



Moise A. Khayrallah Center for
Lebanese Diaspora Studies



Electrical and
Computer Engineering



LAH

Lebanese Association for History
الهيئة اللبنانية للتاريخ

محرّف

Muharaf: 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

Funded by



NATIONAL
ENDOWMENT
FOR THE
HUMANITIES

FAIN: ZPA-283823-22



Moise A. Khayrallah Center for
Lebanese Diaspora Studies



Outline

- Muharaf dataset
- Contribution
- Data collection
- Handwritten text recognition (HTR) results
- Future directions

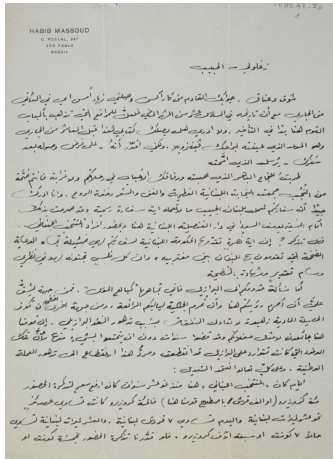
Muharraf Dataset

- Purpose: Train handwritten text recognition system for historic Arabic manuscripts with casual writing styles
- Composition: Annotated and transcribed text lines
- Muharraf: 1,644 images (1,216 public, 428 restricted)

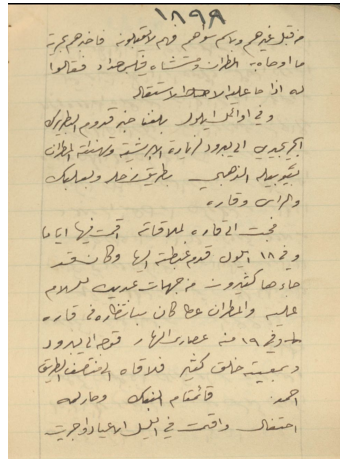
- Time period: 1800-2011
- Text lines: 36,311
- Text regions: 4,867



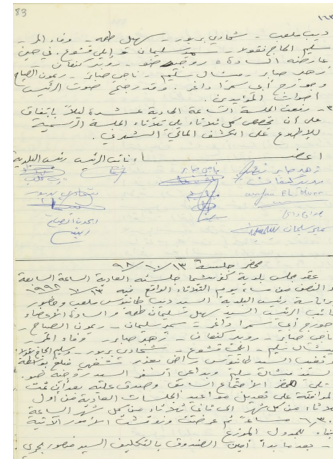
Al Batroun Collection



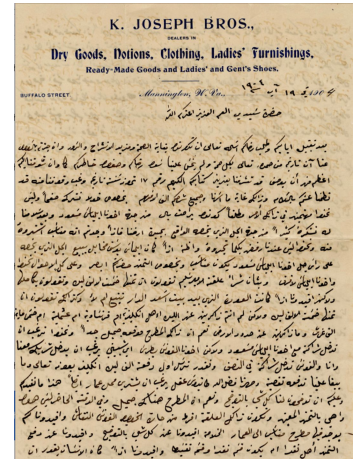
Joseph El Hachem Collection



Hanna Gaith Collection



Papers of Kfarchima Municipality



K. Joseph Collection

Contribution

- Largest publicly available historic Arabic dataset for handwritten text recognition
- Unlike IAM and KHATT (scribed under controlled experimental conditions)
- Ruq'ah: Casual/informal style of writing
- Uses: HTR, text-line segmentation, layout detection, writer identification
- Challenges: Ink bleeds, crossed-out text, barely legible handwriting, torn paper

Dataset	Page Count	Text Regions	Line Count
IAM	1,539	1,539	13,353
RASAM	300	676	7,540
RASM	120	132	2,613
KHATT	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

Experimental Results

- CNN based Start, Follow, Read (Wigington et al., 2018) - Adapted for Arabic
- End-to-end full page handwritten text recognition

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

Future Work

- Identification of writers
- Classification of styles
- Train more state-of-the-art HTR systems
- Extract linguistic knowledge and identify historic colloquial form of Arabic

Thank you!

Reach out to us for any questions/comments on Github:

<https://github.com/MehreenMehreen/muharaf>

