KHATT: Arabic Offline Handwritten Text Database

Sabri A. Mahmoud[a], Irfan Ahmad[a], Mohammad Alshayeb[a], Wasfi G. Al-Khatib[a], Mohammad Tanvir Parvez[b], Gernot A. Fink[c], Volker Märgner[d], and Haikal El Abed[d]
[a] *King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia*
{smasaad, irfanics, alshayeb, wasfi}@kfupm.edu.sa
[b] *Qassim University, Qassim, Saudi Arabia*
{m.parvez@qu.edu.sa}
[c] *Technische Universität Dortmund, 44227 Dortmund, Germany*
Gernot.Fink@tu-dortmund.de
[d] *Technische Universität Braunschweig, 38092 Braunschweig, Germany*
{v.maergner, elabed}@tu-bs.de

Abstract

In this paper, we report our comprehensive Arabic offline Handwritten Text database (KHATT) after completion of the collection of 1000 handwritten forms written by 1000 writers from different countries. It is composed of an image database containing images of the written text at 200, 300, and 600 dpi resolutions, a manually verified ground truth database that contains meta-data describing the written text at the page, paragraph, and line levels. A formal verification procedure is implemented to align the handwritten text with its ground truth at the form, paragraph and line levels. Tools to extract paragraphs from pages and segment paragraphs into lines are developed. Preliminary experiments on Arabic handwritten text recognition are conducted using sample data from the database and the results are reported.

The database will be made freely available to researchers world-wide for research in various handwritten-related problems such as text recognition, writer identification and verification, etc.

1. Introduction

Researchers consider the lack of freely available and comprehensive Arabic handwritten databases as one of the reasons for the scarcity of research on Arabic text recognition compared with other languages [1][2].

KHATT is a transcription of the Arabic word (*خط*) which means 'handwriting'. It is also shorthand for **K**FUPM **H**andwritten **A**rabic **T**ex**T**.

This paper complements the earlier report on this database [3]. The earlier paper presented details on the design of the handwriting sample forms and paragraphs where only 300 forms were reported. Here we report the statistics of the full database, which consists of 1000 forms written by distinct writers; our verification procedure, and experimental results using parts of the database. Each form consists of 4 pages. The first page collects information about the writer (viz. the name, age category, upbringing country, qualification, gender, left/right-handedness, and a section for management purposes). The second page consists of two paragraphs; a fixed paragraph that consists of minimal text covering all the shapes of Arabic characters. This paragraph may be used to extract all Arabic characters in all their forms. It was shown in [4] that when the number of samples for some Arabic shapes are not enough, recognition rate can be improved by including 50 samples of a minimal text paragraph that covers all shapes of Arabic characters to train HMM. The second paragraphs in this page were collected randomly from a large corpus to make the database a statistical representation of the corpus. Each random paragraph is unique over all forms. The third page consists of another unique random paragraph and the fixed text paragraph of the second page. The fourth page is designed for free writing with ruled lines provided at the lower half of the page. The last page allows more subjects to be added to the database. It also may be used for research

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.224

on the effect of lined pages on the quality of handwriting of individuals. We would like to make a note that the design of the forms was based on our analysis of available databases on Arabic and Latin languages. In particular, the IAM-database [5], the CEDAR letter [6] and the Firemaker dataset [7] were analyzed. It was designed to serve research in Arabic writer identification & verification in addition to Arabic handwritten text recognition.

This paper is organized as follows: Section 2 presents the literature review related to developing Arabic offline handwriting databases. Data collection and statistics are presented in Section 3. Task definition is presented in Section 4. Data verification is described in Section 5. In Section 6, the experimental results of Arabic text recognition using part of the data are presented. Finally, we present the conclusions in Section 7.

2. Literature Review

Most of the efforts of developing Arabic handwritten databases are limited in being closed vocabulary or made of words, digits, characters, and limited sentences.

IFN/ENIT database [8][9] was developed by the Institute of Communications Technology (IFN) at Technical University Braunschweig in Germany and The National School of Engineers of Tunis (ENIT) in 2002. It is made of Tunisian town/village names written by 411 writers. It is one of the most widely used databases. The use of this database resulted in improving the state of the art in Arabic handwritten text recognition. However, it lacks the naturalness of handwritten Arabic text as it essentially contains names of towns and villages of Tunisia and as such is a limited vocabulary database. Al ISRA database [10] contains Arabic words, digits, signatures, and free form Arabic sentences gathered from five hundred randomly selected students at Al Isra University in Amman, Jordan. This database was collected by a group of researchers at the University of British Columbia. This database has the same limitation regarding Arabic text.

Al-Ohali et al. [11] developed an Arabic check database that includes Arabic legal and courtesy amounts that were extracted from 3000 bank checks of AlRajhi Bank, Saudi Arabia. This database can be mainly used for bank check applications. AHDB database includes words that are used in writing legal amounts on Arabic checks and some free handwriting pages from 100 writers [12]. Khedher et al. [13] developed a database of unconstrained Arabic handwritten characters.

AlAmri [14] developed a database containing Arabic dates, isolated digits, numerical strings, letters, words and some special symbols. ADBase database [15] developed a database of handwritten Arabic (Indian) digits which is suitable for research in Arabic digit recognition.

Table 1 below, shows selected databases on Arabic handwritten characters, words, text, and digits.

Table 1. Selected Arabic handwritten databases

Database	Description	Writers #
KHATT database	1000 forms, 2000 (random and fixed paragraphs) & free paragraphs	1000
IFN/ENIT [8]	26,459 images of Tunisian city names	411
Al–Isra [10]	37,000 words, 10,000 digits, 2,500 signatures, 500 sentences	500
CENPARMI [11]	3,000 checks (Legal and courtesy amounts and digits)	–
AHDB [12]	10,000 words for check processing	100
Khedher et al. [13]	Characters	48
Alamri et al.[14]	46,800 digits, 13,439 numerical strings, 21,426 letters, 11,375 words, 1,640 special symbols	328
AD/MADBase [15]	700,000 digits	700

We are not aware of any comprehensive and open vocabulary Arabic handwritten text database of adequate size that reflects the naturalness of Arabic text. KHATT database was made to fulfill this need. It will be made freely available to interested researchers.

3. Data Collection and Statistics

The data of the forms was collected from 46 different sources; these sources cover 11 different subjects. Table 2 shows sources' topics and the number of collected paragraphs in each topic.

The forms were filled by people raised in 18 countries. Table 3 shows the number of forms filled by people raised in each country. The 'others' include countries in which we collected less than 10 forms, these countries include: Canada, Syria, Sudan, Algeria, Australia, Bahrain, Lebanon, Libya and Oman.

The forms were filled by people of different ages and qualifications. Table 4 shows the number of forms filled by people in each age category and educational qualifications. Out of the 1000 writers, 677 were male while 323 were female and 928 were right handed while 72 were left handed.

Table 2. Source data's topics and paragraphs

Topic	# of Sources	# of Paragraphs
Art	3	399
Economy	4	81
Education	1	46
Health	3	103
History	7	177
Literature	5	636
Management	3	75
Nature	6	134
Social	5	128
Technology	5	189
World	3	32
Total	**45**	**2000**

Table 3. Writers' upbringing country

Country	#	Country	#
Saudi Arabia	676	USA	16
Morocco	90	Egypt	13
Jordan	79	Tunisia	13
Yemen	45	Kuwait	11
Palestine	29	*Others*	28
		Total: 1000	

Table 4. Writers' age and qualifications

Age	#	*Qualification*	#
<15	126	Elementary school	88
16 - 25	643	High School	537
26 - 50	215	University	375
> 50	16		
	1000		**1000**

The ground truth for the data was constructed in two formats, text and XML. One text file was generated for each paragraph with the same structure as written by the writer. We also generated XML files for the truth values similar to the format used by IfN/ENIT [8]. The data was manually verified by a process explained in Section 4.

The data was divided into three sets; 70% training, 15% testing and 15% for validation. The country of upbringing was classified into three regions: Region 1 includes the gulf and Middle East countries, namely: Saudi Arabia, Jordan, Yemen, Kuwait, Palestine, Syria, Bahrain, Lebanon and Oman. Region 2 includes the African countries, namely: Tunisia, Morocco, Egypt, Sudan, Algeria and Libya and Region 3 consists of USA, Canada and Australia. Table 5 shows the statistics with regard to this classification.

Table 5. Statistics on training, testing and validation set

	Train	Test	Validate
Region 1	595	128	128
Region 2	87	18	18
Region 3	18	4	4
Right Handed	650	139	139
Left Handed	50	11	11
Male	476	100	101
Female	224	50	49
Elementary school	65	11	12
High School	371	83	83
University	264	56	55
<15	85	22	19
16 - 25	446	97	100
26 - 50	157	29	29
> 50	12	2	2

4. Task Definition

In this section, statistics on the database is presented. The database is split into three exclusive parts, viz. training, validation, and test sets. Table 6 shows the uni-, bi-, and tri-grams of these sets and of the full database. The out of vocabulary (OOV) of the validation and test data sets compared to the training data sets is 43.86% and 44.38%, respectively. The OOV of KHATT pilot testing set compared to the pilot training set is 55.38%.

For the sake of initial experimentation, KHATT pilot data is formed from lines extracted from paragraphs 3 & 4 of the forms. KHATT pilot training data, which consists of 1400 lines, were written by 258 writers and a test data of 233 lines were written by another 40 writers.

Table 6. N-gram statistics of the database

Set	Word count	Unit-gram	Bi-gram	Tri-gram
Training	125180	19605	59602	68052
Validation	26916	6739	14542	15155
Testing	26159	6510	13999	14555
All data	178255	25194	82846	96416

5. Data Verification

Each form included in the KHATT database has undergone three phases of verification. The first phase verifies it at the form level to ensure the suitability of including it in the database. The second phase verifies

the correctness of the ground truth of the scanned form by ensuring that the ground truth matches the handwritten text on the form. The third stage ensures the absence of any errors resulting from the scanning process or the image segmentation of the scanned forms at the paragraph and line levels.

5.1 Verification at the Form Level

The digitized form is checked for the overall quality of the scanning, ensuring the absence of any cropped portions from the pages. In addition, the handwriting in each form is classified based on its "quality" into three categories: Readable, Challenging and Unreadable. The objectives of this classification are to discard truly "bad" forms that do not properly reflect the Arabic handwriting and to distribute readable and challenging forms into the training, validation and verification data sets according to their size. Figure 1 (a) shows sample text from one of the random paragraphs and Figure 1(b) shows how it was actually written. Many forms that were deemed unreadable were written by high school and middle school students. These forms were replaced with other forms. Although determining whether a certain written text is unreadable is highly subjective, the reviewers tried their best to reduce the subjectivity by attempting to read "apparently" difficult words from the handwritten text without relying on the context.

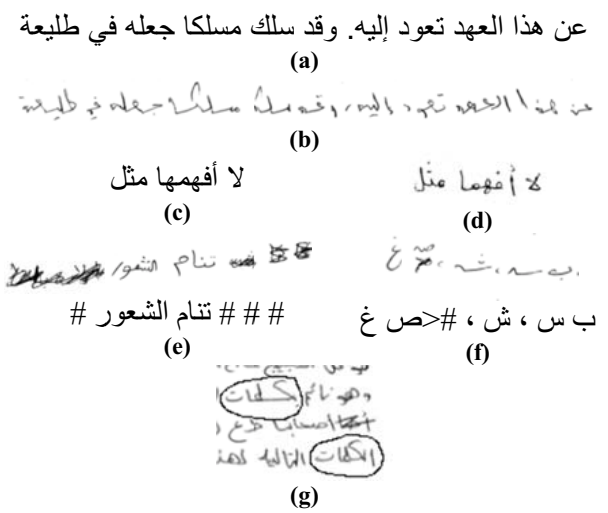

Figure 1. Various writing samples that clarify our verification process.

5.2 Verification of the Ground Truth Correctness

In this phase, the scanned forms were divided into disjoint sets, where each set was given to a separate reviewer to verify that the Ground Truth (GT) exactly matches the written text. Since it is customary that writers may slightly deviate from what is written in the form, the reviewers tried to ensure that the GT conforms to the handwritten text. The reviewers carried out ground truth verification according to the following rules:

1- If the pre-printed word (originally on the form) is different from the written word, the GT is replaced by the written word. Figure 1(c) shows sample original text and Figure 1(d) shows how a writer has written it.

2- When a writer scratches a word, it is encoded as a single "#" character. As long as the scratch appears as a single connected component, irrespective of its size, it will be encoded with one mark, as shown in Figure 1(e). When the scratch is below a word in a certain line, as shown in Figure 1(f), it is encoded as ">#". Similarly, if it is above the word, it is encoded as "#<".

3- If there are apparently more spaces between two words in the handwritten text as compared to the regular one space in the GT, the reviewer estimates the number of empty spaces and adds them to the GT. In some text, elongation is sometimes practiced by some writers, as shown in Figure 1(g). The number of Kashida (dash) is estimated by the reviewer to the best of his judgment and is added to the GT.

4- One should differentiate between Taa Marbootah ـة and Haa ـه. For example, if it is written as ـة in the GT, whereas the writer has written it as ـه, the ـة is replaced by a ـه. Otherwise, the number of dots, and hence the corresponding character is kept as in the original text regardless of the location of the dot or the "apparent" number of them.

5- In some cases, the writer may write certain letters in a different way. If that way is significantly different from the original encoded character, it is encoded according to what is written. Otherwise, if it is similar but not different, it is kept as the original character.

6- If there is a word in the GT that was missed by the form writer, that word is deleted from the ground truth.

5.3 Verification at the Paragraph and Line Levels

In the third phase, scanned forms were verified at the paragraph and line levels to ensure that there were no errors in the extracted paragraph and line images. This process includes the following steps:

1- The paragraph and line numbers of the extracted images must match their corresponding numbers in the original form.
2- If portions of the previous line or the next line exist in the image file of the current line, this has been noted. This is also carried out at the paragraph level. Similarly, a note is taken when part of the line or the paragraph in the corresponding image file is missing,

After this process is carried out, which we call the initial verification process, forms are given to a different reviewer to carry out the final verification process following phases 2 and 3. The purpose of this is to correct any mistakes that were overlooked in the initial verification process.

6. Experiments 'KHATT PILOT'

In this section we present the preliminary experiments we conducted for handwritten text recognition using KHATT pilot data.

The form images were scanned at 300 DPI. The form pages were binarized using the algorithm of Otsu [16]. Then, salt and pepper noise is removed by using median filtering. Forms were skew corrected and the paragraphs were extracted. Text lines were extracted from the paragraphs using the algorithm as described in [17]. Baseline correction was performed using the technique presented in [18] which is adapted for Arabic handwritten text. We performed the slant correction using the algorithm as described in [17]. Finally lines are normalized to a height of 96 pixels keeping the aspect ratio fixed.

Hidden Markov model (HMM) was used as classifier. The general trend for cursive text recognition is to use HMM [19]. The use of other classifiers requires the segmentation of cursive text into characters which is error prone and is implicitly done by HMM. HTK tools [20] were used in the experimentation using HMM. A left-to-right discrete HMM with Bakis topology is used for handwritten Arabic text recognition with same number of states for every class/model. We treated every character shape as a different class during training and recognition. The recognized shapes of the same character were merged into one class as they represent one character. This is opposed to some researchers treating different character shapes as the same class. We had a total of 149 classes modeled in HMM corresponding to all the possible character shapes and the digits. The output text is merged into 49 classes.

We extracted a number of statistical features of the line image using the sliding window technique. We tried different window sizes and overlap but the best results were obtained using window width of four pixels with a two pixel overlap.

Pixel density features from the text line image and its horizontal and vertical edge derivatives were computed. Edge derivatives were calculated using Sobel operator. We implemented some additional statistical features as described in [18] and [21] (adapted for Arabic handwritten text) and gradient features. Once the features were extracted, they were quantized into linear codebook using a top-down clustering process and nearest neighbor for clustering. After training the HMM, the recognition was performed on the test set. Table 7 reports the best recognition results obtained at character level. The code book sizes and the number of states for the best results are also shown in the table.

The low recognition rate may be attributed to limited data in the KHATT PILOT that was used for experimentation. The data is real, unconstrained natural handwriting, it is an open-vocabulary problem, and the language model and dictionary were not used. Moreover some characters and character shapes (like غ ض ؤ ز ظ ج), having relatively very few training samples, were having very low recognition rates. So it is expected that using complete database will improve the recognition rates. Nevertheless comparatively low recognition rate explains the need for future research in this challenging area and is expected to encourage researchers to address this difficult task.

Table 7. Recognition rates on sample data using different features.

Features Used	Corr. (%)	Acc. (%)	Code book	States
Image adaptive pixel density features & horizontal & vertical edge derivatives	54.8	47.8	256	12
Statistical features adapted from [18] & [21]	53.1	46.0	350	14
Adaptive Gradient Features	55.9	51.2	196	14

7. Conclusions

In this paper we presented a comprehensive Arabic off-line handwritten text database written by 1000 distinct writers from different upbringing country, age group, qualification, gender, and left/right-handedness.

Each writer filled a form of 4 pages which are scanned at 200, 300, 600 dpi resolution. The ground truth was verified manually. It was verified at the form, paragraph, and line levels. The forms are segmented into paragraphs and paragraphs into lines. Tools to extract paragraphs, lines, skew correction, etc. are developed. Preliminary experiments on Arabic handwritten text recognition are conducted using KHATT pilot data and the results are reported.

This database addresses the need by the research community interested in Arabic text recognition, writer identification and verification, forms analysis, segmentation, etc.

The database will be made freely available to interested researchers.

Acknowledgement

The authors would like to acknowledge the support provided by King Abdul-Aziz City for Science and Technology (KACST) through the Science & Technology Unit at King Fahd University of Petroleum & Minerals (KFUPM) for funding this work through project no. 08-INF99-4 as part of the National Science, Technology and Innovation Plan. In addition, we would like to thank all the writers and persons who contributed to this database.

References

[1] B. Al-Badr and S. Mahmoud, "Survey and bibliography of Arabic optical text recognition," *Signal Processing*, vol. 41, no. 1, pp. 49-77, Jan. 1995.

[2] M. T. Parvez and S. A. Mahmoud, "Off-line Arabic Handwritten Text Recognition: A Survey," *Accpted in ACM Computing Surveys*, 2011.

[3] S. Mahmoud, I. Ahmad, M. Alshayeb, and W. Al-Khatib, "A Database for Offline Arabic Handwritten Text Recognition," in *Image Analysis and Recognition - 8th International Conference, (ICIAR 2011)*, 2011, vol. 6754, pp. 397-406.

[4] H. Al-Muhtaseb, S. Mahmoud, and R. Qahwaji, "Recognition of off-line printed Arabic text using Hidden Markov Models," *Signal Processing*, vol. 88, no. 12, pp. 2902-2912, 2008.

[5] U.-V. Marti and H. Bunke, "The IAM-Database: an English Sentence Database for Offline Handwriting Recognition," *International Journal on Document Analysis and Recognition*, vol. 5, no. 1, pp. 39-46, Nov. 2002.

[6] S. Cha and S. Srihari, "Assessing the Authorship Confidence of Handwritten Items," in *Fifth IEEE Workshop on Applications of Computer Vision (WACV'00)*, 2000, pp. 42-47.

[7] L. Schomaker and L. Vuurpijl, "Forensic Writer Identification: A Benchmark Data Set and a Comparison of Two Systems (research report)," Nijmegen, 2000.

[8] M. Pechwitz, S. S. Maddouri, V. Märgner, N. Ellouze, and H. Amiri, "IFN/ENIT - Database of Handwritten Arabic Words," in *7th Colloque International Francophone sur l'Ecrit et le Document , CIFED 2002*, 2002, p. 129--136.

[9] H. El Abed and V. Märgner, "The IFN/ENIT-database - a tool to develop Arabic handwriting recognition systems," in *9th International Symposium on Signal Processing and Its Applications. ISSPA 2007.*, 2007, pp. 1-4.

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

[11] Y. Al-Ohali, M. Cheriet, C. Y. Suen, and M. B, "Databases for recognition of handwritten Arabic cheques," *Pattern Recognition*, vol. 36, no. 1, pp. 111-121, Jan. 2003.

[12] S. Al-ma'adeed, D. Elliman, C. A. Higgins, and J. Campus, "A Data Base for Arabic Handwritten Text Recognition Research," *Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR'02)*, 2002.

[13] M. Z. Khedher and G. Abandah, "Arabic character recognition using approximate stroke sequence," in *Third Int'l Conf. on Language Resources and Evaluation (LREC 2002)*, 2002, pp. 28-34.

[14] H. Alamri, J. Sadri, C. Y. Suen, and N. Nobile, "A Novel Comprehensive Database for Arabic Off-Line Handwriting Recognition Huda Alamri," in *Eleventh International Conference on Frontiers in Handwriting Recognition*, 2008.

[15] E. A. El-Sherif and S. Abdelazeem, "A two-stage system for Arabic handwritten digit recognition tested on a new large database," in *International Conference on Artificial Intelligence and Pattern Recognition*, 2007, pp. 237–242.

[16] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 66, 62, Jan. 1979.

[17] M. T. Parvez, "Arabic Handwritten Text Recognition," PhD Thesis, KFUPM, 2010.

[18] M. Wienecke, G. A. Fink, and G. Sagerer, "Toward automatic video-based whiteboard reading," *International Journal on Document Analysis and Recognition*, vol. 7, no. 2, pp. 188-200, 2005.

[19] T. Plötz and G. A. Fink, "Markov models for offline handwriting recognition: a survey," *International journal on document analysis and recognition*, vol. 12, no. 4, pp. 269-298, 2009.

[20] "HTK Speech Recognition Toolkit." .

[21] U.-V. Marti and H. Bunke, "Handwritten sentence recognition," in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, 2000, pp. 463-466.