



Technische
Universität
Braunschweig



Institut für Nachrichtentechnik



An Historical Handwritten Arabic Dataset for Segmentation-Free Word Spotting – HADARA80P

Werner Pantke, September 2, 2014

Table of Contents

1. Motivation
2. Dataset
3. Evaluation
4. Conclusion



Table of Contents

1. Motivation
2. Dataset
3. Evaluation
4. Conclusion

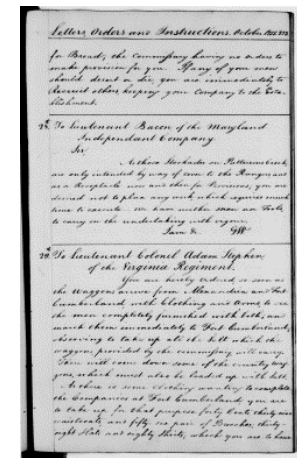
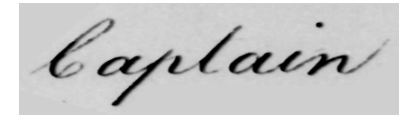


Motivation (1/2)

- Word Spotting (WS)
 - task of **searching** for specific words in images
 - w/o necessarily recognizing the textual content
- query type
 - **string**
 - example image (**template**)
- operating **level**
 - segmentation-based
 - requires preceding segmentation
 - possible levels: text lines, words, ...
 - segmentation-free
 - w/o **any** segmentation
 - emerging trend
 - need for datasets to develop and evaluate these WS systems



Captain



Source: George Washington Dataset [Lavrenko2004]

Motivation (2/2)

- Context: HADARA project
 - Team: computer scientists, linguists, and historians
 - Goal:
 - **digitization** of hard accessible books
 - development of tools to **process** and archive scanned manuscripts
 - Focus: analysis of **historical handwritten Arabic** documents
- Existing non-segmented WS datasets
 - only **few** available w/ word coordinates
 - mostly for **Latin** scripts (languages such as Latin, English, Medieval German)
 - found none for Arabic

Table of Contents

1. Motivation
- 2. Dataset**
3. Evaluation
4. Conclusion



Dataset

Characteristics of Arabic Language

- Diacritics

- mandatory (req. to distinguish characters)
- optional (e.g., short vowels)

ب *b* ت *t*
بَ *ba*

- Character shape depends on

- position in a word (begin, middle, end, isolated)

ي ي ي ي *y*

- Suffixes and prefixes

- person, number, tense, case
- some prepositions and conjunctions („in“, „for“, „and“, ...)

- Consider substring matches?

- keyword present, but prefixed/suffixed – still relevant?

طاعون *tāʿūn*

الطاعون *alṭāʿūn*

Dataset

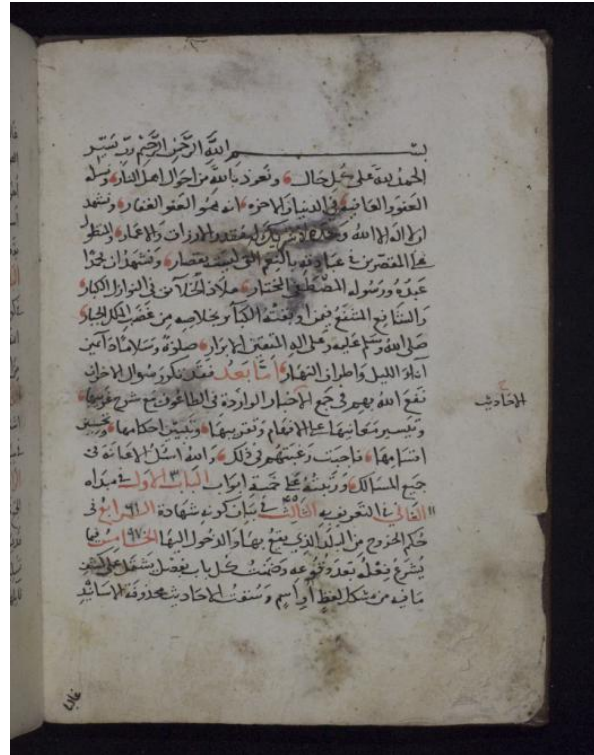
Underlying Manuscript (1/2)

- Author from Egypt/Palestine
- Title **بَدْلُ الْمَاعُونِ فِي فَضْلِ الطَّاعُونَ** *badlu almā'ūn fi faḍlu alṭā'ūn*
- Approx. 250 text pages (5 chapters)
- Published in
 - 06. 833 AH (Islamic calendar)
 - Feb. 1430 AD (Gregorian calendar)

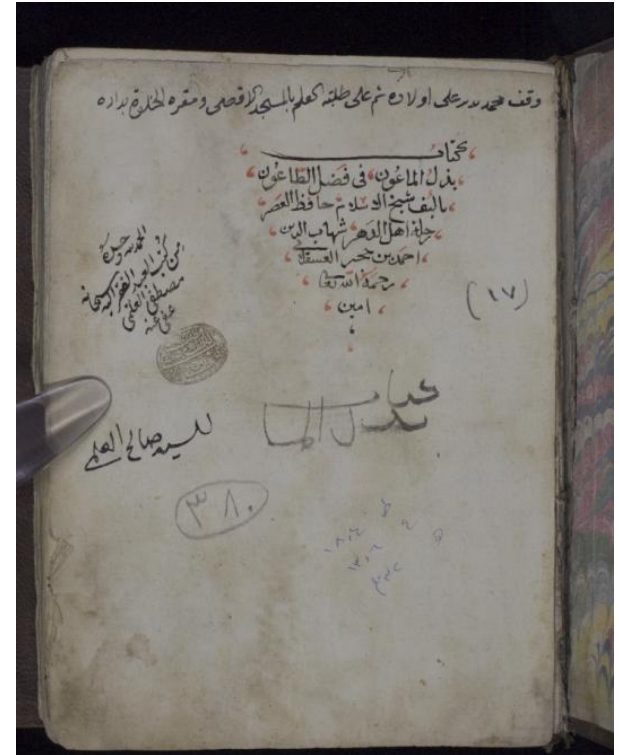
Dataset

Underlying Manuscript (2/2)

- 1st page: metadata
- 2nd page: typical page
- Few side notes
- Degraded paper
- Text color
 - black
 - red



Page 2



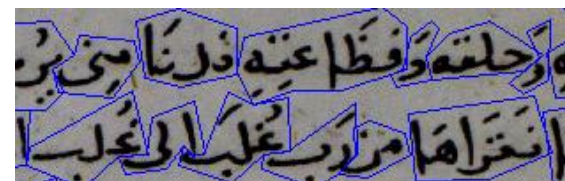
Page 1

Dataset

Data Acquisition

- Scanning
 - **Full-frame** CMOS camera w/ **normal prime lens**
 - 1 page/image, **~380 dpi**
 - **48-bit TIFF** images w/ 16 bits per color channel (true 12 bits per color channel)
 - **Lossless** compression (TIFF deflate) -> ~51 Mbytes/image

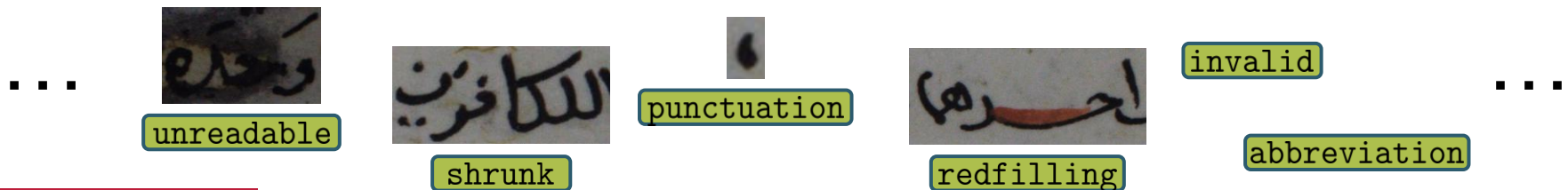
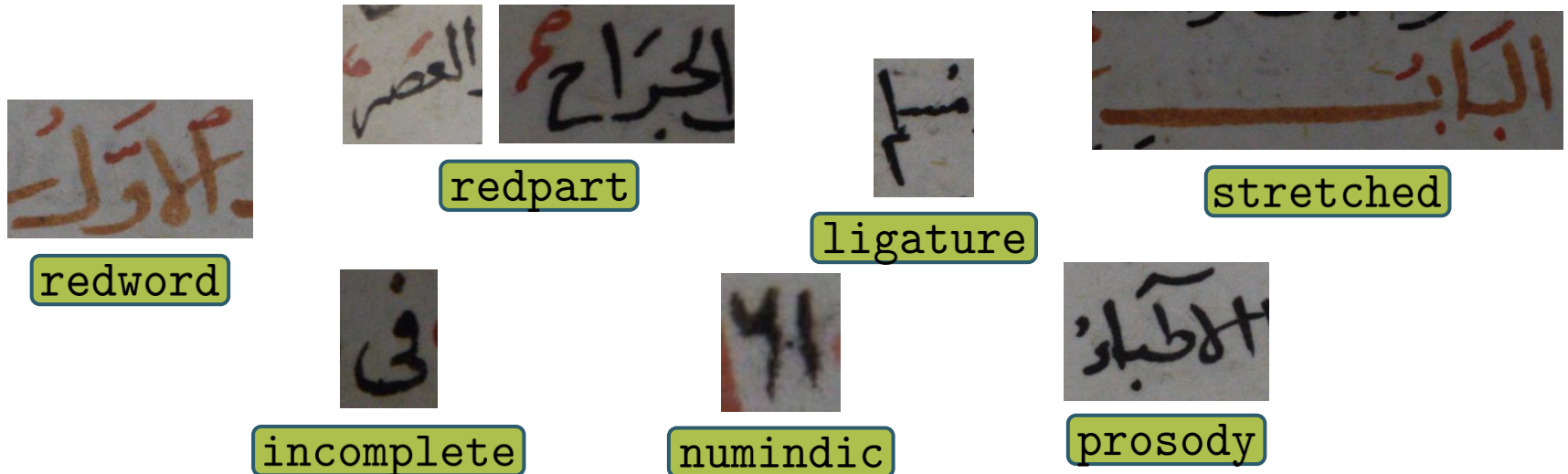
- Annotation (ground truth)
 - XML-based data format
 - Word segments (**polygons**)
 - contain all corresponding diacritics
 - contain no strokes from other words (wherever possible)
 - Narrow transcription (UTF-8)



Dataset

Word-Level Tags

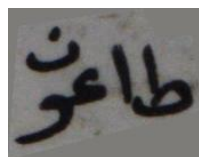
- Developed set of 18 tags
- May be used to focus WS development/evaluation on specific issues



Dataset

Pre-defined Small Keyword Set

- Full dataset evaluation often computationally not feasible
- For fast **comparison of WS systems**:
 - 25 pre-defined keywords
 - different properties
 - # occurrences in the text (many, few)
 - text color (black or red)
 - linguistic properties



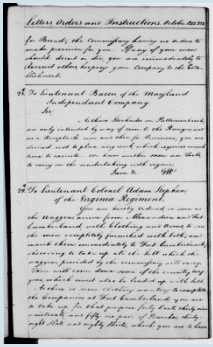

shrunk



redword



Dataset Comparison

| | GW20P [Lavrenko2004] | HADARA80P |
|-----------------|--|---|
| example image |  |  |
| color depth | 8-bit grayscale | 48-bit color (true 36 bits) |
| language/script | English/Latin | Arabic/Arabic |
| # pages | 20 | 80 (79 text pages) |
| # words | 4856 | 16935 (w/o side notes) |
| word details | transcription rectangles | transcription polygons detailed info by tags |

[Lavrenko2004] V. Lavrenko, T. M. Rath and R. Manmatha: Holistic Word Recognition for Handwritten Historical Documents. In: Proc. of the Int'l Workshop on Document Image Analysis for Libraries (DIAL), Palo Alto, CA, January 23-24, 2004, pp. 278-287.

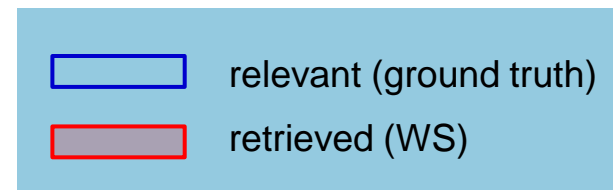
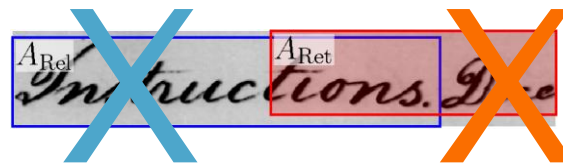
Table of Contents

1. Motivation
2. Dataset
- 3. Evaluation**
4. Conclusions



Evaluation Protocol

- Evaluation protocol of [Pantke2013], ensuring stable and fair results
- Traditional recall and precision (p_{IR}) measures
 - relevance criterion (here): hard overlap threshold of 1 pixel w/ relevant area
- Precision measure $\overline{\gamma_{LA}}$ w/o hard decision
 - developed for segmentation-free WS
 - judging location quality γ_{LA} of a retrieved word
 - continuous, ranging from 0 to 1
 - punishing **non-retrieved** and **non-relevant** areas



[Pantke2013] W. Pantke, V. Märgner, and T. Fingscheidt, “On evaluation of segmentation-free word spotting approaches without hard decisions,” in Proc. Int. Conf. Document Analysis and Recognition (ICDAR 2013), Washington DC, USA, 2013, pp. 1300–1304.

Evaluation Results

- HADARA word spotting system
- Mean average precision (MAP) over all queries

| MAP using | GW20P | HADARA80P |
|--------------------------|-------|-----------|
| p_{IR} | 0.61 | 0.41 |
| $\overline{\gamma_{LA}}$ | 0.44 | 0.31 |

- Results
 - HADARA80P may be a **more complex** task than GW20P
 - Soft precision measure $\overline{\gamma_{LA}}$ punishes (here) too small retrieved word areas

Table of Contents

1. Motivation
2. Dataset
3. Evaluation
- 4. Conclusion**



Conclusion

- HADARA80P dataset
 - 80 pages from historical handwritten **Arabic** manuscript
 - for development and evaluation of **segmentation-free** word spotting systems
 - high resolution and color depth, **polygons**
 - additional information by word-level tags
 - pre-defined set of 25 keywords

HADARA80P will be free of charge for research

Thank you.

Werner Pantke

pantke@ifn.ing.tu-bs.de



Technische
Universität
Braunschweig



Institut für Nachrichtentechnik