Muharaf (محرف): Manuscripts of Handwritten Arabic Dataset for Cursive Text Recognition

Mehreen Saeed, Adrian Chan, Anupam Mijar, Joseph Moukarzel, Georges Habchi, Carlos Younes, Amin Elias, Chau-Wai Wong, Akram Khater

NC STATE UNIVERSITY Moise A. Khayrallah Center for Lebanese Diaspora Studies

NC STATE UNIVERSITY Electrical and Computer Engineering

USEK

LAH Lebanese Association for History الهيئة اللبنانية للتاريخ

NEURAL INFORMATION PROCESSING SYSTEMS

Manuscripts of Handwritten Arabic (Muharaf) dataset is a machine learning dataset of more than 1,600 historic handwritten page images transcribed by experts in archival Arabic with annotated text lines and various page elements.

Motivation

- Modern Standard Arabic: ~400 million native speakers.
- Massive existing collections, e.g., British library has ~15,000 works in ~14,000 volumes of Arabic manuscripts.
- Limited Arabic resources for handwritten text recognition (HTR) of scanned historic manuscripts.
- Most historic Arabic datasets in neat calligraphic writing.

Muharaf dataset

- “Muharaf” is Arabic for “typeface”.
- Historic Arabic manuscripts: 1800-2011.
- Images: 1,644.
- Text lines: 36,311.
- Text regions: 4,867 (main text, headings, floating).
- Muharaf-Public images: 1,216.
- Muharaf-Restricted images: 428.

Sample images from Muharaf dataset from (a) USEK Al Batroun Collection, (b) USEK Joseph El Hachem Collection, (c) KCLDS El-Khouri Collection, (d) KCLDS Nasrallah Collection, (e) USEK Papers of Kfarchima Municipality Collection, (f) KCLDS K. Joseph Collection, (g) USEK Amin Farhat Collection, (h) KCLDS T. Attallah Collection.

Contribution

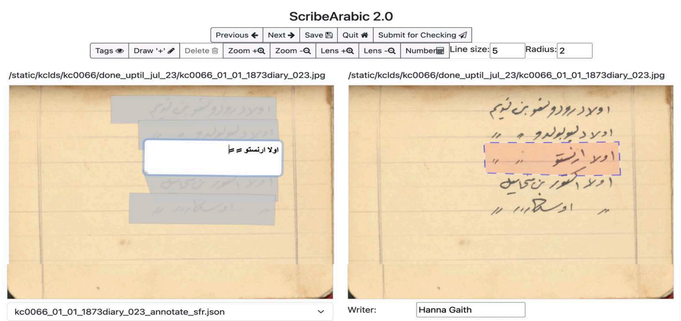
- Largest publicly available historic Arabic HTR dataset.
- Scanned manuscripts (not scribed like IAM [1], KHATT [2]).
- Ruq‘ah: casual/informal style of writing.
- Not calligrahic like RASM [3], RASAM [2].
- Type of manuscripts: letters, notes, poems, dialogs, legal records, church documents, official correspondences.
- Uses: handwritten text recognition, text-line segmentation, layout detection, writer identification.
- Challenges: ink bleeds, crossed-out text, barely legible handwriting, damaged paper.

Dataset	Page Count	Text Regions	Line Count
IAM [1]	1,539	1,539	13,353
RASAM [2]	300	676	7,540
RASM [3]	120	132	2,613
KHATT [4]	4,000	4,000	13,435
Muharaf-public	1,216	3,479	24,495
Muharaf-restricted	428	1,388	11,816
Muharaf	**1,644**	**4,867**	**36,311**

Comparison of Muharaf and other publicly available datasets.

ScribeArabic

Custom software used to annotate and transcribe images



Screenshot of ScribeArabic software.

Expert transcription team

- History professor from Lebanese Association for History.
- Two expert Arabic manuscript archivists at USEK.
- QA: history professors at USEK and KCLDS.

Funded by

NATIONAL ENDOWMENT FOR THE HUMANITIES

FamilySearch

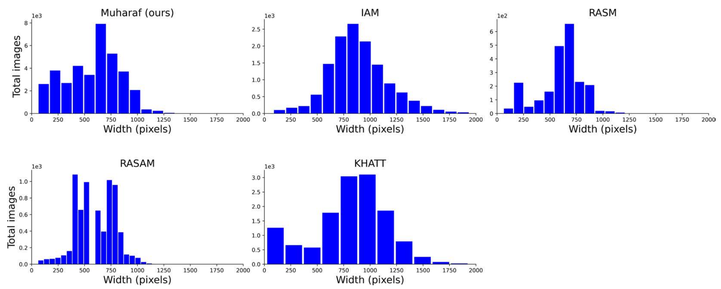
NC STATE UNIVERSITY Moise A. Khayrallah Center for Lebanese Diaspora Studies

NC STATE UNIVERSITY Electrical and Computer Engineering

FAIN: ZPA-283823-22

Packaged dataset

- Images (JPG) + Ground truths (JSON + PAGE-XML).
- Compatible with: PAGE-XML viewer [5] + Aletheia tool [6].
- Separate directory with warped, preprocessed line images and ground truths.



Comparison of the distribution of width of line images of Muharaf and other publicly available datasets.

Page elements

- Annotated page elements:
  - Main paragraph regions.
  - Floating text regions.
  - Page numbers.
  - Signature areas.
  - Graphics (logos, stamps).
- Page elements viewer included.

Viewer of different page elements.

Experimental results

Results using CNN based start, follow, read network [7]

Dataset	Split (Train, Validate, Test)	Level	CER	WER
Muharaf-public	(1100, 50, 66)	Page	0.157 ± 0.008	0.398 ± 0.007
		Line	0.181 ± 0.009	0.430 ± 0.011
Muharaf	(1500, 50, 96)	Page	0.134 ± 0.007	0.353 ± 0.012
		Line	0.149 ± 0.004	0.380 ± 0.004

References

[1] Urs-Viktor Marti and Horst Bunke. The IAM-database: An English sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5:39–46, November 2002.

[2] Chahan Vidal-Gorène, Noëmie Lucas, Clément Salah, Aliénor Decours-Perez, and Boris Dupin. RASAM – A dataset for the recognition and analysis of scripts in Arabic maghrebi. In Elisa H. Barney Smith and Umapada Pal, editors, *Document Analysis and Recognition – ICDAR 2021 Workshops*, Cham, 2021. Springer International Publishing.

[3] Christian Clausner, Apostolos Antonacopoulos, Nora Mcgregor, and Daniel Wilson-Nunn. ICFHR 2018 competition on recognition of historical Arabic scientific manuscripts — RASM2018. In *Proceedings of the 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2018.

[4] Sabri A. Mahmoud, Irfan Ahmad, Mohammad Alshayeb, Wasfi G. Al-Khatib, Mohammad Tan- vir Parvez, Gernot A. Fink, Volker Märgner, and Haikal El Abed. KHATT: Arabic offline handwritten text database. In *Proceedings of the International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 449–454, 2012.

[5] Christian Clausner, Stefan Pletschacher, and Apostolos Antonacopoulos. Aletheia — An advanced document layout and text ground-truthing system for production environments. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 48–52, 2011.

[6] Pattern Recognition and Image Analysis Research Lab (PRImA). PAGE XML for page content. https://www.primaresearch.org/schema/PAGE/gts/pagecontent/ 2019-07-15/pagecontent.xsd. [Last Accessed: 13 November 2024].

[7] Curtis Wigington, Chris Tensmeyer, Brian Davis, William Barrett, Brian Price, and Scott Cohen. Start, Follow, Read: End-to-end full-page handwriting recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.