KFUPM **H**andwritten **A**rabic **T**ex**T** (KHATT) database

Version 1.0 (September 2012 Release)

The database contains handwritten Arabic text images and its ground-truth developed for research in the area of Arabic handwritten text.

Overview:

- Forms written by 1000 different writers.
- Scanned at different resolutions (200, 300, and 600 DPIs).
- Writers are from different countries, gender, age groups, handedness and education level.
- Natural writings with unrestricted writing styles.
- 2000 unique paragraph images and their segmented line images (source text from different topics like arts, education, health, nature, technology).
- 2000 paragraph images containing similar text, each covering all Arabic characters and shapes and their segmented line images.
- Free paragraphs written by writers on any topic of their choice.
- Paragraph and line images are supplied with manually verified ground-truths.
- The database divided into three disjoint sets viz. training (70%), validation (15%), and testing (15%).
- Promote research in areas like writer identification, line segmentation, and binarization and noise removal techniques beside handwritten text recognition.

KHATT v1.0 Database Details

The current version of the database contains the following:

1. **Folder Name: FormImages_v1.0**

   This folder contains the grayscale images of the forms. Each form contains four pages. The first page contains the details on the writer. Pages two, three, and four contain the paragraphs written by a writer. The forms are saved as TIFF files at 300 DPI resolutions with LZW compression. Forms with 200 and 600 DPI resolutions are also available and can be provided on request.

Folder Name	Description	Size
Training	700 Scanned forms where each form has 4 pages.	2.4 GB
Validation	150 Scanned forms where each form has 4 pages.	523 MB
Testing	150 Scanned forms where each form has 4 pages.	503 MB

2. **Folder Name: ParagraphImages_v1.0**

   The version (1.0) of the paragraph images contains the following sub-folders:

   a. **Folder Name: FixedTextParagraphs**

      This folder contains 300 DPI grayscale images of paragraphs one and four. They are the fixed (similar text paragraphs) written by each writer two times i.e. paragraph one and paragraph four. Thus for the complete 1000 forms, the database contains 2000 fixed text paragraphs. Moreover the text of the paragraph contains all the unique shapes for all the Arabic letters.

Folder	Description	Size
Training	1400 Paragraph Images	348 MB
Validation	300 Paragraph Images	74.9 MB
Testing	300 Paragraph Images	73.8 MB

   b. **Folder Name: UniqueTextParagraphs**

      This folder contains 300 DPI grayscale images of paragraphs two and three. They are the unique (there is no repetition of text) text written by each writer. Each writer wrote two paragraphs with unique text. Thus for the complete 1000 forms, the database contains 2000 unique text paragraphs.

Folder	Description	Size
Training	1400 Paragraph Images	268 MB

Validation	300 Paragraph Images	58.3 MB
Testing	300 Paragraph Images	56.6 MB

3. **Folder Name: LineImages_v1.0**

   The version (1.0) of the line images contains the following sub-folders:

   a. **Folder Name: FixedTextLineImages**

      This folder contains 300 DPI binary images of the lines extracted from paragraphs one and four.

Folder	Description	Size
Training	4686 Line Images	65.3 MB
Validation	966 Line Images	13.7 MB
Testing	1042 Line Images	13.6 MB

   b. **Folder Name: UniqueTextLineImages**

      This folder contains 300 DPI binary images of the lines extracted from paragraphs two and three.

Folder	Description	Size
Training	4837 Line Images	60.2 MB
Validation	940 Line Images	12.0 MB
Testing	967 Line Images	11.5 MB

4. **Folder Name: GroundTruth_v1.0**

   The ground-truth folder contains the following files

   a. **Folder Name: FixedTextUnicodeTruthValues-v1.0**
      It contains the ground-truth in Unicode (Arabic text) at the line level for the lines from the fixed text paragraphs. It is distributed under folders training, validation, and test.

   b. **Folder Name: UniqueTextUnicodeTruthValues-v1.0**
      It contains the ground-truth in Unicode (Arabic text) at the line level for the lines from the unique text paragraphs. It is distributed under folders training, validation, and test.

   c. **Folder Name: FixedTextLinesLatinTransliteration**
      It contains the ground-truth in Latin transliteration at the line level for the lines from the fixed text paragraphs. It is distributed under three separate files for training, validation, and test respectively.

d. **Folder Name: UniqueTextLinesLatinTransliteration**
It contains the ground-truth in Latin transliteration at the line level for the lines from the unique text paragraphs. It is distributed under three separate files for training, validation, and test respectively.

e. **File: KHATTUnicodeTruthValues-v1.0.xlsx**
It contains the ground-truth in Unicode (Arabic text) for the whole database except the last page (containing paragraphs five and six).

5. Folder: KHATT-Pilot

The KHATT pilot contains the line images and their ground-truth. It was used for the pilot experimentation as reported in the paper: S. A. Mahmoud, I. Ahmad, M. Alshayeb, W. G. Al-Khatib, M. T. Parvez, G. A. Fink, V. Margner, and H. EL Abed, "KHATT: Arabic Offline Handwritten Text Database," In Proceedings of the 13th International Conference on Frontiers in Handwriting Recognition (ICFHR 2012), Bari, Italy, 2012, pp. 447-452, IEEE Computer Society.

Folder	Description	Size
LineImages	1633 line images	21.1 MB

a. **File: Groundtruth-Unicode.xlsx**
It contains the ground-truth of the line images in Unicode (Arabic text).
b. **File: Groundtruth-Latin-TrainingSet.txt**
It contains the Latin representation of the ground-truth for the training set based on the lookup-table as presented in the following section.
c. **File: Groundtruth-Latin-TestSet.txt**
It contains the Latin representation of the ground-truth for the test set based on the lookup-table as presented in the following section.

Arabic-Latin Lookup Table

Latin Label	Arabic Label	Latin Label	Arabic Label
hh	ء	sa	ص
am	آ	de	ض
ae	أ	to	ط
ah	إ	zha	ظ
aa	ا	ay	ع
ba	ب	gh	غ
ta	ت	fa	ف
teE	ة	ka	ق
th	ث	ke	ك
ja	ج	la	ل
ha	ح	ma	م
kh	خ	na	ن
da	د	he	ه
dh	ذ	wa	و
ra	ر	wl	ؤ
za	ز	ya	ي
se	س	ee	ى
sh	ش	al	ئ

n0	0	com	,
n1	1	qts	?
n2	2	exc	!
n3	3	dot	.
n4	4	bro	(
n5	5	brc	)
n6	6	fsl	/
n7	7	bsl	\
n8	8	equ	=
n9	9	hyp	-
atr	@	usc	_
col	:	scr	#
dbq	"	per	%

Note: Some insignificant number of other symbols like Latin characters was not transliterated.

Some Useful Statistics on KHATT Database

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**

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	

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**

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

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

References

1. S. A. Mahmoud, I. Ahmad, M. Alshayeb, W. G. Al-Khatib, M. T. Parvez, G. A. Fink, V. Margner, and H. EL Abed, “KHATT: Arabic Offline Handwritten Text Database,” In Proceedings of the 13th International Conference on Frontiers in Handwriting Recognition (ICFHR 2012), Bari, Italy, 2012, pp. 447-452, IEEE Computer Society.

2. Sabri A. Mahmoud, Irfan Ahmad, Mohammed Alshayeb, and Wasfi G. Al-Khatib, "A Database for Offline Arabic Handwritten Text Recognition", M. Kamel and A. Campilho (Eds.): ICIAR 2011, Image Analysis and Recognition Part II, LNCS 6754, pp. 397--406. Springer, Heidelberg (2011). DOI: 10.1007/978-3-642-21596-4_40