

NC STATE Electrical and UNIVERSITY Computer Engineering







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









Outline

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





Muharaf Dataset

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

Time period: 1800-2011

- Text lines: 36,311
- Text regions: 4,867

لت مدة المان سان من داد من شد تركيد الدي مع دلالة بنا بعد مد وكن يداد در ميا شاند المان كان بنك بن درم ما الله ترت ما عالما تله من من من الله من الله من من ما الله ترت من من من من من من من من ما الله من من ما ما ما من من من م	الم خاط محاط المراحين من جامع الموالسين من جامع الموالسين من حيث المرك المحاط المراحين من حيث المرك المحاط من من حيث المرك من من حيث المرك المحاط من من حيث المرك من من حيث المرك من الموالي من المرك من الموالي من من حيث المحاط من من من حيث المحاط من من من حيث المحل من من من حيث المحل من حيث المحل من حيث المحل من حيث المحل من حيث الموالي من من حيث الموالي من من حيث الموالي من من حيث الموالي من	علم والمطان عط كان سائل والام الملب والمطان عط كان سائل من قرار المسية حامت كمشر فلاقاء المنت مناطرت جمد قائمة > المنبك وحارف استنال واقعت في المول الما والحرب استنال واقعت في الول الما والحرب	مرافعت من مع القدار النام في في محت معت و لائم الما يسلم المعان مع القد السعون على المعان معت معت و لائم المعان المعان المعرب المعان المعان معان المعان المعان المعان المعان المعان المعرب المعان على المعان ال	ملاشر هما الفن العديم التوابيد بيت أحد الدائر بيني ما كركك تعليما الم الولي منافع مدين ومكن المرتكرين عن يون المواكلين متناه (مطلق المعظمة المع الولي منافع المع المع المدين المدين الموالي عن المجلي في الما ويعلم يحصل ان الملك تعليما في الحلق رفته المعالين عن المجلي في المع المع مليك مح المعال المع المع المع المع العلم ومنافع المع المع ملي المع المع المع من المترفين الماكن المعالي والمعاص ومن المدين المدين المع المع المع من المترفين المحالي المعاد العلم ومنافع المع المع المع المع من المترفين المع المعال المعال المع المع على المع المع المع من المول علمة المحمد المحال المعاد المعاد على المحمل المع المع من المول علمة المحمد المع المع المع المع المع المع المع المع
مع معن من من مع مي الا سانة الافتده - معن من من مورسية عرفيه من ما رفته مع من من من من موتار والفار سمت مع اليدة الله من مدار من من ما ما من سمة اليدة العلم من مع من من من ما ما من سمة مع الات كل من مع من من من ما من الما من مع مع من من من ما من الما والما من مع مع من من من ما من ما من الما من مع من من من من ما من الما من من مع من من من ما من ما من الما من مع من من من من من من من من من مع من من من من من من من من من من مع من من من من من من من من من من مع من من من من من من من من من من مع من م	تغلاب البيب - تعادی مادند . ماکند المبی - ماد مادند . ماکند المبا ماد ماد ماد . ماد ماد . ماد ماد ماد . ماد ماد با ماد ماد ماد ماد ماد ماد ماد ماد . ماد ماد ماد با ماد ماد ماد ماد ماد . ماد ماد ماد ماد ماد ماد ماد ماد . ماد ماد ماد ماد ماد ماد . ماد ماد ماد ماد ماد ماد . ماد ماد ماد ماد ماد ماد ماد . ماد ماد ماد ماد ماد ماد ماد . ماد ماد ماد . ماد ماد ماد . ماد ماد ماد ماد . ماد ماد ماد ماد . ماد ماد ماد ماد . ماد ماد ماد . ماد ماد ماد . ماد ماد ماد . ماد	مترفيم ريم مرم فه مترم فاخرم . ما وحاج الملار يحت ، فلهما ، فاخرم . به اذا ما علا موصل يستلم و فاداتل يول من مر قري المراك المرجعي اليور المادة المرتبة قائلة المان و مراب وقار . فريسا المرس علم تعام ولا المان و فراد المراد علام . فريسا المرس في مرجع ولا المان و فراد المراد علام . و فراد المراد علم المان المراح و فراد المراد مرجل حسب مرب ماد حال مشرو مرجل حسب مرب	مع معلمه مع معادي روب مراس مع مع د مع	د. بالمحمد المحمد ملمحمد ملمد المحمد المحمد المحمد المحمد المحمد المحمد المحمد المحمد المحمد المحم

K. Joseph Collection

Al Batroun Collection

- لاقع iel'as 2ê e مقرق

ارد المراج

مرد في أورة ひんでしょ

Hanna Gaith Collection

Joseph El Hachem Collection

Contribution

bise A. Khayrallah Center for NC STATE

Electrical and Computer Engineering

• Largest publicly available historic Arabic dataset for handwritten text recognition

Lebanese Association for

- 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

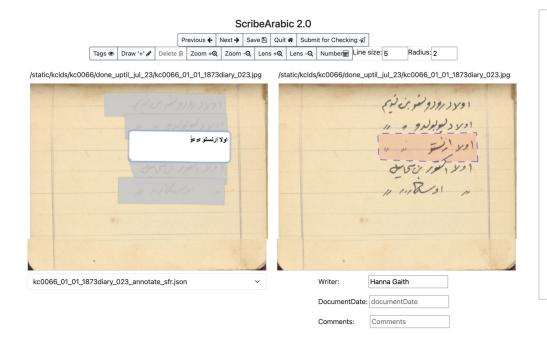
Dataset	Page Count	Text Regions	Line Count
IAM	1,539	1,539	13,353
RASAM	300	676	7,540
RASM	120	132	2,613
KHATT	4,000	4,000	13,435
Muharaf-public	1,216	3,479	24,495
Muharaf-restricted	428	1,388	11,816
Muharaf	1,644	4,867	36,311

محرف Muharaf





ScribeArabic - Software for Data Collection



Lebanese Association for His

Moise A. Khayrallah Center for NC STATE

banese Diaspora Studies

Electrical and

Computer Engineering

Transcription Team

- History professor from Lebanese Association for History.
- Two expert Arabic manuscript archivists at Holy Spirit University of Kaslik (USEK).
- QA: History professors at USEK and NC State.





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 Line	$\begin{array}{c} 0.157 \pm 0.008 \\ 0.181 \pm 0.009 \end{array}$	$\begin{array}{c} 0.398 \pm 0.007 \\ 0.430 \pm 0.011 \end{array}$
Muharaf	(1500, 50, 96)	Page Line	$\begin{array}{c} 0.134 \pm 0.007 \\ 0.149 \pm 0.004 \end{array}$	$\begin{array}{c} 0.353 \pm 0.012 \\ 0.380 \pm 0.004 \end{array}$





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/Mehreen/Mehreen/muharaf

