Transliteration	hh	am	ae	ah	aa	ba	ta	tee	th	ja	ha	kh	da
Arabic Character	ذ	ر	ز	س	ش	ص	ض	ط	ظ	ع	غ	ف	ق
Transliteration	dh	ra	za	se	sh	sa	de	to	zha	ay	gh	fa	ka
Arabic Character	ك	ل	م	ن	ه	و	ؤ	ي	ى	ئ			
Transliteration	ke	la	ma	na	he	wa	wl	ya	ee	al			
Symbol	0	1	2	3	4	5	6	7	8	9	@	:	"
Transliteration	n0	n1	n2	n3	n4	n5	n6	n7	n8	n9	atr	col	dbq
Symbol	,	؛	،	?	!	.	(	)	/	\	=	-	_
Transliteration	com	com	com	qts	exc	dot	bro	brc	fsl	bsl	equ	hyp	usc
Symbol	#	%	Blank space										
Transliteration	Scr	Per	Sp										
Diacritics	◌	◌	◌	◌	~	◌	◌	◌					
Transliteration	D	H	K	F	X	B	N	Z					

Table 8. Different types of errors made by the writers of Arabic online text along with their codes for transliteration and examples.

Error	**Code**	**Description**	**Example**
Missing dots	#1-*n*#	*n* dots of the character dots are missing.	
Added Dots	#1+*n*#	*n* dots are incorrectly added to the character.	
Broken letter	#2*n*#	The character is broken into *n* segments.	
Disconnected letter	#3#	The character is disconnected from the previous character in the same PAW.	
Touching letters	#4#	A character is touching the character that follows it.	

(Table 8) contd.....

Error	Code	Description	Example
Touching diacritics/dots	#5-*n*-type#	*n* character dots are touching certain aspects of the text (*e.g.* , dots, diacritics, character). This error has three types: 1. Character: *n* character dots are touching the following character. This error has the error code #5-*n*-1# 2. Diacritics: *n* character dots are touching the diacritics of the following character. This error has the error code #5-*n*-2# 3. Dots: *n* character dots are touching the dots of the following character. This error has the error code #5-*n*-3#	
Ligature character	#6-*n*#	The next *n* characters are ligature.	
Missed Notch	#7-*n*#	The character is written with *n* missing notch(s).	

7. CLASSIFICATION OF ARABIC ONLINE CHARACTERS

As a pilot recognition task, we built a character recognition system that recognizes the primary part (*i.e.*, stroke) of the online Arabic characters (*i.e.*, not including delayed strokes). Those characters have 28 different basic shapes. The segmented data described in Section 5 was used for this task. To make the data more useful for certain recognition tasks (*e.g.* , managing delayed strokes), a semi-automatic annotation was performed at the stroke level. The annotated dataset contains 19,780 samples of 108 characters. Because the collected text is natural text, the samples of the characters are not uniformly distributed, and certain characters occur less frequently. Certain samples were also discarded due to errors in the segmentation process. Combining characters with similar primary strokes resulted in high counts for certain models; and 100 samples for training and 30 samples for testing for other models. To normalize the results, we ran the experiments with 100 random samples for the training set and 30 samples from the testing set, and averaged the results of each class.

In the preprocessing phase, a number of operations were performed. A simplification process was performed by eliminating duplicate successive points within a stroke. A weighted-average smoothing filter was used to reduce noise and eliminate hardware imperfections and variations in the input handwriting signal caused by the acquisition devices. Due to variations in writing speed, the acquired points were not distributed evenly along the stroke trajectory. Interpolation and re-sampling operations were used to recover missing data or to force points to lie at uniform distances. The linear interpolation introduced in [55] was used for this purpose.

For classification, we used the DBN-based HMM classifier *via* the Bnet MATLAB tool version 7, which was written by Kevin Murphy and acquired from the following link: https://code.google.com/p/bnt. An HMM is a stochastic finite automaton that is denoted in the first order by λ and defined by the triple (π, A, B), where π is the vector of the initial state probabilities, A the state transition matrix, and B the observation probability distribution:

- $\pi(i) = \{\pi_i | \pi_i = P(S_1 = i)\}$.
- $A = \{a_{ij} | a_{ij} = P(S_t = j | S_{t-1} = i)\}$.
- $B = \{ b_j(o_k) \mid b_j(o_k) = P(O_t = o_k | S_t = j)\}$.

Bayesian Networks (BNs) or belief networks, also known as Probabilistic Networks (PNs) are representations of domains involving uncertain relations among a group of random variables. The extension of Bayes nets can be done using the Dynamic Bayesian Network (DBN) [57] to model semi-infinite collections of random variables, Z_1, Z_2,…. The partition of the variables are $Z_t = (U_t, X_t, Y_t)$ representing the input, hidden and output variables. A DBN is a pair $(B_1, B_{\rightarrow})$, where B_1 is a BN which defines the prior $P(Z_1)$, and $B_{\rightarrow}$ is a two-slice temporal Bayes net (2TBN) which defines $P(Z_t | Z_{t-1})$ by means of a DAG (directed acyclic graph). An HMM can be represented as an instance of a DBN unrolled for 3 slices. Reference may be made to [57] for more details on the used classification technique.

In this study, we experimented with several types of features and certain variations of the classifier parameters, including the number of HMM states and the number of Gaussian mixtures. Several experiments were performed using the average of the local writing direction, which was represented by the cosine and sine trigonometric functions over a sliding window with a length of 15 points with an overlap of 5 points. We used 11 states and 16 Gaussian mixtures for all classes. The average recognition rate was 54% and the top-3 accuracy was 83%. Table **9** shows the confusion matrix,

where '*char'* is the character class and 'r%' is the recognition rate percent. For the sake of readability, the confusion value is highlighted with a red color scale starting by white for zero values (*i.e.* no confusion) and goes up with higher scales of red color for higher values (more confusion). We can notice that highest highlighted values are on or close to the diagonal. This is expected as it comes from recognizing the character correctly (on the diagonal) or confusing it with similar shapes (close to the diagonal) as can be seen for the characters Ha-B, and Ha-M.

Table 9. Confusion matrix for classification of the basic shapes of Arabic characters.

Id	Char	Label	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	r%
1		Aa-I	22	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	1	0	0	0	0	0	0	0	0	73%
2		Aa-E	0	16	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	6	3	0	0	1	0	0	0	0	0	53%
3		Ba-I	0	0	19	0	0	3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	5	0	0	0	0	0	2	63%
4		Ba-B	0	0	6	5	2	1	0	0	3	1	1	3	1	1	0	1	1	0	0	0	0	2	0	1	0	0	0	1	17%
5		Ba-M	0	0	6	2	10	3	0	0	0	2	0	0	3	0	0	0	1	1	0	0	0	0	0	1	0	0	0	1	33%
6		Ba-E	0	0	6	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	3	0	0	0	0	0	1	53%
7		Ha-B	0	0	0	0	0	0	15	7	0	0	0	0	0	0	2	0	2	0	0	0	1	0	2	0	0	0	1	0	50%
8		Ha-M	0	0	0	0	0	1	8	11	5	0	0	0	0	2	0	0	0	0	0	0	0	0	0	2	0	1	0	0	37%
9		Da-I	0	0	1	0	0	0	2	3	19	0	0	0	0	0	0	0	0	0	0	0	1	2	0	2	0	0	0	0	63%
10		Da-E	0	0	1	0	6	3	0	0	0	12	0	2	0	2	0	0	0	1	0	1	0	2	0	0	0	0	0	0	40%
11		Ra-I	1	0	0	0	1	0	0	0	1	1	24	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	80%
12		Ra-E	0	0	2	0	0	2	0	0	1	1	9	8	0	1	0	0	0	2	0	2	0	1	0	0	0	0	0	1	27%
13		Se-M	0	0	4	1	1	2	0	0	0	2	0	0	17	1	0	0	0	0	0	0	2	0	0	0	0	0	0	0	57%
14		Ay-B	0	0	0	0	0	0	0	1	1	0	0	0	0	25	0	0	0	0	0	0	0	1	0	0	0	0	0	2	83%
15		Ay-M	0	0	0	0	0	0	0	2	0	0	0	0	0	0	13	2	6	0	0	0	0	0	0	2	0	0	5	0	43%
16		Fa-B	0	0	0	0	1	0	0	1	0	0	0	0	0	0	3	13	2	0	0	0	0	0	0	0	0	7	3	0	43%
17		Fa-M	0	0	0	0	0	0	0	0	0	1	0	0	0	0	6	1	15	0	0	0	0	0	0	2	0	0	5	0	50%
18		La-B	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	18	4	4	0	1	0	0	0	0	0	1	60%
19		La-M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	23	6	0	0	0	0	0	0	0	0	77%
20		La-E	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	5	24	0	0	0	0	0	0	0	0	80%
21		Ma-B	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	2	1	0	0	18	0	0	0	0	4	0	2	60%
22		Na-I	0	0	11	0	0	5	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	8	0	3	0	0	0	2	27%
23		He-B	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	7	0	0	0	0	73%
24		He-M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	5	21	0	0	1	0	70%
25		He-E	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	27	0	0	0	90%
26		Wa-I	0	0	0	0	0	1	0	0	1	1	0	0	0	0	1	4	0	0	0	0	0	0	0	0	0	12	10	0	40%
27		Wa-E	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	4	0	0	0	0	0	0	0	0	13	10	0	33%
28		Ee-E	0	0	2	0	0	12	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0	2	0	0	0	0	0	11	37%
		Total																													**54%**

After analyzing the recognition results, the sources for confusion can be categorized as follows:

- Intra-class errors: These errors result from the similarity of the positional shapes of the same character. For

example, the beginning and middle forms of most Arabic letters are similar, their primary difference is a connecting stroke called "Kashida ("_")". Similarly, the isolated and end forms are similar for most of the Arabic letters. Fig. (**12**) shows samples of this type of error. This type of error occurred 119 times (*i.e.*, approximately 30% of all errors). Combining the labels of the different forms of the same character into the same code, as performed in [3], the recognition rate of the proposed method was 69%. This type of error may be less problematic because word processors use contextual analysis and display character shapes based on its position within the text (*i.e.*, beginning, middle, end, or isolated).

- Inter-class errors: These errors are caused by the similarity between characters of different classes due to character segmentation. This type of error occurred 68 times (*i.e.*, 17% of all errors). (Fig. **13**) shows samples of these errors. Integrating shape context in the recognition of Arabic online text is expected to eliminate these errors. These errors disappear during online text recognition because the shapes of the beginning, middle, end, and isolated forms are different.
- Writing distortion and variation errors: This type of error results from writing distortions and variations that come from the unconstrained writing of online text. This type includes variations in the direction of writing movements (*e.g.* , Ma-B vs. Fa-B) or due to curvature (*e.g.* , Ba-E with the shape of end ب without a dot, vs. Ee-E). This type of error occurred 122 times (*i.e.*, 31% of all errors), and examples of this type of error are shown in Fig. (**14**). The differentiation of Ay-M and Fa-M was addressed in [56]. In general, most of these errors are expected to disappear during online text recognition. For example, Ma-B and Fa-B are differentiated by a dot over the Fa symbol and a lack of a dot in the Ma symbol. Writing these characters properly will result in different features as their shapes are different, allowing their correct classification. In other cases, more than one model should be used for certain classes to address the different ways of writing those classes.
- Genuine errors: These errors are due to different characters that look similar. These errors account for approximately 22% of all errors. An example of such an error is the confusion of Ha-B with He-B. To address these errors, more discriminant features and classifiers are required. We are currently addressing this issue in an extension of this study in online character and text recognition.

There is a trend to use deep learning for several computer vision problems including pattern recognition tasks such as handwriting recognition. For Arabic handwriting recognition, deep learning is successfully utilized to recognize offline handwritten Arabic characters as in [58 - 60]. Moreover, the applicability of deep learning is also examined for the online case as in [61], we plan to use for future work.

a) Samples of "Ha-M" (left) recognized as "Ha-B" (right)

b) Samples of "Wa-I" (left) recognized as "Wa-E" (right)

c) Samples of "Fa-B" (left) recognized as "Fa-M" (right)

Fig. (12). Examples of intra class errors while recognizing Arabic characters based on primary shape only.

a) Samples of "Fa-B" (left) recognized as "Wa-I" (right)

b) Samples of "Da-E" (left) recognized as "Ba-M" (right)

c) Samples of "Wa-E" (left) recognized as "Fa-M" (right)

Fig. (13). Examples of inter class errors while recognizing Arabic characters based on primary shape only.

a) Samples of "Ay-M" (left) recognized as "Fa-M" (right)

b) Samples of "Ma-B" (left) recognized as "Fa-B" (right)

Fig. (14). Examples of errors originating from writing distortions (bad handwriting).

CONCLUSION

In this study, we presented an Arabic online text database (Online-KHATT). The Online-KHATT database contains 10,040 lines of Arabic online text with more than 80,000 Arabic words written by 623 writers. The source text of the Online-KHATT database covers several domains to ensure a wide range of topics. We applied a three-level verification procedure to the source text and to the written online text compared to the ground truth. The review process matched the ground truth with the written text images. We also included the transliteration of the database text and tracked writing errors at the character level, which researchers can use in their studies. These added resources could allow researchers of Arabic online text recognition to easily evaluate the performances of their systems.

A pilot Arabic online character recognition system that uses the segmented characters of the Online-KHATT database was presented. The results are encouraging because the data are unconstrained online data, in which variations are large, and different ways of writing the characters are used, where certain forms are unnatural (*e.g.* , writing in the reverse direction).

To the best of our knowledge, Online-KHATT database is the largest Arabic online text database in terms of the number of lines written with electronic pens using natural Arabic text. This database is freely available upon request for interested researchers. With its large Arabic online text and associated ground truth, transliteration, and other resources, we believe that the Online-KHATT database can be used as a benchmark database for studies related to online text recognition and related areas.

CONSENT FOR PUBLICATION

Not applicable.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEGEMENTS

The authors would like to thank the referees for their constructive criticism and stimulating remarks. The modification of the original manuscript to address those remarks improved the revised manuscript. This work was funded by the National Plan for Science, Technology and Innovation (MAARIFAH)-King Abdul-Aziz City for Science and Technology-through the Science & Technology Unit at King Fahd University of Petroleum & Minerals (KFUPM)-the Kingdom of Saudi Arabia, project no. 11-INF2153-4.

REFERENCES

[1] S.A. Mahmoud, I. Ahmad, W.G. Al-Khatib, M. Alshayeb, M. Tanvir Parvez, V. Märgner, and G.A. Fink, "KHATT: An open Arabic offline handwritten text database", *Pattern Recognit.,* vol. 47, pp. 1096-1112, 2014. [http://dx.doi.org/10.1016/j.patcog.2013.08.009]

[2] A. Mahdi, *"Spell Checking and Correction for Arabic Text Recognition"*, M.S. thesis, King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia, 2012.

[3] H. Al-Muhtaseb, S. Mahmoud, and R. Qahwaji, "Recognition of off-line printed Arabic text using Hidden Markov Models", *Signal Processing,* vol. 88, pp. 2902-2912, 2008. [http://dx.doi.org/10.1016/j.sigpro.2008.06.013]

[4] F. Biadsy, J. El-Sana, and N.Y. Habash, "Online arabic handwriting recognition using hidden markov models", In: *Proc. 10th Int. Work. Front.* Handwrit. Recognit., La Baule, France, 2006, pp. 85-90.

[5] F. Biadsy, R. Saabni, and J. El-Sana, "Segmentation-free online arabic handwriting recognition", *Int. J. Pattern Recognit. Artif. Intell.,* vol. 25, pp. 1009-1033, 2011. [http://dx.doi.org/10.1142/S0218001411008956]

[6] J. Sternby, J. Morwing, J. Andersson, and C. Friberg, "On-line Arabic handwriting recognition with templates", *Pattern Recognit.,* vol. 42, pp. 3278-3286, 2009. [http://dx.doi.org/10.1016/j.patcog.2008.12.017]

[7] M. Kherallah, A. Elbaati, H.E. Abed, and A.M. Alimi, "The on/off (LMCA) dual Arabic handwriting database", In: *11th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Montréal, Québec, Canada, 2008.

[8] M. Kherallah, L. Haddad, A.M. Alimi, and A. Mitiche, "On-line handwritten digit recognition based on trajectory and velocity modeling", *Pattern Recognit. Lett.,* vol. 29, pp. 580-594, 2008. [http://dx.doi.org/10.1016/j.patrec.2007.11.011]

[9] M. Kherallah, F. Bouri, and A.M. Alimi, "On-line Arabic handwriting recognition system based on visual encoding and genetic algorithm", *Eng. Appl. Artif. Intell.,* vol. 22, pp. 153-170, 2009.
[http://dx.doi.org/10.1016/j.engappai.2008.05.010]

[10] M.A. Abuzaraida, A.M. Zeki, and A.M. Zeki, "Online database on Quranic hadwritten words", *J. Theor. Appl. Inform. Technol.,* vol. 62, pp. 485-492, 2014.

[11] M. Pechwitz, S.S. Maddouri, V. Margner, N. Ellouze, and H. Amiri, "IFN/ENIT-database of handwritten Arabic words", In: *Proc. of CIFED*, 2002, pp. 127-136.

[12] H. El Abed, M. Kherallah, V. Märgner, and A.M. Alimi, "On-line Arabic handwriting recognition competition", In: *International Conference on Document Analysis and Recognition (ICDAR)*, 2011, pp. 15-23.

[13] H. El Abed, V. Margner, M. Kherallah, and A.M. Alimi, "ICDAR 2009 online Arabic handwriting recognition competition", *International Conference on Document Analysis and Recognition (ICDAR'09),* 2009pp. 1388-1392

[14] H. Ahmed, and S.A. Azeem, "On-line Arabic handwriting recognition system based on HMM", In: *International Conference on Document Analysis and Recognition*, Beijing, China, 2011, pp. 1324-1328.
[http://dx.doi.org/10.1109/ICDAR.2011.266]

[15] S. Abdelazeem, and H.M. Eraqi, "On-line Arabic handwritten personal names recognition system based on HMM", In: *International Conference on Document Analysis and Recognition*, Beijing, China, 2011, pp. 1304-1308.
[http://dx.doi.org/10.1109/ICDAR.2011.262]

[16] H. Boubaker, A. Chaabouni, M. Kherallah, A.M. Alimi, and H. El Abed, "Fuzzy segmentation and graphemes modeling for online Arabic handwriting recognition", In: *International Conference on Frontiers in Handwriting Recognition*, Kolkata, India., 2010, pp. 695-700.
[http://dx.doi.org/10.1109/ICFHR.2010.113]

[17] H. Boubaker, A. El Baati, M. Kherallah, A.M. Alimi, and H. Elabed, "Online Arabic handwriting modeling system based on the graphemes segmentation", In: 20th International Conference on Pattern Recognition. , Istanbul, Turkey, 2010, pp. 2061-2064.
[http://dx.doi.org/10.1109/ICPR.2010.507]

[18] H. Boubaker, A. Chaabouni, M. Ben Halima, A. El Baati, and H. El Abed, "Arabic diacritics detection and fuzzy representation for segmented handwriting graphemes modeling", In: *6th International Conference on Soft Computing and Pattern Recognition*, Tunis, Tunisia, 2014, pp. 71-76.
[http://dx.doi.org/10.1109/SOCPAR.2014.7007984]

[19] H. Boubaker, N. Tagougui, H. El Abed, M. Kherallah, and A.M. Alimi, "Graphemes Segmentation for Arabic Online Handwriting Modeling", *J. Inf. Process. Syst.,* vol. 10, pp. 503-522, 2014.
[http://dx.doi.org/10.3745/JIPS.02.0006]

[20] N. Tagougui, H. Boubaker, M. Kherallah, and A.M. Alimi, "A Hybrid NN/HMM Modeling Technique for Online Arabic Handwriting Recognition", *Int. J. Comput. Linguist. Res.,* vol. 4, pp. 107-118, 2013.

[21] N. Tagougui, H. Boubaker, M. Kherallah, and A.M. Alimi, "A hybrid MLPNN/HMM recognition system for online Arabic Handwritten script", In: *World Congress Computer and Information Technology,* Sousse, Tunisia, 2013, pp. 1-6.
[http://dx.doi.org/10.1109/WCCIT.2013.6618744]

[22] S.A. Azeem, and H. Ahmed, "Combining online and offline systems for Arabic handwriting recognition", In: *21st International Conferance Pattern Recognition,* Tsukuba, Japan, 2012, pp. 3725-3728.

[23] I. Hosny, S. Abdou, and A. Fahmy, "Using advanced hidden markov models for online Arabic handwriting recognition", In: *First Asian Conferance Pattern Recognition*, 2011, pp. 565-569.
[http://dx.doi.org/10.1109/ACPR.2011.6166664]

[24] H. Al-Barhamtoshy, S. Abdou, and F.A. Al-Wajih, "A Toolkit for Teaching Arabic Handwriting", *Int. J. Comput. Appl.,* vol. 49, pp. 17-23, 2012.

[25] I. Abdelaziz, S. Abdou, and H. Al-Barhamtoshy, "A large vocabulary system for Arabic online handwriting recognition", *Pattern Anal. Appl.* pp. 1-13, 2015.

[26] G. Kour, and R. Saabne, "Fast classification of handwritten on-line Arabic characters", In: *6th Int. Conf. Soft Comput. Pattern Recognit.*, Tunis, Tunisia, 2014, pp. 312-318.
[http://dx.doi.org/10.1109/SOCPAR.2014.7008025]

[27] R. Saabni, and J. El-Sana, "Comprehensive synthetic Arabic database for on/off-line script recognition research", *Int. J. Doc. Anal. Recognit.,* vol. 16, pp. 285-294, 2013.
[http://dx.doi.org/10.1007/s10032-012-0189-5]

[28] S. Njah, B. Ben Nouma, H. Bezine, and A.M. Alimi, "MAYASTROUN: A Multilanguage Handwriting Database", In: 2012 *Int. Conf. Front. Handwrit. Recognit.*, Bari, Italy, 2012, pp. 308-312.

[29] R.I. Elanwar, M. Rashwan, and S. Mashali, "OHASD: the first on-line Arabic sentence database handwritten on tablet PC", In: *Proc. World Acad. Sci. Eng. Technol. (WASET), Int. Conf. Int. Conf. Signal Image Process. ICSIP,* Singapore, 2010, pp. 910-915.

[30] R.I. Elanwar, M. Rashwan, and S. Mashali, "On-Line Arabic Handwriting Text Line Detection Using Dynamic Programming", In: *Int. Conf. Comput. Math. Nat. Comput.*, Penang, Malaysia., 2011, pp. 588-593.

[31] R.I. Elanwar, M. Rashwan, and S. Mashali, "Unconstrained arabic online handwritten words segmentation using new hmm state design", *Int. Sch. Sci. Res. Innov.,* vol. 6, pp. 1189-1197, 2012.

[32] I. Abdelaziz, and S. Abdou, "AltecOnDB: A Large-Vocabulary Arabic Online Handwriting Recognition Database", arXiv Prepr. arXiv1412.7626. 2014.

[33] R. Saabni, and J. El-Sana, "Efficient generation of comprehensive database for online arabic script recognition", In: *10th Int. Conf. Doc. Anal. Recognition, ICDAR'09.*, Barcelona, Spain, 2009, pp. 1231-1235.
[http://dx.doi.org/10.1109/ICDAR.2009.258]

[34] R. Saabni, and J. El-Sana, "Hierarchical on-line arabic handwriting recognition", In: *10th Int. Conf. Doc. Anal. Recognition, ICDAR'09,* Barcelona, Spain, 2009, pp. 867-871.

[35] N. Mezghani, A. Mitiche, and M. Cheriet, On-line recognition of handwritten arabic characters using a kohonen neural network In: *Proc. Eighth Int. Work. Front. Handwrit. Recognit.*, Ontario, Canada, 2002, pp. 490-495.

[36] N. Mezghani, M. Cheriet, and A. Mitiche, "Combination of pruned kohonen maps for on-line arabic characters recognition", In: *Seventh Int. Conf. Doc. Anal. Recognit.*, Edinburgh, UK, 2003, pp. 900-904.

[37] N. Mezghani, A. Mitiche, and M. Cheriet, "On-line character recognition using histograms of features and an associative memory", In: *IEEE Int. Conf. Acoustics, Speech, Signal Process. ICASSP-04,* Montreal, Canada, 2004, pp. 841-844.

[38] N. Mezghani, A. Mitiche, and M. Cheriet, "A new representation of character shape and its use in on-line character recognition by a self organizing map", In: *Int. Conf. Image Process. ICIP'04.*, Singapore, 2004, pp. 2123-2126.
[http://dx.doi.org/10.1109/ICIP.2004.1421505]

[39] N. Mezghani, A. Mitiche, and M. Cheriet, "A new representation of shape and its use for high performance in online Arabic character recognition by an associative memory", *Int. J. Doc. Anal. Recognit.,* vol. 7, pp. 201-210, 2005.
[http://dx.doi.org/10.1007/s10032-005-0145-8]

[40] S. Izadi, and C.Y. Suen, "Online Writer-Independent Character Recognition Using a Novel Relational Context Representation", In: *Seventh Int. Conf. Mach. Learn. Appl. ICMLA'08.*, San Diego, California, USA, 2008, pp. 867-870.

[41] N. Mezghani, A. Mitiche, and M. Cheriet, "Bayes classification of online arabic characters by Gibbs modeling of class conditional densities", *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 30, no. 7, pp. 1121-1131, 2008.
[http://dx.doi.org/10.1109/TPAMI.2007.70753] [PMID: 18550897]

[42] A.T. Al-Taani, and S. Al-Haj, "Recognition of on-line arabic handwritten characters using structural features", *J. Pattern Recognit. Res.,* vol. 1, pp. 23-37, 2010.
[http://dx.doi.org/10.13176/11.217]

[43] K. Addakiri, and M. Bahaj, "On-line handwritten arabic character recognition using artificial neural network", *Int. J. Comput. Appl.,* vol. 55, pp. 42-46, 2012.

[44] M.A.H. Omer, and S.L. Ma, "Online Arabic handwriting character recognition using matching algorithm", In: *2nd Int. Conf. Comput. Autom. Eng.*, Singapore, 2010, pp. 259-262.
[http://dx.doi.org/10.1109/ICCAE.2010.5451492]

[45] I. Khodadad, M. Sid-Ahmed, and E. Abdel-Raheem, "Online Arabic/Persian character recognition using neural network classifier and DCT features", In: *IEEE 54th Int. Midwest Symp. Circuits Syst.*, 2011, pp. 1-4.

[46] A. Ramzi, and A. Zahary, "Online Arabic handwritten character recognition using online-offline feature extraction and back-propagation neural network", In: *1st Int. Conf. Adv. Technol. Signal Image Process.*, Sousse, Tunisia, 2014, pp. 350-355.

[47] M. Harouni, D. Mohamad, and A. Rasouli, "Deductive method for recognition of on-line handwritten Persian/Arabic characters", In: *2nd Int. Conf. Comput. Autom. Eng.*, Singapore, 2010, pp. 791-795.
[http://dx.doi.org/10.1109/ICCAE.2010.5451869]

[48] S.A. Azeem, and H. Ahmed, "Recognition of Segmented Online Arabic Handwritten Characters of the ADAB Database", In: *10th Int. Conf. Mach. Learn. Appl. Work. (ICMLA).*, Hawaii, USA., 2011, pp. 204-207.
[http://dx.doi.org/10.1109/ICMLA.2011.120]

[49] H.M. Eraqi, and S.A. Azeem, "An On-line Arabic Handwriting Recognition System: Based on a New On-line Graphemes Segmentation Technique", In: *Int. Conf. Doc. Anal. Recognit.*, Beijing, China, 2011, pp. 409-413.
[http://dx.doi.org/10.1109/ICDAR.2011.90]

[50] A.T. Al-Taani, "An efficient feature extraction algorithm for the recognition of handwritten arabic digits", *Int. J. Comput. Intell,* vol. 2, pp. 107-111, 2005.

[51] A.T. Al-Taani, and H. Maen, "Recognition of on-line handwritten Arabic digits using structural features and transition network", *Informatica.,* vol. 32, pp. 275-281, 2008.

[52] S.A. Azeem, M. El Meseery, and H. Ahmed, "Online Arabic Handwritten Digits Recognition", In: *Int. Conf. Front. Handwrit. Recognit.*, Bari, Italy, 2012, pp. 135-140.

[53] M.A. Abuzaraida, A.M. Zeki, and A.M. Zeki, "Online Recognition System For Handwritten Arabic Digits", In: *ICIT15 7th Int. Conf. Inf. Technol.*, Amman, Jordan, 2015, pp. 45-49.
[http://dx.doi.org/10.15849/icit.2015.0007]

[54] M. Parvez, and S.A. Mahmoud, "Arabic handwriting recognition using structural and syntactic pattern attributes", *Pattern Recognit.,* vol. 46, pp. 141-154, 2013.
[http://dx.doi.org/10.1016/j.patcog.2012.07.012]

[55] B.Q. Huang, Y.B. Zhang, and M-T. Kechadi, "Preprocessing techniques for online handwriting recognition", In: *Intell. Text Categ. Clust.*, Springer, 2009, pp. 25-45.

[56] S.A. Mahmoud, "Arabic character recognition using fourier descriptors and character contour encoding", *Pattern Recognit.,* vol. 27, pp. 815-824, 1994.
[http://dx.doi.org/10.1016/0031-3203(94)90166-X]

[57] K.P. Murphy, *Dynamic bayesian networks: representation, inference and learning.,* University of California: Berkeley, 2002.

[58] S.B. Ahmed, S. Naz, M.I. Razzak, and R. Yousaf, "Deep learning based isolated Arabic scene character recognition", In: *Arabic Script Analysis and Recognition (ASAR)*, 1st International Workshop on, 2017, pp. 46-51.
[http://dx.doi.org/10.1109/ASAR.2017.8067758]

[59] A. Ashiquzzaman, and A.K. Tushar, "Handwritten Arabic numeral recognition using deep learning neural networks", In: *Imaging, Vision & Pattern Recognition (icIVPR), IEEE International Conference,* 2017, pp. 1-4.

[60] C. Boufenar, and M. Batouche, *"Investigation on deep learning for off-line handwritten Arabic Character Recognition using Theano research platform," Intell. Syst. Comput. Vision.,* ISCV, 2017, pp. 1-6.

[61] N. Tagougui, and M. Kherallah, "Recognizing online Arabic handwritten characters using a deep architecture", In: *Proc. SPIE 10341, Ninth International Conference on Machine Vision,* 2017, pp. 103410L.