When Tesseract Does It Alone
Optical Character Recognition of Medieval Texts

Vít Novotný

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
witiko@mail.muni.cz
https://mir.fi.muni.cz/

Abstract. Optical character recognition of scanned images for contemporary printed texts is widely considered a solved problem. However, the optical character recognition of early printed books and reprints of Medieval texts remains an open challenge. In our work, we present a dataset of 19th and 20th century letterpress reprints of documents from the Hussite era (1419–1436) and perform a quantitative and qualitative evaluation of speed and accuracy on six existing OCR algorithms. We conclude that the Tesseract family of OCR algoritms is the fastest and the most accurate on our dataset, and we suggest improvements to our dataset.

Keywords: Optical character recognition, OCR, Historical texts

1 Introduction

The aim of the AHISTO project is to make documents from the Hussite era (1419–1436) available to the general public through a web-hosted searchable database. Although scanned images of letterpress reprints from the 19th and 20th century are available at the Czech Medieval Sources online (CMS online) web site,[1] accurate optical character recognition (OCR) algorithms are required to extract searchable text from the scanned images.

In this paper, we compare the speed and accuracy of six OCR algorithms on the CMS online dataset both quantitatively and quantitatively. In Section 2, we describe the OCR algotithms. In Section 3, we describe our dataset, how it was pre-processed and used in our quantitative and qualitative evaluation. In Section 4, we discuss the results of our evaluation. In Section 5, we offer concluding remarks and ideas for future work in the OCR of Medieval texts.

2 Related Work

Optical character recognition makes it possible to convert scanned images to digital text. For the AHISTO project, an OCR algorithm should:

[1] https://sources.cms.flu.cas.cz/

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2020, pp. 3–12, 2020. © Tribun EU 2020

1. support different European languages, mainly Czech, German, and Latin,
2. detect language of the text to enhance the searchability of our database, and
3. use a standard output format to prevent a vendor lock-in.

In this section, we will describe and discuss the OCR algorithms we considered.

2.1 Google Cloud Vision AI

Google Cloud Vision AI is a paid OCR service made available by Google in 2015.[2]

Google Vision AI supports Czech, German, and Latin, among other languages,[3] and detects language at the level of individual letters. Regrettably, only a non-standard JSON output format is supported, leading to a vendor lock-in.

Google Cloud Vision AI provides two features: `DOCUMENT_TEXT_DETECTION` and `TEXT_DETECTION`. `DOCUMENT_TEXT_DETECTION` performs the OCR of printed documents, whereas `TEXT_DETECTION` solves the more general task of detecting text in arbitrary images, such as camera images of traffic signs. In our experiments, we used the `DOCUMENT_TEXT_DETECTION` feature.

2.2 Tesseract

Tesseract [6] has been developed by Hewlett-Packard in the 1980s and released under a free open-source license in 2005. Since 2006, the development of the project has been funded by Google, and it is presumed that parts of Tesseract are also used in Google Cloud Vision AI.

Initially, Tesseract only supported English. Since version 2, Tesseract has also supported Western languages, including German. Since version 3, Tesseract has also supported Czech, [7], detected language at the level of words, and added support for the standard HOCR XML output format. [1] Since version 3.04, Tesseract has also supported Latin. Since version 4, Tesseract has supported a new LSTM-based OCR engine,[4] which is regrettably not GPU-accelerated.

In our experiments, we used Tesseract 4.00 in three configurations:

1. `--oem 0`, which uses the non-LSTM OCR engine (f.k.a. Tesseract 3),
2. `--oem 1`, which uses the LSTM OCR engine (f.k.a. Tesseract 4), and
3. `--oem 2`, which ensembles the two OCR engines (f.k.a. Tesseract 3 + 4).

For all configurations, we used the `--psm 3` page segmentation mode, the Czech, German, and Latin (`-l ces+deu+lat`) language models, and the medium-size pre-trained models,[5] which, compared to the best models,[6] support the non-LSTM OCR engine of Tesseract 3.

[2] https://cloud.google.com/vision/docs/release-notes

[3] https://cloud.google.com/vision/docs/languages#supported-langs

[4] https://tesseract-ocr.github.io/tessdoc/NeuralNetsInTesseract4.00

[5] https://github.com/tesseract-ocr/tessdata.git

[6] https://github.com/tesseract-ocr/tessdata_best.git

2.3 OCR-D

OCR-D [2] is a free open-source project that has been funded by the German Research Foundation and developed by the Berlin-Brandeburg Academy of Sciences, the Herzog August Bibliothek, the Berlin State Library, and the Karlsruhe Institute of Technology. The main goal of the project is to digitize the German cultural heritage 16th–19th century in connection with the VD16, VD17, and VD18 cataloging and archival projects. In February 2020, the project has entered phase 3 and OCR-D is now being deployed to intent organizations.

Unlike Google Cloud Vision AI and Tesseract, which provide fully-automated OCR workflows, OCR-D's workflows are fully-configurable[7] and include image enhancement, binarization, cropping, denoising, deskewing, dewarping, segmentation, clipping, line OCR, text alignment, and post-correction. Another distinguishing feature of OCR-D is the OCR4all web frontend,[8] which enables semi-automatic OCR with human input.

As a part of the line OCR workflow step, OCR-D supports the LSTM-based and GPU-accelerated Calamari engine,[9] which has achieved the state-of-the-art performance on historical texts typeset in Fraktur. [9,4] With Calamari, OCR-D regrettably does not detect the language of the text, but it supports the standard HOCR XML output format [1] like Tesseract.

In our experiments, we used the recommended workflow for OCR-D,[10] which includes Calamari as the line OCR engine. Since Calamari is GPU-accelerated, we tested OCR-D both with and without a GPU to see the difference in speed.

3 Methods

In this section, we will describe the CMS online dataset, how it was pre-processed and then used to evaluate the OCR algorithms discussed in the previous section.

3.1 Data Preprocessing

In the CMS online dataset, we received 302,909 low-resolution and 168,113 high-resolution scanned images of letterpress reprints from the 19th and 20th centuries. For all low-resolution images, we received Google Cloud Vision AI OCR outputs, although it is unknown if these were produced by the `TEXT_DETECTION` feature, or the more appropriate `DOCUMENT_TEXT_DETECTION` feature, and if they were produced using the low-resolution or the high-resolution images.

Although the high-resolution images have an estimated average density of 414 DPI and are suitable for OCR, the low-resolution images have an estimated average density of only 145 DPI. Therefore, we had to design an algorithm that would link the low-resolution images to the matching high-resolution images in

[7] https://ocr-d.de/en/workflows

[8] https://www.uni-wuerzburg.de/en/zpd/ocr4all/

[9] https://github.com/Calamari-OCR/calamari

[10] https://ocr-d.de/en/workflows#best-results-for-selected-pages

order to produce a test dataset containing high-resolution images together with corresponding Google Cloud Vision AI OCR outputs as ground truth.

In designing the algorithm, we assumed that there was no other difference between matching low-resolution and high-resolution images other than downscaling. Additionally, we manually inspected the low-resolution and high-resolution images to find a subset of 187,267 (62%) low-resolution images guaranteed to cover all the books from which the high-resolution images originated.

A pseudocode of our linking algorithm is given in Algorithm 1. Using the algorithm, we linked 65,348 (39%) high-resolution scanned images with Google Vision AI ground truth OCR outputs, which served as our test dataset.

Algorithm 1: Linking low-resolution and high-resolution images.

```
Result: Linked low-resolution and high-resolution images.
Preprocess all images by rescaling them to 512 × 512 px and binarizing;
Index the preprocessed high-resolution images in a vector database;
foreach preprocessed low-resolution image do
    Retrieve the ground truth Google Vision AI output;
    Retrieve the 100 preprocessed high-resolution images nearest to the
      preprocessed low-resolution image by the Hamming distance;
    foreach neighboring preprocessed high-resolution image do
        Process the high-resolution image by Tesseract 4;
    end
    Rerank the 100 nearest high-resolution images by TF-IDF cosine similarity
      between the Google Vision AI ground truth and the Tesseract 4 output;
    if the nearest high-resolution image is the same after reranking then
        Prelink the low-resolution image with its nearest high-resolution image;
    end
end
foreach book do
    if all low-resolution images in the book are prelinked then
        Link all low-resolution images in the book with their prelinked
          high-resolution images;
    end
end
```

3.2 Quantitative Evaluation

Speed To evaluate the speed of the OCR algorithms, we measured and report the wall clock time in days to process the test dataset on a single CPU/GPU. For Tesseract, we used the `apollo`, `asteria04`, `mir`, `hypnos1`, `turnus01`, and `nymfe{23..74}` nodes at the Faculty of Informatics, Masaryk University, totalling 492 CPU cores. For OCR-D, we used the the `epimetheus{1..4}` nodes for evaluating the CPU speed, and `turnus03` for evaluating the GPU speed. For OCR-D, we also measured and report the speed of each workflow step separately.

Accuracy To evaluate the accuracy of the OCR algorithms, we measured and report the Character Error Rate (CER) and the Word Error Rate (WER): [8]

$$\mathrm{ErrorRate}(A, B) = \frac{\mathrm{EditDistance}(A, B)}{\mathrm{Maximum}(|A|, |B|)} \times 100\%.$$

For WER, we lower-cased and deaccented the texts in addition to tokenization in order to better model our full-text search use case. Although CER and WER are correlated, we were mainly interested in WER, which is harder and better corresponds to our full-text search use case.

To discover the origin of the ground truth, we evaluated the accuracy of the Google Cloud Vision AI using both the low-resolution images (f.k.a. G-low) and the high-resolution images (f.k.a. G-high), not just the high-resolution images.

3.3 Qualitative Evaluation

To better understand the strengths and weaknesses of the individual OCR algorithms, we compared their outputs on three different scanned images:

1. A random image from the test dataset.
2. The image with the best WER using the least accurate OCR algorithm.
3. The image with the worst WER using the most accurate OCR algorithm.

To inspect the quality of the ground truth, we manually classified the differences between the ground truth and the OCR output on the three images as either improvements or errors, and we report the ratio of improvement to error.

4 Results

In this section, we will describe the results of the evaluation described in the previous section. We report both quantitative and qualitative evaluation results.

4.1 Quantitative Evaluation

Speed Table 1 shows that Tesseract 3 is the fastest OCR algorithm. The second fastest Tesseract 4 is more than twice as slow as Tesseract 3 because of the higher computational complexity of the non-GPU-accelerated LSTM-based OCR engine. OCR-D achieves speed that is comparable with Tesseract 4 despite the GPU acceleration of Calamari, which is likely because of the fully-configurable workflows and the high disk I/O caused by the locking and the common updates

Table 1. The wall clock time to process the test dataset with different OCR algorithms ordered from the fastest to the slowest. Since Google Cloud Vision AI is not self-hosted, it was not included in the speed evaluation.

	Tesseract 3	Tesseract 4	OCR-D (GPU)	Tesseract 3 + 4	OCR-D (CPU)
Time (days)	*61.12*	127.69	140.39	172.11	174.07

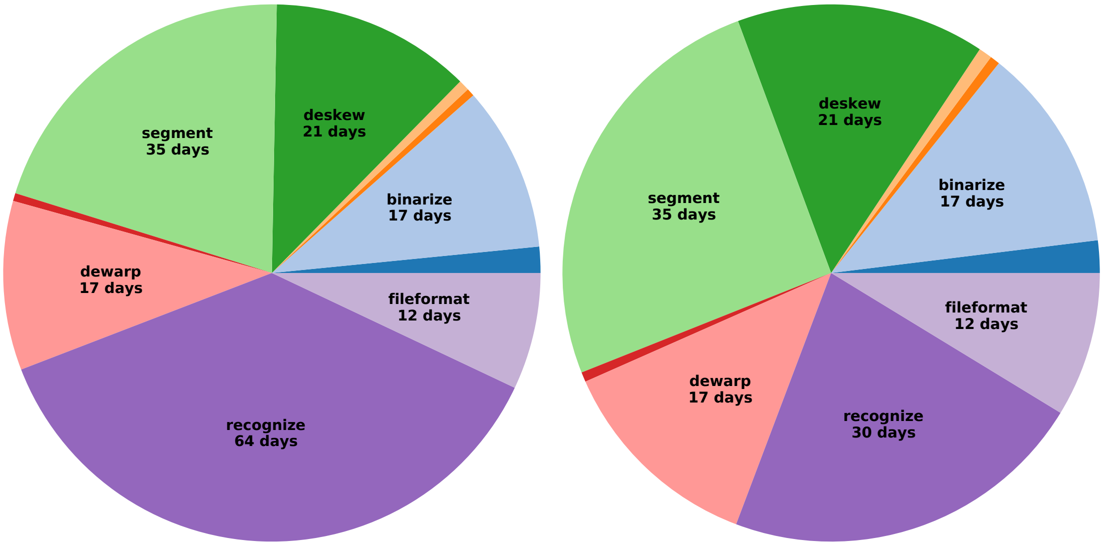

Fig. 1. The wall clock time to process the test dataset on the level of the individual workflow steps of OCR-D on a single CPU (left) and on a single GPU (right).

to a large METS XML document containing the history of all workflow steps executed for all scanned images in a book. Tesseract 3 + 4 is the second slowest with the time roughly equivalent to the sum of the times of Tesseract 3 and Tesseract 4. OCR-D without GPU acceleration is the slowest OCR algorithm.

Figure 1 shows that without GPU acceleration, more than one third of OCR-D's time is taken up by the Calamari line OCR workflow step. GPU acceleration causes more than 2× speed increase of this workflow step, whereas the other steps do not benefit from GPU acceleration.

Accuracy Table 2 shows that Google Cloud Vision AI achieved perfect accuracy with neither low-resolution nor high-resolution images, which leads us to the conclusion that the ground truth had been produced using the less appropriate `TEXT_DETECTION` feature. The higher accuracy of Google Cloud Vision AI with low-resolution images leads us to believe that the ground truth had been produced using the low-resolution images rather than the high-resolution images. Both findings raise doubts about the quality of the ground truth.

Table 2 also shows that Tesseract 4 is the most accurate OCR algorithm with the second most accurate being Tesseract 3. Ensembling Tesseract 3 and Tesser-

Table 2. The accuracy of different OCR algorithms on the test dataset ordered from the most accurate to the least accurate. G-low and G-high correspond to the Google Cloud Vision AI with low-resolution images and high-resolution images, respectively.

	G-low	G-high	*Tesseract 4*	Tesseract 3	Tesseract 3 + 4	OCR-D
WER (%)	3.85	5.24	*13.77*	16.04	18.70	30.94
CER (%)	2.44	3.38	*9.81*	10.60	12.00	19.88

act 4 leads to the second worst accuracy, which suggests a poor confidence heuristic of Tesseract in selecting words from the suggestions of both OCR algorithms. OCR-D is the least accurate OCR algorithm, likely because we used the recommended Calamari models trained on German Fraktur,[11] whereas most of our images are scans of letterpress prints in Antiqua typefaces with Czech accents.

4.2 Qualitative Evaluation

Random image We randomly selected page 5 from *Archiv český 01*,[12] with the following ground truth text:

> I. PSANJ ČESKÁ CJSARE SIGMUNDA
> od roku 1414 do 1437.
>
> Panu Čeńkowi z Wartenberka a z Weselj, neyw. purkrabj Pražskému: aby králowně Sofii na statejch gegjch wěnnjch nátisku činiti nedopauštěl.
>
> Z Teplice u Ferrary, 1414, 24 Mart.
>
> Sigmund z božie milosti Rimský a Uherský oc. král.
>
> Urozený wěrný mily! Slyšíme, že někteří páni w Čechách najoswiecenější kněžnu pant Sofii, králewnu Českú, sestru naši milú, mienie a chystaji sie, jie na jejím wèně mimo práwo a mimo panský nález tisknúti; jenžto nerádi slyšíme, anižbychom toho rádi dopustili, by sie jie to od koho mělo státi. Protož od tebe žádáme i prosime, byloliby žeby jmenowanú králewnu, sestru naši milú, mimo prawo kto tisknúti, a nebo na jejiem wèně překážeti chtěl, aby podle ní stál, a jie wěrně pro nás pomohl, aby od swého nebyla tištěna. Na tom nám zwláštní službu učiniš a ukážeš. Dán w Teplici u Ferrarii, weder matky božie Annuntia-tionis, léta králowstwie našich Uherského uc. w XXVII., a Římskéhow čtwrtém létě,
>
> Ad mandatum D. Regis: Michael de Priest.
>
> Urozenému Čeńkowi z Wesele
> wěrnému nám zwláště milému.

Tesseract 3 achieved 6.36% WER and 3.96% CER. Over 42% of changes were improvements, raising doubts about the quality of the ground truth:

>]. PSANJ ČESKÁ CJSAŘE SIGMUNDA
> od roku 1414 do 1437.
>
> Panu Čeňkowi z Wartenberka a z Weselj, neyw. purkrabj Pražskému: aby králowně Sofii na. statcjch gegjch wénnjch nátisku činiti nedopauštěl.
>
> Z Teplice uFermry, 1414, 24 Mart.
>
> Sigmund z božie milosti Římský a Uherský oc. král.
>
> Urozený wěrný milý! Slyšíme, že někteří páni w Čechách najoswiecenějši kněžnu paní Sofii, králewnu Česků, sestru nasi milů, mienie a chystají sie, jie na jejím wené mimo práwo a mimo panský nález tisknúti; jenžto nerádi slyšíme, anižbychom toho rádi dopustili, by sie jie to od koho mělo státi. Protož od tebe žádáme i prosime, byloliby žeby jmenowanů králewnu, sestru naši milů, mimo práwo kto tisknúti, a nebo na jejiem wené překážeti chtěl, aby podlé ní stál, a jie wěrně pro nás pomohl, aby od swého nebyla tištěna. Na tom nám zwlástní službu učiníš & ukážeš. Dán W Teplici u Ferrarii, wečer matky božie Annuntia-tionis, létá králowstwie našich Uherského oc. w XXVII., & Římského w čtwrtém létě.
>
> Ad mandatum D. Regis: Michael de Priest.
>
> Urozenému Čeňkowi z Weselé
> Wěrnému nám zwláště milému.

Tesseract 4 achieved 8.57% WER, 4.95% CER. 28% changes were improvements:

> I. PSANJ ČESKÁ CJSAŘE SIGMUNDA
> od roku 1414 do 1437.
>
> Panu Čeňkowi z Wartenberka a z Weselj, neyw. purkrabj Pražskému: aby králowně Sofii na statcjch gegjich wénnjch nátisku činiti nedopauštěl.
>
> Z Teplice u Ferrary, 1414, 24 Mart.
>
> Sigmund z bozie milosti Římský a Uherský oc. král.
>
> Urozený wěrný milý! Slyšiíme, že někteří páni w Čechách najoswiecenější kněžnu paní Sofii, králewnu Česků, sestru naši milů, mienie a chystají sie, jie na jejím wěené mimo práwo a mimo panský nález tisknůti; jenžto nerádi slyšíme, anižbychom toho rádi dopustili; by sie jie to od koho mělo státi. ProtoZ od tebe žádáme i prosime, byloliby Zeby jmenowanü králewnu, sestru naài milü, mimo práwo kto tisknüti, a nebo na jejiem wéné prekázeti chtél, aby podlé ni stàl, a jie wěrně pro nás pomohl, aby od swého nebyla tištěna. Na tom nám zwláštní sluZbu ucini$ a ukážeš. Dán w Teplici u Ferrarii, wečer matky boZie Annuntia-tionis, léta králowstwie našich Uherského sc. w XXVIL, a Římského w čtwrtém létě,
>
> Ad mandatum D. Regis: Michael de Priest.
>
> Urozenému Ceünkowi z Weselé
> wérnému náàm zwlásté milému.

[11] https://ocr-d-repo.scc.kit.edu/models/calamari/GT4HistOCR/model.tar.xz

[12] https://sources.cms.flu.cas.cz/src/index.php?cat=10&bookid=792&page=5

Tess. 3 + 4 achieved 12.43% WER, 8.35% CER. 12% changes were improvements:

> I. PSANJ ČESKÁ CJSAŘE SIGMUNDA
> od roku 1414 do 1437.
>
> Panu Čeňkowi z Wartenberka a z Weselj, neyw. purkrabj Pražskému: aby královně Sofii na statcjch gegjch wěnnjch nátisku činiti nedopauštěl.
>
> Z Teplice u Ferrary, 1414, 24 Mart.
>
> Sigmund z božie milosti Římský a Uherský oc. král.
>
> Urozeny wörny mily! Sly&tme, ze nékterí páni w Cechách najoswiecenéjsi knéé£nu pani Sofii, kráülewnu Ceskü, sestru nasi milü, mienie a chystaji sie, jie na jejim wéné mimo práwo a mimo pansky nález tisknüti; jenZto nerádi slysime, aniZbychom toho rádi dopustili, by sie jie to od koho mélo státi. ProtoZ od tebe Zádáme i prosime, byloliby Zeby jmenowanü královnu, sestru naài milü, mimo práwo kto tisknüti, a nebo na jejiem wéné prekázeti chtél, aby podlé ni stàl, a jie wérné pro nás pomohl, aby od swého nebyla tisténa. Na tom nām zwlāstni sluZbu ucini$ a ukààZes. Dáàn w Teplici u Ferrari, wecer matky bozie Annuntia-tionis, léta králowstwie nasich Uherského oc. w XXVIL, a Rimskeho w ótwrtém lété,
>
> Ad mandatum D. Regis: Michael de Priest.
>
> Urozenému Ceünkowi z Weselé
> wérnému náàm zwlásté milému.

OCR-D achieved 27.17% WER, 13.95% CER. Only 5% of changes were improvements, which indicates that the ground truth can still differentiate good and bad OCR outputs. Due to the training on German Fraktur without Czech accents, most errors originate from accented letters:

> J. PaAN SINA
> od roku 1414 do 1437.
>
> Panu Cekowi 2 artenberka a 2 Veselj, neyw. purkrabj Praiskému: aby krâlown Sofii na statejch gegjeh wènnjeh nâtisku öiniti nedopaustẽl.
>
> Teplice u Ferrar, 1414, 24 Mart.
>
> Sigmund z boie milosti Rimsky a Uhersk oc. krâl.
>
> ⫽rozeny wörny milyl Slysime, e nekteri pâni w Cechâch najoswiecenèjsi knênu pani Sofii, kralewnu Cesk, sestru nasi milu, mienie a chystaji sie, jie na jejim wenè mimo prâwo a mimo pansky nalez tisliti; jenàto nerâdi slysime, anizbychom toho radi dopustili, by sie jie to od koho mêlo stâti. Proto od tebe adame i prosime, byloliby eby jmenowanu kralewnu, sestru nasi milu, mimo prâwo kto tiskniti, a nebo na jejiem wènè prekâieti chtél, aby podlé ni stal, a jie wvrnẽ pro nas pomohl, aby od swého nebyla tistêna. Na tom nâm zwlastni sluibu ucinis a ukâes. Dân w Teplici u Ferrarii, wecer matky boie Annuntia-tionis, léta krâlowstwie nasich Uherského oe. w XXVJ., a Rimského w êtwrtém létè.
>
> Ad mandatum D. Regis: Michael de Priest.
>
> Urozenému Ceükowi 2 Veselé
> wêrnému nam zlastè milému.

Best image OCR-D achieved the best accuracy (0% WER, 0.29% CER) on page 59 of *CIM I*.[13] None of the changes were improvements, confirming our thesis that the ground truth can reliably distinguish good and bad OCR outputs:

> vllam exigere, petere aut recipere volumus pecuniam, aut exigi, peti
> …
> vngelta in dicta ciuitate videlicet pannorum, mercium institarum et braxaturas ceruisie cum prouentibus ipsorum iuxta placita, per nos et ipsos ciues nostros hincinde habita, infra spacium dictorum annorum cum adicione quinti anni pro se recipere debeant et habere, quousque omnia et singula debita per ipsos ciues nostros quocumque modo contracta in integrum fuerint persoluta, impedimento nostro et cuius- libet non obstante. Debet quoque ipsum vngeltum sic recipi atque dari, videlicet quod vendens pannos siue merces institarum ciuis vel hospes de qualibet sexagena grossorum Sex paruos denarios vsuales et emens merces easdem totidem paruos in prima vendicione et em- pcione tantum et non vlterius soluere est astrictus, et quilibet braxans ceruisiam in ipsa ciuitate de vna braxatura ceruisie vnum grossum pro vngelto ipsis ciuibus dare debet. Si quis vero sub vna sexagena grossorum vendiderit vel emerit in mercibus, vt predicitur, quidquam, de huiusmodi vendicione et empcione pro vngelto nichil dabit. Addi- cimus eciam, quod omnes mercatores Pragam cum pannis quibus- cunque uel mercibus institarum ibidem non emptis transire volentes,
> …
>
> Liber vetustissimus statutorum c. 993 str. 61 V archivu m. Prahy.

Worst image Tesseract 4 achieved the worst accuracy (100% WER, 81.56% CER) on page 1297 of *CIM II*.[14] All changes were improvements, since the ground truth did not detect two-column page layout, unlike Tesseract 4. This confirms our thesis that although the ground truth can distinguish good and bad OCR outputs, it cannot distinguish OCR outputs that are better than the ground truth.

[13] https://sources.cms.flu.cas.cz/src/index.php?cat=12&bookid=117&page=230

[14] https://sources.cms.flu.cas.cz/src/index.php?cat=12&bookid=118&page=1327

5 Conclusion and Future Work

In our work, we have compared the speed and the accuracy of six OCR algorithms on the CMS online dataset from the AHISTO project.

Based on our results, we conclude that the ground truth OCR outputs in our dataset are low-quality and should be replaced or supplemented either by human judgements, or by synthetic data produced from searchable PDF documents.

As far as the ground truth can be trusted, Tesseract 4 is the second fastest and the most accurate OCR algorithm, which also detects language at the level of words. Pre-detecting the language of paragraphs using *n*-gram frequency analysis and then using Tesseract 4 only with the language models corresponding to the detected languages can further improve its accuracy. [3, Section 4.4]

OCR-D with the Calamari line OCR is comparable to Tesseract 4 in speed and is likely to produce more accurate results than Tesseract 4. However, the performance of Calamari was undermined by its poor pre-trained models. Additionally, Calamari does not detect language. However, OCR-D can align OCR outputs from Calamari and Tesseract 4, which can bring the best of both worlds.

Applying super-resolution algorithms [5] to the low-resolution scanned images can make them suitable for OCR and the AHISTO project.

Acknowledgments. First author's work was funded by the South Moravian Centre for International Mobility as a part of the Brno Ph.D. Talent project and also by TAČR Éta, project number TL03000365. I also sincerely thank my colleagues from the NLP Centre and the MIR research group at the Masaryk University for their insight.

References

1. Breuel, T.M.: The hOCR microformat for OCR workflow and results. In: ICDAR 2007. vol. 2, pp. 1063–1067. IEEE (2007)
2. Neudecker, C., Baierer, K., Federbusch, M., Boenig, M., Würzner, K.M., Hartmann, V., Herrmann, E.: OCR-D: An end-to-end open source OCR framework for historical printed documents. In: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage. pp. 53–58 (2019)
3. Panák, R.: Digitalizace matematických textů. Master's thesis, Faculty of Informatics, Masaryk University (2006), https://is.muni.cz/th/pspz5/
4. Reul, C., Springmann, U., Wick, C., Puppe, F.: State of the art optical character recognition of 19th century fraktur scripts using open source engines (2018), https://arxiv.org/pdf/1810.03436.pdf
5. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1874–1883 (2016)
6. Smith, R.: An overview of the Tesseract OCR engine. In: ICDAR 2007. vol. 2, pp. 629–633. IEEE (2007)

7. Smith, R., Antonova, D., Lee, D.S.: Adapting the Tesseract open source OCR engine for multilingual OCR. In: Proceedings of the International Workshop on Multilingual OCR. pp. 1–8 (2009)
8. Soukoreff, R.W., MacKenzie, I.S.: Measuring errors in text entry tasks: An application of the levenshtein string distance statistic. In: CHI'01 extended abstracts on Human factors in computing systems. pp. 319–320 (2001)
9. Wick, C., Reul, C., Puppe, F.: Calamari-a high-performance tensorflow-based deep learning package for optical character recognition (2018), https://arxiv.org/pdf/1807.02004.pdf