MUNI
FI

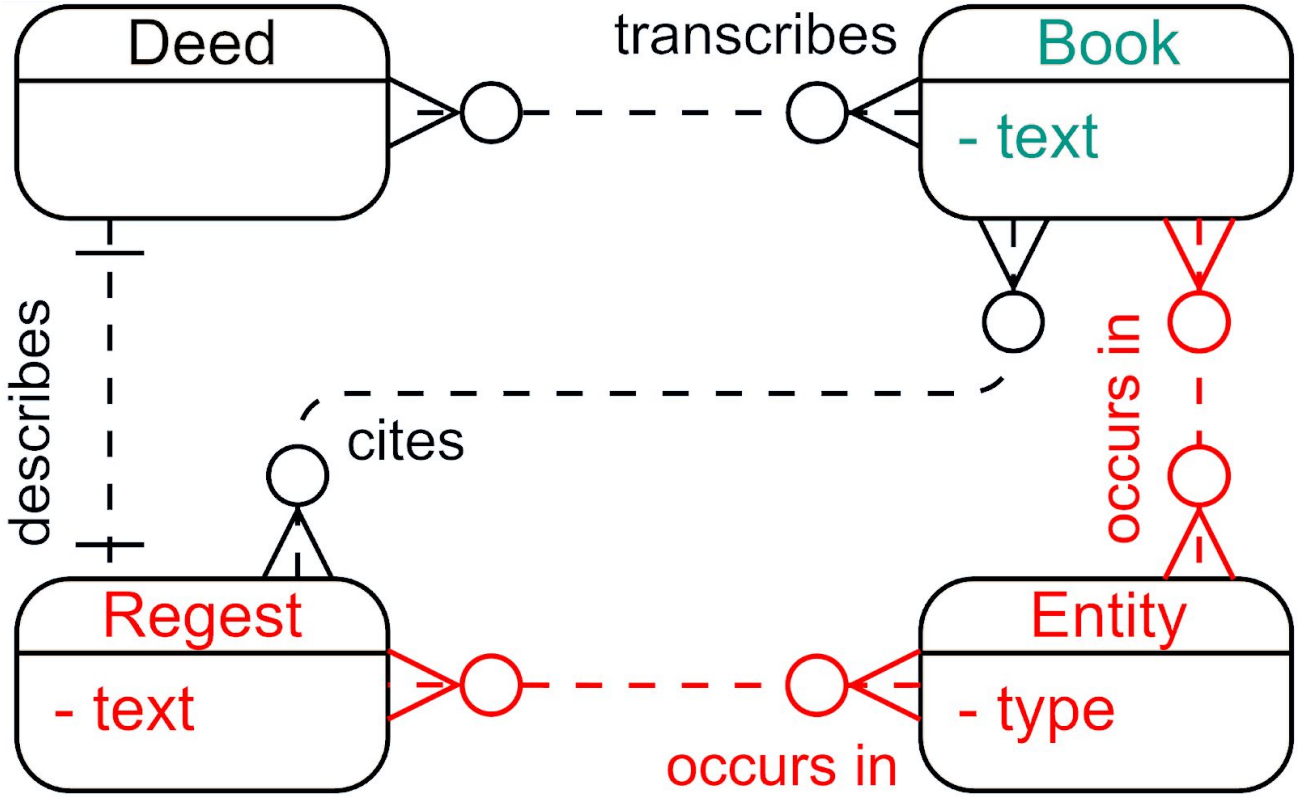
When Tesseract Meets PERO: Open-Source Optical Character Recognition of Medieval Texts

Vít Novotný and Aleš Horák

2022-12-10, RASLAN

Introduction

Entity-Relationship Diagram of AHISTO



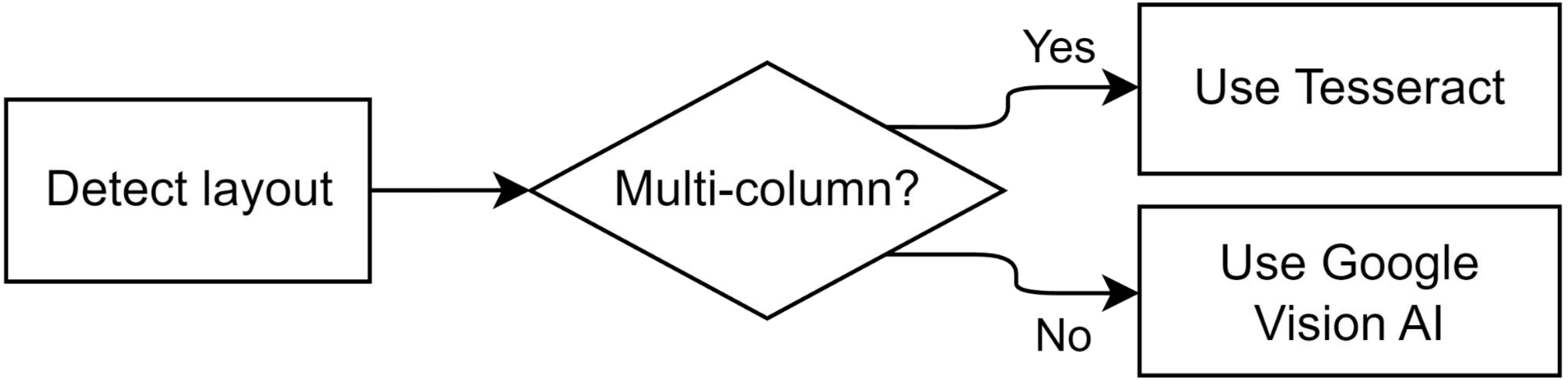
- 814 *books* transcribe medieval *deeds*.
- During 2020–2021, we focused at recognizing text in scanned book *pages*:
  - NOVOTNÝ. When Tesseract Does It Alone. In Horák et al. *RASLAN 2020*. [8]
  - NOVOTNÝ et al. When Tesseract Brings Friends. In Horák et al. *RASLAN 2021*. [10]
- Meanwhile, our colleagues from ARTS wrote 2 094 *regests* that describe *deeds*, denote *entities* (people and places) that occur in deeds, and cite related *books*.
- In 2022, we focused at retrieving *entities* from *regests* in *book* texts and recognizing new *entities* in *book* texts.

MUNI FI

Introduction

Previous AHISTO OCR system

- Previously [10], we combined free Tesseract with Google Vision AI:



- **Problem:** Closed-source Google Vision AI is difficult to reproduce.

MUNI FI

Related Work

PERO OCR system

- At ICDAR 2021, Michal Hradiš et al. introduced PERO OCR:
  - KODYM et al. Page layout analysis system for unconstrained historic documents. [5]
  - KIŠŠ et al. AT-ST: Self-training adaptation strategy for OCR in domains with limited transcriptions. [4]
  - KOHÚT et al. TS-Net: OCR trained to switch between text transcription styles. [6]

- PERO OCR is available as web demo and open-source at GitHub.

MORAVSKÁ ZEMSKÁ KNIHOVNA

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

MINISTERSTVO KULTURY

MUNI FI

Methods

OCR system combinations

- We replace Google Vision AI with web demo / GitHub PERO OCR.
- Previously [10], we used Google Vision AI from 2020-10-02.
- For fair comparison, we also use Google Vision AI from 2022-08-11.
- As baselines, we also use Google Vision AI and PERO OCR alone.

Quantitative evaluation

- We evaluate word error rate (WER) on 110 human-annotated pages.
- As previously, we lower-case and deaccent text before WER.

Results

Quantitative evaluation

- Google Vision AI (2022) is significantly better at multi-column pages.
- Two variants of PERO OCR achieve different WER, GitHub is worse.
- Replacing Google Vision (2022) with PERO (GitHub) improves WER.

	Google Vision AI		**PERO OCR**		**AHISTO OCR**	
	2020-10-02	*2022-08-11*	*Web demo*	*GitHub*	*with Google*	*with PERO*
Single-column	4.88%	3.79%	2.83%	**2.08%**	3.79%	**2.08%**
Multi-column	78.35%	10.52%	31.51%	49.38%	**7.43%**	9.93%
All pages	16.23%	4.83%	7.26%	9.39%	4.35%	**3.29%**

MUNI FI

Conclusion

- PERO OCR makes AHISTO OCR reproducible and more accurate.
- We release open-source AHISTO OCR system [9] at GitLab FI MU.
- We release open dataset [11] of outputs from different OCR systems.


FACULTAS ARTIS INFORMATICÆ
UNIVERSITAS MASARYKIANA