MUNI FI | MUNI ARTS

When Tesseract Brings Friends

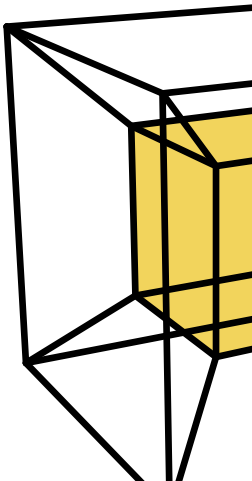
Layout Analysis, Language Identification, and Super-Resolution in the Optical Character Recognition of Medieval Texts

Vít Novotný[1] **Kristýna Seidlová**[2] **Tereza Vrabcová**[1] **Aleš Horák**[1]
{witiko,449852,485431}@mail.muni.cz, hales@fi.muni.cz

[1]NLP Centre, Faculty of Informatics, Masaryk University

[2]Auxiliary Historical Sciences and Archive Studies, Faculty of Arts, Masaryk University

December 10, 2021



Introduction

- The AHISTO project makes documents from the Hussite era publicly available online.
- Accurate OCR is required to extract actionable *texts* from *scanned images*.
- Previously, we showed [1] that *Tesseract 4* was the most accurate of five OCRs.
- We study the impact of eight preprocessing techniques on the accuracy of Tesseract 4.
- We publish dataset of scanned images, OCR texts, and *annotations* for three tasks.

Related Work

- OCRs *preprocess* scanned images to fix artefacts of:
  1. printing process
  2. scanning process
  3. aging of paper
- We additionally preprocess scanned images to:
  4. adapt to strengths and weaknesses of Tesseract 4
  5. add missing information to low-quality scans
- We use preprocessing techniques based on:
  - layout analysis [2–12]
  - language identification [13, 14]
  - image super-resolution [15–22]

Methods

Optical Character Recognition

- Besides Tesseract 4, we also use Tesseract 3, Tesseract 3 + 4, and Google Vision AI:
  - as baseline OCRs
  - for *fusion of several OCRs* (on the next slide)
- For more information about different OCRs, see our previous article [1, Section 2].

Scanned Image Dataset

- Previously, we developed a dataset [1, Section 3.1] of 65,348 scanned images.
- For *reproducibility*, we use a subset of 51,351 images from *public-domain books*:

```
~$ wget https://lindat.mff.cuni.cz/repository/xmlui/bitstream/
   /handle/11234/1-4615/scanned-images.zip
Length: 50609673540 (47G) [application/zip]
Saving to: ‘scanned-images.zip’

scanned-images.zip  0% [>     ]  299.10M  73.1MB/s    eta 9m 6s
```

Methods

Layout Analysis

- Previously, we showed [1, Section 4.2]:
  - Google Vision AI can be more accurate than Tesseract 4.
  - Google Vision AI fails to properly segment *multi-column pages*.
- We developed two techniques to decide whether a page is single- or multi-column:

a) Geometry

3 multi-column rays

4 single-column rays

Probably single-column

b) Machine learning

boundaries of lines

remove outliers

3 clusters of x-coords

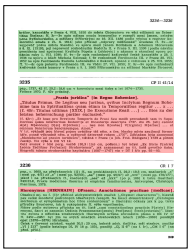
Probably multi-column

- We use our techniques to decide whether a page is single- or multi-column.
- Single-column pages are processed by Google Vision AI, multi-column by Tesseract 4.

Methods

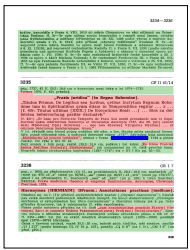
Language Identification

- Panák [23, Section 4.4] showed that *two-pass processing* can improve OCR accuracy:
  1. We use OCR to *identify languages*.
  2. We use OCR with identified languages.
- We developed two techniques to identify the languages in a page:



a) Paragraph-based
Pr(Czech) = 60%
Pr(Latin) = 40%

More robust
Ignores lone words



a) Word-based
Pr(Czech) = 55%
Pr(Latin) = 40%
Pr(German) = 5%

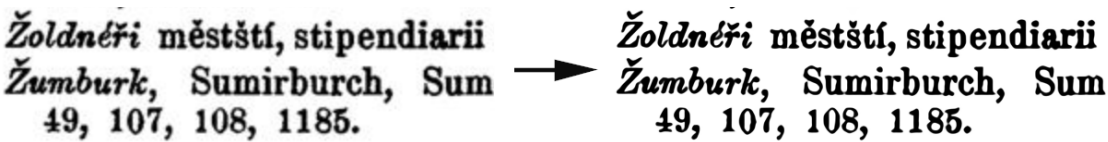
Detects more langs

- In first pass, we identify languages using *three* or *nine* hand-picked language models.
- In second pass, we use languages L, where $\Pr(L) \geq \textit{threshold}$ for different thresholds.

Methods

Image Super-Resolution

- Many of our scanned images are only available in *low resolution*.
- We use baseline techniques to upscale them:
  1. bilinear interpolation
  2. Potrace vectorizer [24]
- We also use *image super-resolution techniques*:



- We used four different models for image super-resolution:
  3. two pre-trained SRCNN models: Waifu2x with *low* and *high noise removal*
  4. two SRGAN models: pre-trained and trained on a PDF book of medieval texts [25]

Methods

Evaluation

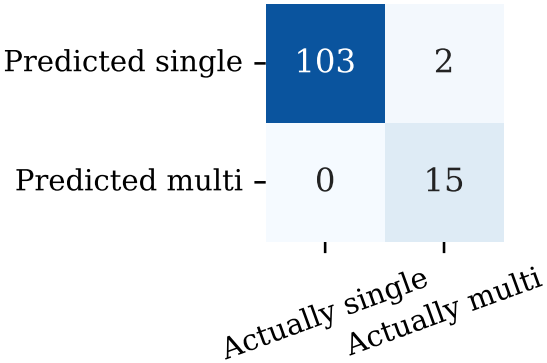
- We evaluate our preprocessing techniques:
  - intrinsically on the *layout analysis* and *language identification* tasks
  - extrinsically on the *OCR accuracy*
- For layout analysis, we report *confusion matrices* for single- and multi-column classification.
- For language identification, we report the percentage of pages (*Accuracy@1*), where we correctly guessed the primary language of a page during the first pass.
- For OCR accuracy, we report the percentage of words, which the OCR did not guess correctly (*word error rate*), averaged over all pages.

- For reproducibility, we publish the human annotations for all three tasks.

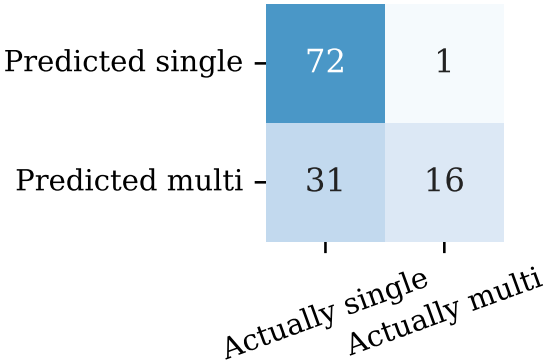
Results

Layout Analysis I

- Simpler *geometry* technique only misclassified two out of 120 pages.



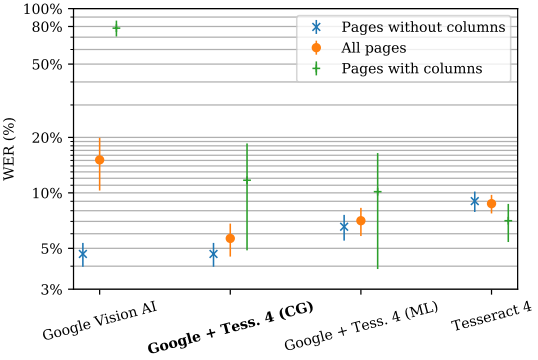
- *ML* technique misclassified 31 / 103 single-column pages as multi-column.



Results

Layout Analysis II

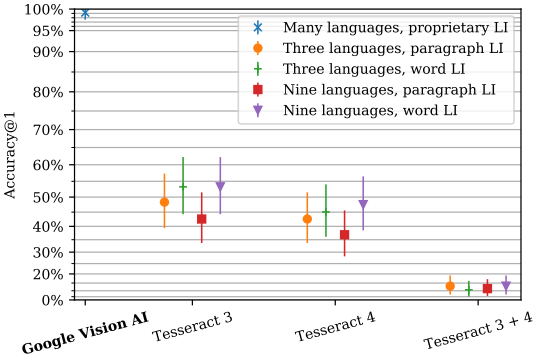
- *Google Vision AI* performs significantly better than Tesseract on single-column pages, but fails catastrophically on multi-column pages.
- We receive significant improvements to OCR accuracy by combining *Google Vision AI* and Tesseract using the geometry layout analysis technique.



Results

Language Identification I

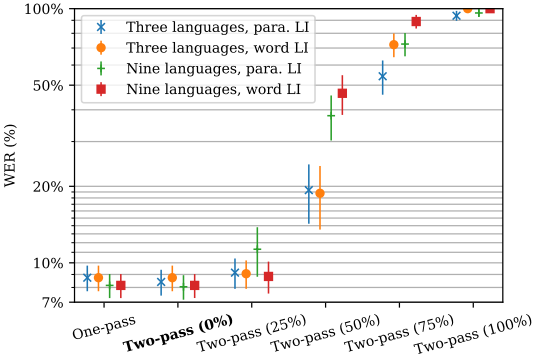
- *Google Vision AI* performed much better than Tesseract on language identification.
- *Tesseract 3* performed slightly better than *Tesseract 4*.
- For Tesseract 3, using *nine* language models with the *word-based* language identification technique consistently outperformed other configurations.



Results

Language Identification II

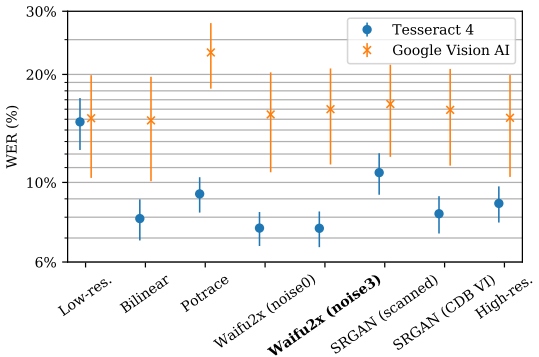
- Using *two-pass processing*, *nine language models*, *paragraph-based* language identification technique, and *0% threshold* improved the OCR accuracy of *Tesseract 4*.
- The 0% threshold is very *conservative*: only languages that we are sure are not in the page are discarded.
- Increasing the threshold sharply reduces OCR accuracy, because we consider pages with no languages to be empty and skip the second OCR pass.



Results

Image Super-Resolution

- *Google Vision AI* does not benefit from image super-resolution techniques.
- *Tesseract 4* always achieves better OCR accuracy with super-resolution than with *low-resolution images*.
- *Tesseract 4* outperforms even *high-resolution images* with the pre-trained *Waifu2x* models and with our trained SRGAN model.



Results

Text Corpus

- We combined our most successful preprocessing techniques:
  - layout detection using *geometry*,
  - language identification using
    - *0% threshold*
    - *nine* language models
    - *paragraph-based* technique
  - image super-resolution using the *Waifu2x* pre-trained with *high noise removal*
- We achieved 5.42% word error rate compared to 8.74% with no preprocessing.

- To enable NLP research of medieval texts, we publish the text corpus of OCR texts.

Conclusion and Future Work

Conclusion

- OCR of scanned images for contemporary printed texts is considered solved problem.
- OCR of early printed books and reprints of medieval texts remains an open challenge.

- We developed eight preprocessing techniques in three different areas.
- We showed that they can improve the OCR accuracy on medieval texts.
- We also published an open dataset of:
  - 51,351 scanned images and OCR texts
  - 120 annotations for layout analysis and OCR evaluation
  - 122 annotations for language identification

Future work

- We only used language identification for whole pages.
- OCR accuracy may be improved by processing smaller areas of the page separately.

- We produced empty OCR outputs when no languages passed the confidence threshold.
- Just disabling the language models in Tesseract may give better results.

Thank You for Your Attention!

Bibliography I

[1] Vít Novotný. “When Tesseract Does It Alone: Optical Character Recognition of Medieval Texts”. In: *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2020*. Ed. by Aleš Horák et al. Tribun EU, 2020, pp. 3–12. ISBN: 978-80-263-1600-8.

[2] Friedrich M. Wahl et al. “Block segmentation and text extraction in mixed text/image documents”. In: *Computer graphics and image proc.* 20.4 (1982), pp. 375–390.

[3] Lawrence O’Gorman. “The document spectrum for page layout analysis”. In: *IEEE Transactions on pattern analysis and machine intelligence* 15.11 (1993), pp. 1162–1173.

[4] Koichi Kise et al. “Segmentation of page images using the area Voronoi diagram”. In: *Computer Vision and Image Understanding* 70.3 (1998), pp. 370–382.

Bibliography II

[5] Ming Chen et al. “Unified HMM-based layout analysis framework and algorithm”. In: *Science in China, Series F: Information Sciences* 46.6 (2003), pp. 401–408.

[6] S. Chowdhury et al. “Segmentation of text and graphics from document images”. In: *ICDAR*. Vol. 2. IEEE. 2007, pp. 619–623.

[7] Henry S. Baird et al. “Image segmentation by shape-directed covers”. In: *ICPR*. Vol. 1. IEEE. 1990, pp. 820–825.

[8] Theo Pavlidis and Jiangying Zhou. “Page segmentation and classification”. In: *CVGIP: Graphical models and image proc.* 54.6 (1992), pp. 484–496.

[9] Thomas M. Breuel. “Two geometric algorithms for layout analysis”. In: *Int. workshop on document analysis systems*. Springer. 2002, pp. 188–199.

Bibliography III

[10] Kwan Y. Wong et al. “Document analysis system”. In: *IBM journal of research and development* 26.6 (1982), pp. 647–656.

[11] G. Nagy and S. C. Seth. “Hierarchical Representation of Optically Scanned Documents”. In: *ICPR*. Montreal, Canada, 1984, pp. 347–349.

[12] Smith. “Hybrid page layout analysis via tab-stop detection”. In: *ICDAR*. IEEE. 2009, pp. 241–245.

[13] Smith. “An overview of the Tesseract OCR engine”. In: *ICDAR*. IEEE. 2007, pp. 629–633.

[14] Dar-Shyang Lee and Ray Smith. “Improving book OCR by adaptive language and image models”. In: *Int. Workshop on Document Analysis Systems*. IEEE. 2012, pp. 115–119.

Bibliography IV

[15] Chao Dong et al. “Learning a Deep Convolutional Network for Image Super-Resolution”. In: *Computer Vision – ECCV 2014*. Ed. by David Fleet et al. 2014, pp. 184–199. ISBN: 978-3-319-10593-2.

[16] Christian Ledig et al. “Photo-realistic single image super-resolution using a generative adversarial network”. In: *Proc. of the IEEE conf. on computer vision and pattern recognition*. 2017, pp. 4681–4690.

[17] Ankit Lat and C. V. Jawahar. “Enhancing OCR accuracy with super resolution”. In: *ICPR*. IEEE. 2018, pp. 3162–3167.

[18] Ryo Nakao et al. “Selective Super-Resolution for Scene Text Images”. In: *ICDAR*. 2019, pp. 401–406.

Bibliography V

[19] Xiangdong Su et al. “Improving Text Image Resolution using a Deep Generative Adversarial Network for Optical Character Recognition”. In: *ICDAR*. 2019, pp. 1193–1199.

[20] Kha Cong Nguyen, Cuong Tuan Nguyen, et al. “A Character Attention Generative Adversarial Network for Degraded Historical Document Restoration”. In: *ICDAR*. 2019, pp. 420–425.

[21] Zhichao Fu et al. “Cascaded Detail-Preserving Networks for Super-Resolution of Document Images”. In: *ICDAR*. IEEE Computer Society. 2019, pp. 240–245.

[22] Anupama Ray, Manoj Sharma, et al. “An End-to-End Trainable Framework for Joint Optimization of Document Enhancement and Recognition”. In: *ICDAR*. 2019, pp. 59–64.

Bibliography VI

[23] Radovan Panák. “Digitalizace matematických textů”. MA thesis. Faculty of Informatics, Masaryk University, 2006. URL: https://is.muni.cz/th/pspz5/.

[24] Peter Selinger. *Potrace: a polygon-based tracing algorithm*. [cited 2021-11-07]. 2003. URL: http://potrace.sourceforge.net/potrace.pdf.

[25] Zbyněk Sviták et al. *Codex diplomaticus et epistolaris Regni Bohemiae. Tomi VI*. Academia, 2006. ISBN: 9788020015228.

MUNI
FACULTY
OF INFORMATICS