International Journal on Digital Libraries (2025) 26:3
https://doi.org/10.1007/s00799-025-00413-z

Enhancing OCR in historical documents with complex layouts through machine learning

David Fleischhacker[1] · **Roman Kern**[1,2] · **Wolfgang Göderle**[3,4]

Received: 5 October 2023 / Revised: 20 August 2024 / Accepted: 19 January 2025

Abstract

This paper explores the challenge of processing and extracting information from large quantities of printed serial sources from the 19th century, which have been largely untapped due to the inadequacies of existing extraction techniques. We focus on the Habsburg Central Europe's *Hof- und Staatsschematismus*, a comprehensive record published between 1702 and 1918 that documents the Habsburg civil service's hierarchy and the evolution of its central administration over two centuries. Our approach sees the significant investment into machine learning-driven layout detection prior to the OCR-process. We generated synthetic data mimicking the *Hof- und Staatsschematismus* style for initial training of a Faster R-CNN model, followed by fine-tuning the model with a smaller dataset of manually annotated historical documents. Subsequently, we optimised Tesseract-OCR for our document style to enhance the combined structure extraction and OCR process. Our evaluation demonstrates significant improvements in OCR performance metrics (WER and CER), with the combined structure detection and fine-tuned OCR process showing a decrease in error rates of 15.68 percentage points for CER and 19.95 percentage points for WER. These findings underscore the potential of ML techniques in facilitating the extraction and analysis of historical documents.

Keywords PDF extraction · Layout detection · OCR fine-tuning · Synthetic training data · Document analysis and recognition

1 Introduction

The long 19th century provides historians and fellow humanists with a wealth of retrodigitized printed sources. A significant share of these is made up of serial publications with highly structured content and complex layouts (Visually Rich Documents, VRDs) [1]. A high-profile example of this is the Habsburg Monarchy's *Hof- und Staatsschematismus*, which was published from 1702 to 1918 [2]. It provides us with a set of serial data on the administrative and representative elites of Habsburg Central Europe in very high quality [3]. A thorough analysis of this data set would contribute decisively to a significantly better understanding of relevant social processes, power dynamics, social networks and careers in modern Central Europe. The *Schematismus* could be used to trace the genesis of state and administrative institutions, their functioning and development, and the professional biographies of tens of thousands of officials and decision-makers over more than two centuries, across political ruptures and social transformations.

However, the complex structure of such publications has made a more comprehensive and quantitative evaluation of this source impossible. Even though OCR-quality has improved dramatically since the early days of the retrodigitisation of historical publications, structure analysis and layout detection remain a challenge. The *Schematismus* is known for its multi-column layouts, deeply branched hier-

Roman Kern and Wolfgang Göderle have contributed equally to this work.

✉ David Fleischhacker
david.fleischhacker@student.tugraz.at

Roman Kern
rkern@tugraz.at

Wolfgang Göderle
wolfgang.goederle@uibk.ac.at

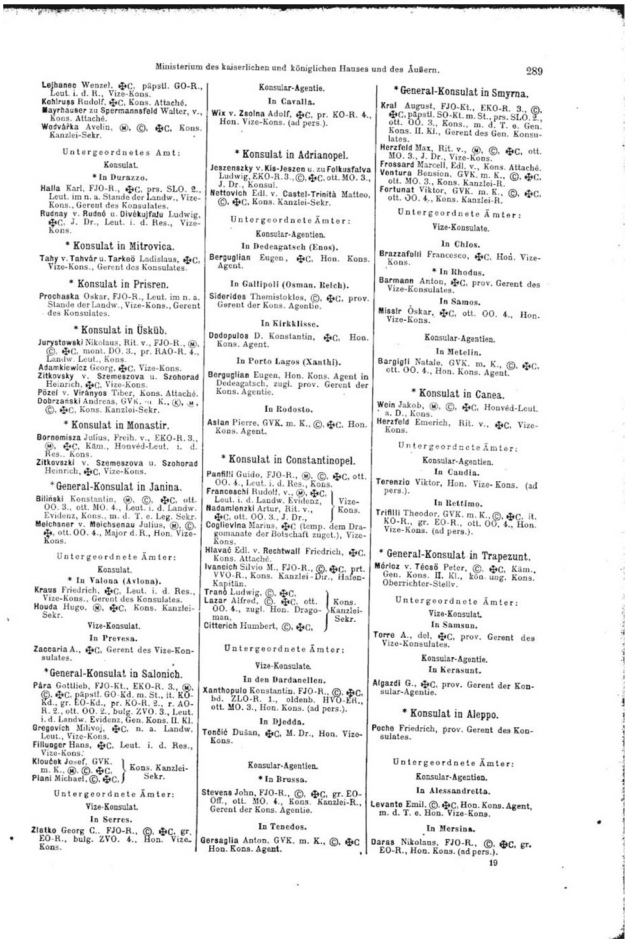
[1] Institute of Interactive Systems and Data Science, Graz University of Technology, Sandgasse 36, 8010 Graz, Austria

[2] Know Center, Sandgasse 36, 8010 Graz, Austria

[3] Department of Austrian History, Institute of History, University of Innsbruck, Innrain 52, 6020 Innsbruck, Austria

[4] Department Structural Changes of the Technosphere, Max Planck Institute of Geoanthropology, Kahlaische Strasse 10, 07745 Jena, Germany

Published online: 17 February 2025



Ministerium des kaiserlichen und königlichen Hauses und des Äußern. 289

Lojbaner Wenzel, ✠C. päpstl. GO-R., Leut. i. d. R., Vize-Kons.
Kohlruss Rudolf, ✠C. Kons. Attaché.
Mayrhauser zu Spermannsfeld Walter, v., Kons. Attaché.
Wodvárka Avelin, Ⓜ, Ⓒ, ✠C. Kons. Kanzlei-Sekr.

Untergeordnetes Amt:

Konsulat.

* In Durazzo.

Halla Karl, FJO-R., ✠C. prs. SLO. 2., Leut. im n. a. Stande der Landw., Vize-Kons., Gerent des Konsulates.
Rudesy v. Ruđné u. Divékújfalu Ludwig, ✠C. J. Dr., Leut. i. d. Res., Vize-Kons.

* Konsulat in Mitrovica.

Tahy v. Tahvár u. Tarkső Ladislaus, ✠C. Vize-Kons., Gerent des Konsulates.

* Konsulat in Prisren.

Prochaska Oskar, FJO-R., Leut. im n. a. Stande der Landw., Vize-Kons., Gerent des Konsulates.

* Konsulat in Üsküb.

Jurystowski Nikolaus, Rit. v., FJO-R. Ⓜ, Ⓒ, ✠C, mont. DO. 3., pr. RAO-R. 4., Landw. Leut., Kons.
Adamkiewicz Georg, ✠C. Vize-Kons.
Zitkovsky v. Szemeszova u. Szohorad Heinrich, ✠C. Vize-Kons.
Pözel v. Virányos Tibor, Kons. Attaché.
Dobrzanski Andreas, GVK., m. K., Ⓚ, Ⓜ, Ⓒ, ✠C. Kons. Kanzlei-Sekr.

* Konsulat in Monastir.

Bornomisza Julius, Freih. v., EKO-R. 3., Ⓜ, ✠C. Käm., Honvéd-Leut. i. d. Res., Kons.
Zitkovszki v. Szemeszova u. Szohorad Heinrich, ✠C. Vize-Kons.

* General-Konsulat in Janina.

Bilinski Konstantin, Ⓜ, Ⓒ, ✠C. ott. OO. 3., ott. MO. 4., Leut. i. d. Landw. Evidenz, Kons., m. d. T. e. Leg. Sekr.
Maichsner v. Maichsenau Julius, Ⓜ, Ⓒ, ✠C. ott. OO. 4., Major d. R., Hon. Vize-Kons.

Untergeordnete Ämter:

Konsulat.

* In Valona (Avlona).

Kraus Friedrich, ✠C. Leut. i. d. Res., Vize-Kons., Gerent des Konsulates.
Houda Hugo, Ⓜ, ✠C. Kons. Kanzlei-Sekr.

Vize-Konsulat.

In Prevesa.

Zaccaria A., ✠C. Gerent des Vize-Konsulates.

* General-Konsulat in Salonich.

Pára Gottlieb, FJO-Kt., EKO-R. 3., Ⓜ, Ⓒ, ✠C. päpstl. GO-Kd. m. St., it. KO-Kd., gr. EO-Kd., pr. KO-R. 2., r. AO-R. 2., ott. OO. 2., bulg. ZVO. 3., Leut. i. d. Landw. Evidenz, Gen. Kons. H. Kl.
Gregovich Milvoj, ✠C. m. a. Landw. Leut., Vize-Kons.
Fillunger Hans, ✠C. Leut. i. d. Res., Vize-Kons.
Klouček Josef, GVK. m. K., Ⓜ, Ⓒ, ✠C. } Kons. Kanzlei-Sekr.
Piani Michael, Ⓒ, ✠C. }

Untergeordnete Ämter:

Vize-Konsulat.

In Serres.

Zlatko Georg C., FJO-R., Ⓒ, ✠C, gr. EO-R., bulg. ZVO. 4., Hon. Vize-Kons.

Konsular-Agentie.

In Cavalla.

Wix v. Zsolna Adolf, ✠C, pr. KO-R. 4., Hon. Vize-Kons. (ad pers.).

* Konsulat in Adrianopel.

Jeszenszky v. Kis-Jeszen u. zu Folkusfalva Ludwig, EKO-R. 3., Ⓒ, ✠C. ott. MO. 3., J. Dr., Konsul.
Nottovich Edl. v. Castel-Trinità Matteo, Ⓒ, ✠C. Kons. Kanzlei-Sekr.

Untergeordnete Ämter:

Konsular-Agentien.

In Dedeagatsch (Enos).

Bergulian Eugen, ✠C. Hon. Kons. Agent.

In Gallipoli (Osman. Reich).

Sideridos Themistokles, Ⓒ, ✠C. prov. Gerent der Kons. Agentie.

In Kirkklisse.

Dodopulos D. Konstantin, ✠C. Hon. Kons. Agent.

In Porto Lagos (Xanthi).

Bergulian Eugen, Hon. Kons. Agent in Dedeagatsch, zugl. prov. Gerent der Kons. Agentie.

In Rodosto.

Aslan Pierre, GVK. m. K., Ⓒ, ✠C. Hon. Kons. Agent.

* Konsulat in Constantinopel.

Pantlli Guido, FJO-R., Ⓜ, Ⓒ, ✠C. ott. OO. 4., Leut. i. d. Res., Kons.
Francaschi Rudolf, v., Ⓜ, ✠C. Leut. i. d. Landw. Evidenz, } Vize-Kons.
Nadamlenzki Artur, Rit. v., ✠C. ott. OO. 3., J. Dr., }
Cogllevina Marius, ✠C (temp. dem Dragomanate der Botschaft zugt.), Vize-Kons.
Hlavač Edl. v. Rechtwall Friedrich, ✠C. Kons. Attaché.
Ivanovich Silvio M., FJO-R., Ⓒ, ✠C. prt. VVO-R., Kons. Kanzlei-Dir., Hofrath-Kapitän.
Tranč Ludwig, Ⓒ, ✠C. } Kons. Kanzlei-Sekr.
Lazar Alfred, Ⓒ, ✠C. ott. OO. 4., zugl. Hon. Dragoman. }
Citterich Humbert, Ⓒ, ✠C. }

Untergeordnete Ämter:

Vize-Konsulate.

In den Dardanellen.

Xanthopulo Konstantin, FJO-R., Ⓒ, ✠C. bd. ZLO-R. 1., oldenb. HVO-Ehr., ott. MO. 3., Hon. Kons. (ad pers.).

In Djedda.

Tenčić Dušan, ✠C. M. Dr., Hon. Vize-Kons.

Konsular-Agentien.

* In Brussa.

Stevens John, FJO-R., Ⓒ, ✠C. gr. EO-Off., ott. MO. 4., Kons. Kanzlei-R., Gerent der Kons. Agentie.

In Tenedos.

Gersaglia Anton, GVK. m. K., Ⓒ, ✠C Hon. Kons. Agent.

* General-Konsulat in Smyrna.

Krai August, FJO-Kt., EKO-R. 3., Ⓒ, ✠C. päpstl. SO-Kt. m. St., prs. SLO. 2., ott. OO. 3., Kons., m. d. T. e. Gen. Kons. II. Kl., Gerent des Gen. Konsulates.
Herzfeld Max, Rit. v., Ⓜ, Ⓒ, ✠C. ott. MO. 3., J. Dr., Vize-Kons.
Frossard Marcell, Edl. v., Kons. Attaché.
Ventura Benoin, GVK. m. K., Ⓒ, ✠C. ott. MO. 3., Kons. Kanzlei-R.
Fortunat Viktor, GVK. m. K., Ⓒ, ✠C. ott. OO. 4., Kons. Kanzlei-R.

Untergeordnete Ämter:

Vize-Konsulate.

In Chios.

Brazzafolli Francesco, ✠C. Hon. Vize-Kons.

* In Rhodos.

Barmann Anton, ✠C. prov. Gerent des Vize-Konsulates.

In Samos.

Miasir Oskar, ✠C. ott. OO. 4., Hon. Vize-Kons.

Konsular-Agentien.

In Metelin.

Bargigli Natale, GVK. m. K., Ⓒ, ✠C. ott. OO. 4., Hon. Kons. Agent.

* Konsulat in Canea.

Wein Jakob, Ⓜ, Ⓒ, ✠C. Honvéd-Leut. a. D., Kons.
Herzfeld Emerich, Rit. v., ✠C. Vize-Kons.

Untergeordnete Ämter:

Konsular-Agentien.

In Candia.

Terenzio Viktor, Hon. Vize-Kons. (ad pers.).

In Rettimo.

Trifilli Theodor, GVK. m. K., Ⓒ, ✠C. it. KO-R., gr. EO-R., ott. OO. 4., Hon. Vize-Kons. (ad pers.).

* General-Konsulat in Trapezunt.

Móricz v. Técső Peter, Ⓒ, ✠C. Käm., Gen. Kons. II. Kl., kön. ung. Kons. Oberrichter-Stellv.

Untergeordnete Ämter:

Vize-Konsulat.

In Samsun.

Torro A., ddl, ✠C. prov. Gerent des Vize-Konsulates.

Konsular-Agentie.

In Kerasunt.

Algazdi G., ✠C. prov. Gerent der Konsular-Agentie.

* Konsulat in Aleppo.

Poche Friedrich, prov. Gerent des Konsulates.

Untergeordnete Ämter:

Konsular-Agentien.

In Alessandretta.

Levante Emil, Ⓒ, ✠C. Hon. Kons. Agent, m. d. T. e. Hon. Vize-Kons.

In Mersina.

Daras Nikolaus, FJO-R., Ⓒ, ✠C. gr. EO-R., Hon. Kons. (ad pers.).

19

Fig. 1 Example page of *Schematismus* from 1910, highlighting the complexity of its structure, including multiple columns, hierarchical relationships, unique characters, and special annotations, such as the curly braces of which two of them are found in the middle of the page [4]

archies of several levels, and multimodal page designs that feature text alongside tables and complex lists. These characteristics represent the significant challenges digital historians and humanists encounter when processing large quantities of such documents. Thus, there has not been a comprehensive extraction of information, entire data sets or structures, such as information on hierarchies, on the detailed composition of administrative authorities, or simply careers or biographies with regard to the *Schematismus* yet. Some studies manually extracted relevant information, which proved tedious, time-consuming and prone to errors.

For years consideration has been given to publishing parts of the *Schematismus*-series as digital editions. Such undertakings have so far reliably failed because of the immense size of the task; the most important series of such handbooks, digitally published by the Austrian National Library[1], comprises 145 volumes compiled between 1702 and 1918 (Table 1).

[1] alex.onb.ac.at.

Table 1 Data sheet of the *Schematismus*

Key statistics
Total of 145 volumes
56 volumes in the 1700s
89 volumes between 1800 and 1918
Complete series from 1816 to 1848 and from 1876 to 1918
Ranging from ∼150 pages (1702) to ∼1900 pages (1918)

We calculated that this series contains between 130,000 and 150,000 printed pages. Neither manual extraction of the information contained therein nor automated processing of the pages, which have very diverse layouts, seemed feasible to us with the solutions currently available. The off-the-shelf solutions we tried, as well as generic layout detection integrated into well-established OCR did not perform too well with the *Schematismus*.

We assumed that a powerful layout detection, which can divide the individual cells of the *Schematismus* into meaningful text blocks, could possibly represent a relatively favourable solution. The complexity and diversity of the layouts led us to consider machine learning models as possible solutions, as we expected them to be able to handle the fractal nature of the page layout better than rule-based models. The first research question in this paper therefore investigates, whether high-quality layout detection as a preprocessing step could improve the performance of downstream OCR, in order to obtain a relatively simple solution for extracting the relevant information. We first identified and tested a suitable deep learning architecture—Faster R-CNN. Then we rebuilt the custom font used in the *Schematismus*. We used this font to synthesize a large amount of *Schematismus*-styled training data in the first step. Care was taken to ensure that the synthetic training data had a similarly complex and varied layout as the original data sets. The synthetically produced training data was also artificially distorted, dirtied, and twisted. We trained the Faster R-CNN model with the synthetic training data and further finetuned the model with a smaller number of manually annotated pages from the *Schematismus*.

Once our layout detection was working, we tested its potential with an off-the-shelf distribution of Tesseract.

Then we addressed the second research question, which investigated, to what degree a custom font finetuned distribution of Tesseract could boost OCR results. We finetuned an off-the-shelf distribution of Tesseract with the custom font we had created of the *Schematismus* and carried out a second test run.

For this study, we only invested limited resources in comparing possible OCR solutions, and we only took a cursory look at the necessary post-processing steps. However, the layout detection solution we present should be able to operate with any downstream OCR solution that can be integrated

into a Python pipeline. We expect that it might perform even stronger with OCR-engines developed specifically for historical fonts, such as Kraken OCR or OCR4all.

We have not devoted significant resources to the post-processing step so far, yet we ran some preliminary tests, using GPT4. The results were promising, and we expect that the latest generation of LLMs provides enormous potential to further improve the quality of the OCR results in a post-processing step.

2 Background & related work

Wide variation in document formats and degradation over time make automated layout detection and the development of generic tools quite challenging, especially for historical documents. For these reasons, a broad range of techniques and methods has been developed and applied over time, in order to be able to provide for a better automated information extraction from historical documents.

In this section, we will first present the state of research in history and more broadly the humanities, then initiatives to improve OCR that have been developed and tested in the field of historical document analysis are discussed.

2.1 Availability and limitations of OCRed primary sources

The retro-digitisation of either parts of or entire historical sources has been an issue among historians and fellow humanists since at least the 1950s, particularly with regard to serial sources [5]. However, for the larger part of the past seven decades, information extraction and the production of digital data that could be processed by computing machines, has been executed manually. Even though OCR software has been widely available by the 1990s at latest, particularly processing historical data has remained a complex issue [6], as the quality of results has been varying strongly [7]. We therefore observe a bifurcation in the field of digital historical and humanist research: On the one hand, large amounts of retro-digitised historical data are processed automatically by important providers of research data, such as www.archive.org [8]. Many national libraries, as for instance the Austrian National Library [9], the National Library of Finland [10] or the Munich Digitization Center of the Bavarian State Library [11], and further large transnational initiatives, as for example Europeana [12], also belong to this group. However, users of these data, such as historical research projects, are still relying on the manual transcription or annotation of digitised primary sources, even at scale.[2] Frequently, this is due to quality issues regarding OCRed documents available online.[3] Relevant databases that were build predominantly on data extracted by manual labour include the *Wiener Datenbank zur Europäischen Familiengeschichte* [13], *The Emperor's Desk* [14], the prosopographical data processed by *The Viennese Court* [15], and further the projects run in *Social Mobility of Elites* [16].

2.2 Common challenges

Historians and digital humanists working with retro-digitised text data are familiar with the phenomenon that OCR-ed texts provided on common platforms may vary in quality. This is mostly due to the fact that texts underwent OCR procedure at different times and different technologies were used. As a result, full-text searches frequently produce inhomogeneous results. Also, a "systematic" OCR error, i.e. distortions that are typical for certain OCR software, can no longer be processed in such a targeted manner if large text corpora consist of parts that were OCRed at different times with different software. In many projects, therefore, raw digitised documents are now re-OCRed, although many historical layouts still pose a challenge for standard OCR but also specialised software [10].[4] Apart from specialised solutions [17], Tesseract OCR and Transkribus are currently regarded as reference standards in the field of historical OCR, but still encounter limitations that require a high level of manual effort, given the special layouts and structures that will be discussed here [18]. Further, OCR4all offers a toolbox for open source OCR applications that has proved highly performant lately [19], as well as kraken OCR, developed by the EHESS and closely associated with the digital research environment eScriptorium.[5] Libraries have begun to use the potential of ML to the classification of large quantities of texts [20], especially in the context of research libraries [21].

For historical research, important primary sources such as censuses of the Habsburg Empire have been mostly digitized manually [22, 23]. Such sources usually feature a relatively limited volume and a highly homogeneous layout and struc-

[2] There is very little documentation on how data in historical research projects is OCRed, so we have to refer to information that we obtained in personal conversation with several dozen colleagues over the past one-and-a-half decade. Based on the information available to us, we would assume that manual transcription of historical text, but especially tabular data, is still the standard procedure, even though this usually starts with a raw document created with standard OCR (Tesseract or Transkribus), which is subsequently improved. The digitisation of tabular data is particularly challenging because the target structure usually has to be created manually beforehand. Cf. [6].

[3] An entire panel in the recent 15th Austrian Contemporary History Day was devoted to the manual and semi-automatized correction of flawed OCR data from large repositories.

[4] Details about the OCR process at large providers of cultural heritage data are only available to a very limited extent, and the publicly accessible documentation leaves much to be desired, even at public institutions.

[5] https://github.com/mittagessen/kraken.

ture. However, because of the enormous volume, this method is not feasible for other complex source works, such as the *Schematismus*. Even in the last initiative we know of, the size and complexity of the *Schematismus* was ultimately considered insurmountable for only partially automated data extraction, and manual edition of a small part was envisaged as an alternative. Due to these obstacles, very little research is engaging with a deeper exploration of the information stored in the *Schematismus*, the work of Bavouzet [24] (building entirely on manually extracted data) is clearly standing out as a beacon here.

Only recently have developments in new fields of research such as document intelligence begun to open up the possibilities of machine learning for this area on a larger scale [25–27]. Impressive progress has been made in some areas [28].

2.3 Pre-OCR steps to enhance OCR quality

The process of extracting information from historical documents can also be considered as a complete extraction pipeline, instead of individual tasks. Monnier and Aubry present work on an extraction tool for historical documents, which provides improvements in terms of robustness and extraction performance due to mutual reinforcement of text line and image segmentation [29]. The task of segmentation is also highlighted by Gruber et al. [30] where the authors also propose to conduct preprocessing of the image before conducting OCR . Such approaches have already been explored in the past, for example by processing the background of the image [31]. Consequently, techniques such as Generative Adversarial Networks have been explored to achieve super-resolving of the input images [32]. Augmentation has been used on several occasions lately in order to develop economic ways to scale up information extraction from historical documents. Grüning et al. use the deep neural network ARU-Net to address the issue of line detection in historical documents [33]. Martínek et al.[18] realise an approach, which combines fine-tuning OCR-engines with comparatively little data, after training the engine with large synthetic data-sets.

Document layout analysis has received increasing attention in the past few years, though this area remains under-explored [34]. Solutions found in this field are not always tackling specifically historical problems, as for instance [35]. It has been recognised though that the increasing availability and usability of deep neural networks, in particular CNNs, offers entirely new opportunities for the development of custom-made solutions for certain document layout analysis tasks [36]. With regard to OCR, line detection is frequently considered more relevant than layout detection [28]. Nevertheless, significant progress has been made in this area in recent years, with layout detection often being understood as part of a more complex single stage process [1, 26, 27]. LayoutLM presents a promising and versatile solution to address many common tasks in this area [37]. We also tried LayoutLM for the *Schematismus*, but it did not prove efficient with this particular type of documents.

Table 2 Approximate number of persons mentioned in each *Schematismus* per year and the number of pages in the respective volume

Year	Persons	Pages
1702	2200	150
1750	4000	430
1801	6700	1030
1820	26,400	1300
1848	28,800	1200
1860	15,000	300
1878	51,200	1100
1910	93,400	1600
1918	110,400	1900

We calculated these values by averaging the number of bounding boxes predicted on each name register page and multiplying this average by the number of name index pages listed in each year's name register

3 Methodology

State Manuals and *Hof- und Staatsschematismen* for the Habsburg Empire in particular are commonly provided in PDF format. These documents are accessible via various historical document repositories such as the Austrian National Library[6], the Munich Digitalization Center (MDZ) of the Bavarian State Library[7] or archive.org[8]. Frequently, however, plain text is not available at all or is available in sub-optimal quality.

Whereas the general issue of suboptimal OCR quality is widely acknowledged in historical and further humanist research, most approaches consider layout a secondary line of attack, when it comes to improving OCR quality. There are few exceptions, as for instance [34, 36, 38]. Even the two probably most important and most widely used out-of-the-box solutions in the field of historical OCR, different distributions of Tesseract and the variety of different OCR and HTR models provided by Transkribus, require significant (manual) effort in data preprocessing, when it comes to extracting information from digitised primary sources. Even conventional document layouts often pose a challenge [6]. Lately, OCR4all has come up with a strong layout detection.

Complex layouts as encountered in the *Schematismus*, however, are still considered a major challenge by the his-

[6] alex.onb.ac.at.

[7] www.digitale-sammlungen.de.

[8] www.archive.org.

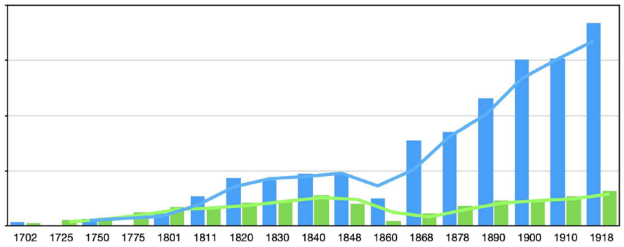


Fig. 2 Number of persons (blue) and pages (green, 10fold augmentation) in different volumes of the *Schematismus*

torical research community. Efficient information extraction from such documents is considered a difficult and complicated task. Several efforts to automatically extract structured data from the *Schematismus* have failed so far. Generally, particularly in historical texts, OCR is performed page-wise, which means that snippets, belonging to different blocks of text, are performed serially, as common OCR processors work line-wise.

3.1 Approach and research questions

For our approach, we identified two points of attack, which correspond with the two research questions we formulated. First, we wanted to find out, whether AI-driven layout detection prior to OCR could significantly boost OCR quality. Second, we were interested to see, to what degree finetuning the OCR engine could further improve its performance. To tackle research question 1 in a first step, we split the individual document pages into their layout elements, in order to preserve the context of the different blocks of text. To this end, we used a deep learning convolutional neural network. This is a complex ML algorithm originally developed for object localization and object recognition. We assumed this approach would be suitable, since we consider the identification of large coherent blocks a computer vision problem.

In the next step, we used an OCR algorithm to process the individual image snippets, rather than the entire document page image. Then we addressed the second research question, to which degree OCR accuracy could be improved by finetuning the standard OCR tool Tesseract on a custom font, which was designed to look as similar as possible to the original font used in the *Schematismus* documents.

We chose a sample from 145 volumes of State Manuals and focused on editions that were published in the second half of the 19th century onward, for two reasons:

1. The task is becoming slightly more complex for the decades prior to 1848, as the fonts that were used are more diverse and complex. We do not consider this a major problem, yet for the proof of concept we were interested in streamlining the entire research process and to eliminate additional complexities that were not in our primary line of attack.
2. Even though State Manuals are available for a period of more than 200 years, the mass of data was produced from the 1850s onward, therefore the yield for a solution capable of dealing with documents of this type is expected to be very high (compare Fig. 2 and Table 2).

To deal with research question 1, we built a machine learning model that can segment retrodigitised PDF-documents of the *Hof- und Staatshandbücher* and split these into their layout-structure elements, such as individual paragraphs and headings. Each of these image snippets was subsequently fed into Tesseract for text extraction. Figure 3 shows a simplified version of this process.

We used two Tesseract OCR models. For research question 2, an instance of Tesseract that had been fine-tuned on our custom font. For comparison and performance assessment, we also used an instance that had not been fine-tuned. OCR accuracy was then calculated by comparing the extracted text with the manually transcribed ground truth. Then, this process was repeated, but without dividing the page into individual segments. To answer the research questions in this study, the accuracies were finally compared in order to evaluate the efficiency of our approach.

The following subsections will describe in detail the implementation of the different methods that we employed to process PDFs of the *Hof- und Staatshandbücher*. This includes training data set generation for the development of the layout detection, layout detection itself, and optical character recognition. Each of these three steps constituted a work package of its own. All the research and analysis was conducted within Jupyter Notebook, the diverse tools that were put to use are listed in the following subsections.

3.2 Data set generation

In order to successfully train a convolutional neural network, a sufficient amount of labeled data is required. Creating a training data set by manually drawing bounding boxes on a large quantity of PDFs drawn from historical source documents is time-consuming and labour-intensive due to the large number of pages that would have to be annotated. Therefore, we developed an alternative approach to artificially generate labeled training data. We wrote a Python script that is designed to generate synthetic documents mimicking the style of the *Hof- und Staatshandbücher*. This Python script generated Latex-code, which was then compiled using luatex [39] to create a PDF file. In the course of this process, the coordinates of the individual text structure elements were determined, which is described in more detail in the following section.

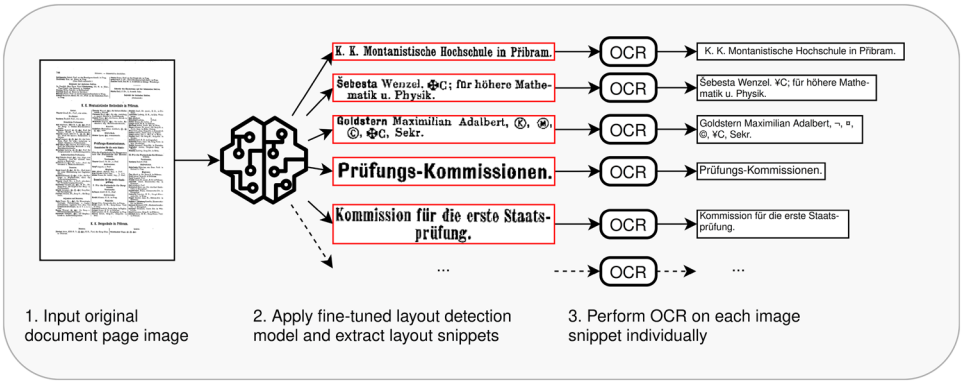


Fig. 3 This flowchart illustrates how each extracted layout element is processed by OCR

In order for the generated data to be used as training data, it was imperative that the created documents appeared as visually realistic as possible compared to the original documents. Reverse engineering the original documents and paying attention to detail were therefore essential to the creation of synthetic training data.

Due to the fact that each paragraph begins with a last name and a first name, a data set containing thousands of first and last names [40] was randomly sampled in order to obtain names. The pool of samples was restricted to Austria, Hungary, Switzerland, and Germany.

A list of abbreviation explanations from the 1910 *Schematismus* document was manually transcribed and randomly selected for the text following the names. To vary the length of individual generated paragraphs, the number of sampled texts was also randomly chosen.

Figure 1 displays a page from the original 1910 *Schematismus* document. Variable column numbers are a key characteristic of such documents and are used almost everywhere. While most sections have three columns, there can be variations in certain sections. For instance, Fig. 8 shows four columns in the name-index section. Thus, generating realistic synthetic documents required the use of the same column layout.

Another key visual element is the relatively distinct font type. Research led us to a font called "Opera-Lyrics-Smooth" that appeared very similar to the original. Even though it represented already a good match, we decided to invest additional effort: Using the open source program "FontForge" [41], we further customised the font in accordance with the original. In order to achieve the best possible match to the original font, screenshots of every letter in the original documents were taken manually, and then the existing letters in the font were adapted according to the screenshots taken.

Another distinctive feature of the *Hof- und Staatsschematismus* is its excessive use of particular symbols, representing orders and similar distinctions of the persons listed in this source. The use of these symbols allowed the further condensation of information stored in the *Schematismus*.

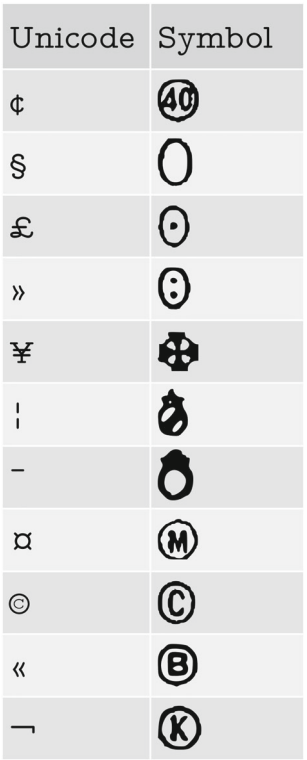


Fig. 4 Illustration of how Unicode characters are mapped to *Schematismus* symbols

Nonstandard symbols were mapped to Unicode symbols, which were unlikely to be needed for document generation. Figure 4 illustrates this mapping.

This method resulted in the creation of three font types: one for general text paragraphs, one for headlines, and one for italics. An example of this can be seen in Figs. 5 and 6, which shows an original paragraph and a corresponding paragraph reproduced using the custom font. Additionally, when reviewing some original *Schematismus* documents, it is apparent that font sizes and alignments vary considerably from section to section or even from page to page. In particular, the difference can be observed when examining the headlines of the original documents. To create realistic head-

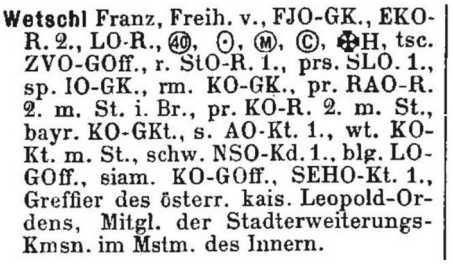

Wetschl Franz, Freih. v., FJO-GK., EKO-R. 2., LO-R., ㊵, ⊙, Ⓜ, ©, ✠H, tsc. ZVO-GOff., r. StO-R. 1., prs. SLO. 1., sp. IO-GK., rm. KO-GK., pr. RAO-R. 2. m. St. i. Br., pr. KO-R. 2. m. St., bayr. KO-GKt., s. AO-Kt. 1., wt. KO-Kt. m. St., schw. NSO-Kd. 1., blg. LO-GOff., siam. KO-GOff., SEHO-Kt. 1., Greffier des österr. kais. Leopold-Ordens, Mitgl. der Stadterweiterungs-Kmsn. im Mstm. des Innern.

Fig. 5 Original paragraph, taken from an existing state manual from the year 1910

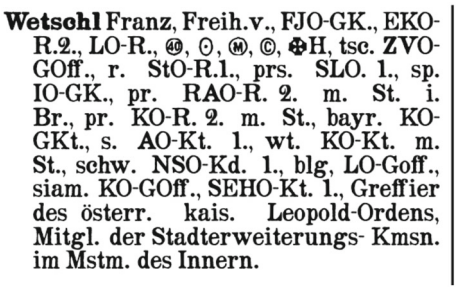

Wetschl Franz, Freih.v., FJO-GK., EKO-R.2., LO-R., ㊵, ⊙, Ⓜ, ©, ✠H, tsc. ZVO-GOff., r. StO-R.1., prs. SLO. 1., sp. IO-GK., pr. RAO-R. 2. m. St. i. Br., pr. KO-R. 2. m. St., bayr. KO-GKt., s. AO-Kt. 1., wt. KO-Kt. m. St., schw. NSO-Kd. 1., blg, LO-Goff., siam. KO-GOff., SEHO-Kt. 1., Greffier des österr. kais. Leopold-Ordens, Mitgl. der Stadterweiterungs- Kmsn. im Mstm. des Innern.

Fig. 6 Corresponding synthetically generated paragraph reproduced with a custom font

lines, four headline types ranging from "H1" to "H4" were used to emulate this feature. The first element was the largest and the rest gradually decreased in size. Table 3 provides a list of all the different classes.

Finally, it is crucial to emphasise some small but very significant visual details. Every paragraph begins with one or two words in bold, the last name of the individual, followed by the first name and some additional titles and awards. Indentation occurs after the first line if the text is too long for a single line. Furthermore, multiple individuals may be grouped together within a large curly bracket, as can be seen in Fig. 1. In name-index pages, multiple individuals with the same last name may be grouped by adding a horizontal line at the beginning, which can be observed in Fig. 7. Additionally, every entry within this section is accompanied by one or more numbers that indicate the page number. Finally, it should be noted that headings are usually centered on the page or within columns and that the end of every text is always marked with a period.

In order to be able to produce text for the synthetic *Schematismus*-style training documents, several sources were consulted. For the purpose of generating large headlines, a simple list of historical Austrian orders and decorations was used [42]. In order to create headlines with smaller font sizes, a combination of years as strings and Austrian municipality names was used. For paragraph generation, two sources were consulted, as previously mentioned. Through the use of all the above methods and visual keys, we were able to create a large number of realistic looking synthetic documents. An example of such a synthetic *Schematismus*-style document can be seen in Figure 7.

In addition to generating a synthetic data set of *Schematismus* documents for training a machine learning model, annotations for each element of the structure had to be generated along with the generation of the document in order to make this data set effective for training. As part of the process of detecting and localising objects, in this case the layout elements, bounding boxes were used to define the location and size of the individual structure elements within an image. The labels accompanying the bounding boxes indicate the class of the corresponding box, such as "paragraph" or "H1", which are necessary for classification tasks performed by the machine learning method.

A latex package called "zref-savepos" is used to save the position of characters on the current page and write these coordinates to an external file at compilation time. Using the coordinates, bounding boxes could be computed by parsing the external file. As the individual text elements had already been generated earlier in the same Python script, it was known which label had to be associated with the respective bounding box. In order to construct the data set, the generated documents, which were compiled by the latex compiler and then saved as PDF files, were converted to images. In addition to the image file, a Pascal VOC XML file containing the corresponding annotations was created and stored in a separate directory. A total of 3766 synthetic *Schematismus* documents have been generated using this approach. Figure 9 shows a synthetic document with its corresponding annotations overlayed.

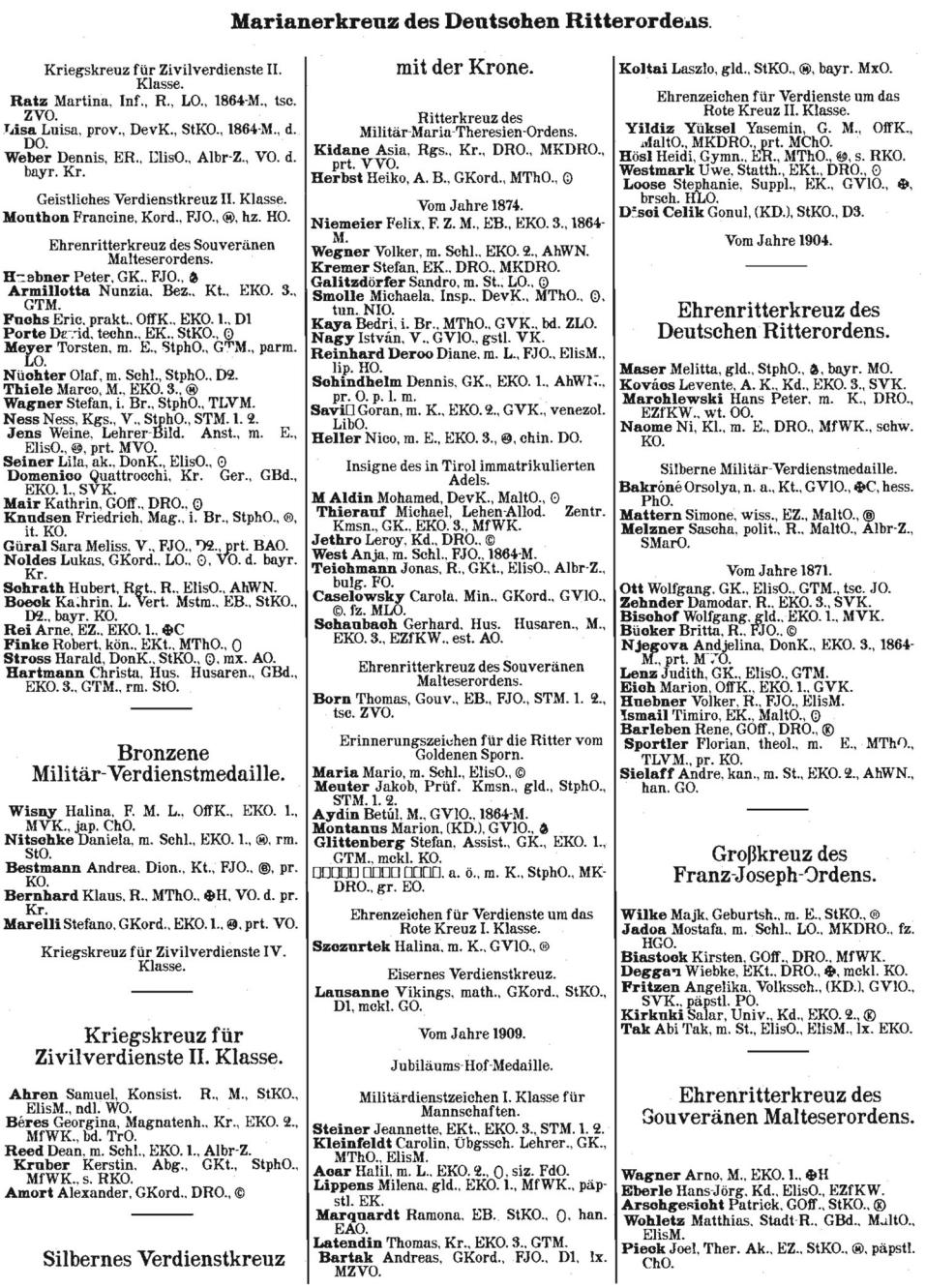
3.3 Layout detection

For the actual layout detection model to be configured in the next step, we chose a faster region-based convolutional neural network (faster R-CNN) built on a ResNet-50 backbone. The model was created and trained using the PyTorch [43] framework. In PyTorch, version-2 of the faster R-CNN implementation was used [44]. Training of the model was conducted on an Nvidia RTX 4090 with 24 GB of video memory.

3.3.1 Model training and settings

Even though we used primarily the default settings of the model, we found that some adjustments had a significant impact on the model's performance. The following paragraphs will provide a detailed description of these adjustments.

1. By setting the pre-trained parameter to "True", the training speed has improved in a way that earlier epochs of training have already begun with a lower training and



Marianerkreuz des Deutschen Ritterordens.

Kriegskreuz für Zivilverdienste II. Klasse.

Ratz Martina, Inf., R., LO., 1864-M., tse. ZVO.
Lisa Luisa, prov., DevK., StKO., 1864-M., d. DO.
Weber Dennis, ER., LlisO., Albr-Z., VO. d. bayr. Kr.

Geistliches Verdienstkreuz II. Klasse.

Mouthon Francine, Kord., FJO., ⊛, hz. HO.

Ehrenritterkreuz des Souveränen Malteserordens.

Hübner Peter, GK., FJO., ♁
Armillotta Nunzia, Bez., Kt., EKO. 3., GTM.
Fuchs Eric, prakt., OffK., EKO. 1., Dl
Porte David, techn., EK., StKO., ⊙
Meyer Torsten, m. E., StphO., GTM., parm. LO.
Nüchter Olaf, m. Schl., StphO., D2.
Thiele Marco, M., EKO. 3., ⊛
Wagner Stefan, i. Br., StphO., TLVM.
Ness Ness, Kgs., V., StphO., STM. 1. 2.
Jens Weine, Lehrer-Bild. Anst., m. E., ElisO., ⊛, prt. MVO.
Seiner Lila, ak., DonK., ElisO., ⊙
Domenico Quattrocchi, Kr. Ger., GBd., EKO. 1., SVK.
Mair Kathrin, GOff., DRO., ⊙
Knudsen Friedrich, Mag., i. Br., StphO., ⊛, it. KO.
Güral Sara Meliss, V., FJO., D2., prt. BAO.
Noldes Lukas, GKord., LO., ⊙, VO. d. bayr. Kr.
Schrath Hubert, Rgt., R., ElisO., AhWN.
Boeck Kathrin, L. Vert. Mstm., EB., StKO., D2., bayr. KO.
Rei Arne, EZ., EKO. 1., ✠C
Finke Robert, kön., EKt., MThO., ⊙
Stross Harald, DonK., StKO., ⊙, mx. AO.
Hartmann Christa, Hus. Husaren., GBd., EKO. 3., GTM., rm. StO.

Bronzene Militär-Verdienstmedaille.

Wisny Halina, F. M. L., OffK., EKO. 1., MVK., jap. ChO.
Nitsohke Daniela, m. Schl., EKO. 1., ⊛, rm. StO.
Bestmann Andrea, Dion., Kt., FJO., ⊛, pr. KO.
Bernhard Klaus, R., MThO., ✠H, VO. d. pr. Kr.
Marelli Stefano, GKord., EKO. 1., ⊛, prt. VO.

Kriegskreuz für Zivilverdienste IV. Klasse.

Kriegskreuz für Zivilverdienste II. Klasse.

Ahren Samuel, Konsist. R., M., StKO., ElisM., ndl. WO.
Béres Georgina, Magnatenh., Kr., EKO. 2., MfWK., bd. TrO.
Reed Dean, m. Schl., EKO. 1., Albr-Z.
Kruber Kerstin, Abg., GKt., StphO., MfWK., s. RKO.
Amort Alexander, GKord., DRO., ©

Silbernes Verdienstkreuz mit der Krone.

Ritterkreuz des Militär-Maria-Theresien-Ordens.

Kidane Asia, Rgs., Kr., DRO., MKDRO., prt. VVO.
Herbst Heiko, A. B., GKord., MThO., ⊙

Vom Jahre 1874.

Niemeier Felix, F. Z. M., EB., EKO. 3., 1864-M.
Wegner Volker, m. Schl., EKO. 2., AhWN.
Kremer Stefan, EK., DRO., MKDRO.
Galitzdörfer Sandro, m. St., LO., ⊙
Smolle Michaela, Insp., DevK., MThO., ⊙, tun. NIO.
Kaya Bedri, i. Br., MThO., GVK., bd. ZLO.
Nagy István, V., GVlO., gstl. VK.
Reinhard Deroo Diane, m. L., FJO., ElisM., lip. HO.
Schindhelm Dennis, GK., EKO. 1., AhWN., pr. O. p. l. m.
Savi Goran, m. K., EKO. 2., GVK., venezol. LibO.
Heller Nico, m. E., EKO. 3., ⊛, chin. DO.

Insigne des in Tirol immatrikulierten Adels.

M Aldin Mohamed, DevK., MaltO., ⊙
Thierauf Michael, Lehen-Allod. Zentr. Kmsn., GK., EKO. 3., MfWK.
Jethro Leroy, Kd., DRO., ©
West Anja, m. Schl., FJO., 1864-M.
Teichmann Jonas, R., GKt., ElisO., Albr-Z., bulg. FO.
Caselowsky Carola, Min., GKord., GVlO., ©, fz. MLO.
Schaubach Gerhard, Hus. Husaren., M., EKO. 3., EZfKW., est. AO.

Ehrenritterkreuz des Souveränen Malteserordens.

Born Thomas, Gouv., EB., FJO., STM. 1. 2., tse. ZVO.

Erinnerungszeichen für die Ritter vom Goldenen Sporn.

Maria Mario, m. Schl., ElisO., ©
Meuter Jakob, Prüf. Kmsn., gld., StphO., STM. 1. 2.
Aydin Betül, M., GVlO., 1864-M.
Montanus Marion, (KD.), GVlO., ♁
Glittenberg Stefan, Assist., GK., EKO. 1., GTM., mckl. KO.
□□□□□ □□□□ □□□□, a. ö., m. K., StphO., MKDRO., gr. EO.

Ehrenzeichen für Verdienste um das Rote Kreuz I. Klasse.

Szczurtek Halina, m. K., GVlO., ⊛

Eisernes Verdienstkreuz.

Lausanne Vikings, math., GKord., StKO., Dl, mckl. GO.

Vom Jahre 1909.

Jubiläums-Hof-Medaille.

Militärdienstzeichen I. Klasse für Mannschaften.

Steiner Jeannette, EKt., EKO. 3., STM. 1. 2.
Kleinfeldt Carolin, Übgssch. Lehrer., GK., MThO., ElisM.
Acar Halil, m. L., EKO. 2., ⊙, siz. FdO.
Lippens Milena, gld., EKO. 1., MfWK., päpstl. EK.
Marquardt Ramona, EB., StKO., ⊙, han. EAO.
Latendin Thomas, Kr., EKO. 3., GTM.
Bartak Andreas, GKord., FJO., Dl, lx. MZVO.
Koltai Laszlo, gld., StKO., ⊛, bayr. MxO.

Ehrenzeichen für Verdienste um das Rote Kreuz II. Klasse.

Yildiz Yüksel Yasemin, G. M., OffK., MaltO., MKDRO., prt. MChO.
Hösl Heidi, Gymn., ER., MThO., ⊛, s. RKO.
Westmark Uwe, Statth., EKt., DRO., ⊙
Loose Stephanie, Suppl., EK., GVlO., ✠, brsch. HLO.
Dəsei Celik Gonul, (KD.), StKO., D3.

Vom Jahre 1904.

Ehrenritterkreuz des Deutschen Ritterordens.

Maser Melitta, gld., StphO., ♁, bayr. MO.
Kovács Levente, A. K., Kd., EKO. 3., SVK.
Marchlewski Hans Peter, m. K., DRO., EZfKW., wt. OO.
Naome Ni, Kl., m. E., DRO., MfWK., schw. KO.

Silberne Militär-Verdienstmedaille.

Bakróné Orsolya, n. a., Kt., GVlO., ✠C, hess. PhO.
Mattern Simone, wiss., EZ., MaltO., ⊛
Melzner Sascha, polit., R., MaltO., Albr-Z., SMarO.

Vom Jahre 1871.

Ott Wolfgang, GK., ElisO., GTM., tse. JO.
Zehnder Damodar, R., EKO. 3., SVK.
Bischof Wolfgang, gld., EKO. 1., MVK.
Bücker Britta, R., FJO., ©
Njegova Andjelina, DonK., EKO. 3., 1864-M., prt. MVO.
Lenz Judith, GK., ElisO., GTM.
Eich Marion, OffK., EKO. 1., GVK.
Huebner Volker, R., FJO., ElisM.
Ismail Timiro, EK., MaltO., ⊙
Barleben Rene, GOff., DRO., ✠
Sportler Florian, theol., m. E., MThO., TLVM., pr. KO.
Sielaff Andre, kan., m. St., EKO. 2., AhWN., han. GO.

Großkreuz des Franz-Joseph-Ordens.

Wilke Majk, Geburtsh., m. E., StKO., ⊛
Jadoa Mostafa, m. Schl., LO., MKDRO., fz. HGO.
Biastock Kirsten, GOff., DRO., MfWK.
Deggau Wiebke, EKt., DRO., ✠, mckl. KO.
Fritzen Angelika, Volkssch., (KD.), GVlO., SVK., päpstl. PO.
Kirkuki Salar, Univ., Kd., EKO. 2., ✠
Tak Abi Tak, m. St., ElisO., ElisM., lx. EKO.

Ehrenritterkreuz des Souveränen Malteserordens.

Wagner Arno, M., EKO. 1., ✠H
Eberle Hans-Jörg, Kd., ElisO., EZfKW.
Arsohgesicht Patrick, GOff., StKO., ✠
Wohletz Matthias, Stadt-R., GBd., MaltO., ElisM.
Pieck Joel, Ther. Ak., EZ., StKO., ⊛, päpstl. ChO.

Fig. 7 Example of a synthetic *Schematismus*-style document used in training-set

validation loss than with randomly initialised starting weights.

2. A further parameter that has been tweaked is the anchor generation of the region proposal network in the faster R-CNN model. When identifying areas of interest in an image, anchor boxes are critical because they determine where to look. Thus, they play an essential role when it comes to detecting layout elements within a document image. In order to be able to accommodate a variety of different types of objects, anchor boxes were selected with different aspect ratios and scales. It is imperative to note that there are multiple anchor boxes applied to each sliding window position in the region proposal network. Therefore, it is logical to specify these ratios and scales according to the shape of the objects. Thus, the minimum, maximum, and mean ratio and scale of all bounding boxes within the 3766 generated *Schematismus* documents have been calculated and used as a guideline to set the anchor-generation parameters. A general characteristic of the anchor boxes chosen is their elongated and narrow shape, which is understandable, given that text lines have a similar shape. Additionally, some

Table 3 Lists of the class types used to generate synthetic *Schematismus* documents

Class name	Class description
Paragraph	Synthetic documents are primarily composed of paragraphs. Paragraphs always begin with a bolded word and are indented after the first line.
BigParagraph	A big paragraph is a paragraph with a bold starting word, but an inverted indentation compared to a normal paragraph. In addition, this type of structure element may never appear within a column.
H1	H1 headlines represent headings that appear above the start of a column. This class may also never appear within a column.
H2	H2 headlines have a slightly smaller font size than H1 headings, but will always be contained within a column.
H3	An H3 header is a title that has the same font size as a paragraph and may only appear within a column. The class is also used to define keywords on the right-hand side of a curly bracket.
H4	An H4 header is formatted in the same way as an H3, but italicized and always enclosed in parentheses.
NameEntry	Name entries consist of a bold text that is left aligned followed by one or more numbers that are right aligned.
Curly	Using the curly class, multiple paragraphs and a single H3 class can be contained within. In this manner, it should be easier to assign individual paragraphs to a H3 class in the future.

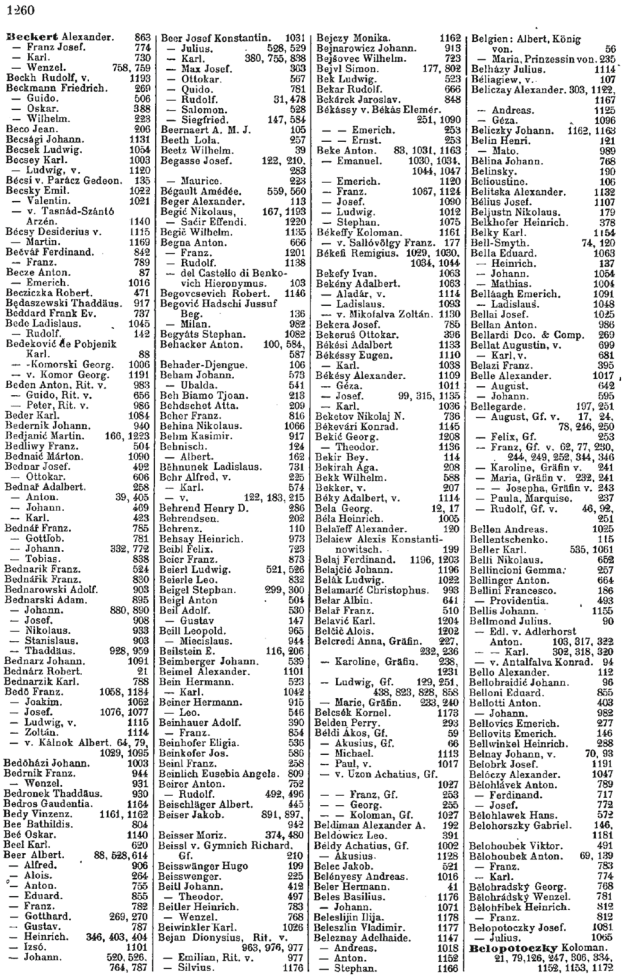


Fig. 8 Example of an original page from the name index section

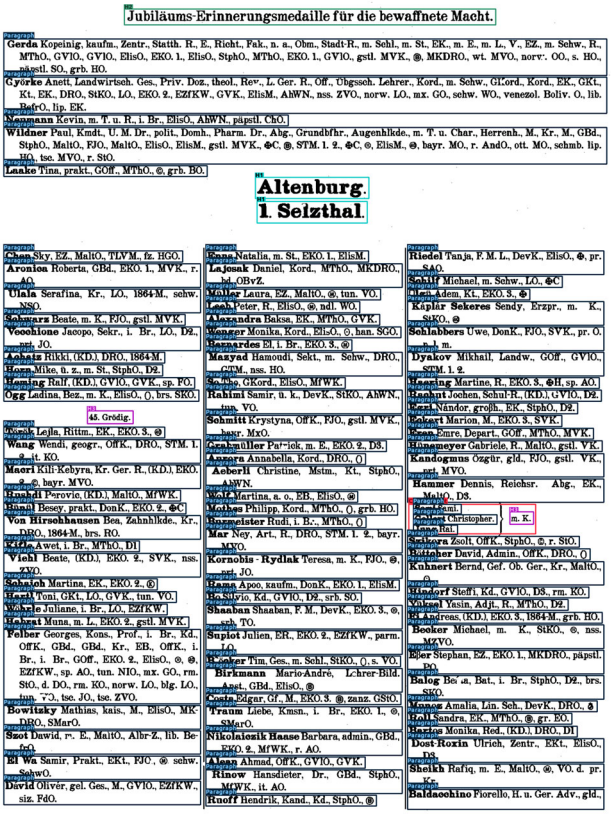


Fig. 9 Example of generated *Schematismus*-style document alongside with corresponding annotations (bouding boxes and class label) used in the training set

objects, such as single lines or headings, were quite small, so the anchor boxes generated were smaller than usual.
3. Further, the maximum number of object detections per image needed to be adjusted. Since most object detection models detect only a few objects at a time within a single image, the default value is 100 objects. It should be noted, however, that since the purpose of this analysis is to detect quite fine-grained layout elements within these *Schematismus* documents, the number of layout elements within one document may easily exceed 250 (some pages, for instance from the name register of the *Schematismus*, feature up to 400 objects on one page). In order to completely disregard this upper limit of detections per image, the parameter was set to 1000.
4. Another significant adjustment has been made to the resolution of the images. It should be noted that while the standard image input resolution of the faster R-CNN model is 1333 × 800 pixels, this resolution results in the model sometimes being unable to detect smaller objects such as single lines within the *Schematismus* documents. The reason for this is that both convolution layers and pooling layers in the model further downscale the input image, resulting in loss of important information. Based on the experiments conducted, a resolution of 1988 × 1405 pixels has been determined as the input resolution.

As mentioned previously, 3766 synthetic *Schematismus*-style documents have been generated for the purpose of creating a fully annotated data set. Among these, 3126 served as the training set and 640 served as the validation set. A stratified split of the full data set was used to select the training and validation sets. In our scenario, the stratification process involved counting the occurrences of each class within each generated document page. This information was then used to split the data into two sets, maintaining a similar class distribution in both the validation and training sets. Even though there were more than 3000 annotated training documents available, dataset augmentation strategies have been applied. Adding random augmentations to existing data allowed the training set to be artificially expanded in terms of document variety without increasing the number of documents and thus increasing training time. Therefore, adding augmentations to the training of a faster R-CNN layout detection model of documents can improve its accuracy and robustness.

3.3.2 Training data augmentation and further steps

As a result of applying random transformations such as rotation, scaling, cropping, optical distortion (to simulate page warping), blur, noise adding and page flipping, the model could learn to recognise and locate different layout elements invariant of their angle or size. Considering that all of these parameters were selected randomly, it is extremely unlikely that two identical documents will be input into the model during training. Additionally, augmentation may help to reduce overfitting, which occurs when a model becomes too specialised in recognising only training examples and performs poorly on new, unknown data. By augmenting the training data, the model is exposed to a wider range of layout variations and becomes more adaptable to new and unseen documents.

As for the actual training process, adjustments have been made to the number of epochs, batch size, and learning rate. During each training iteration, batch size determines the number of samples, and thus images, to be processed by the machine learning model before the weights are updated. It is one of the most influential hyperparameters when training deep learning models, and it can be viewed as a trade-off between accuracy and speed. A larger batch size allows more samples to be processed at once, resulting in faster training times and better hardware utilisation. However, larger batch sizes require more memory and may hinder generalisation [45].

By contrast, a smaller batch size results in fewer samples being processed at once. Despite slower training times, this can also prevent, to some extent, overfitting and produce a more generalisable model. Typically, smaller batch sizes are used when the data set is small or when the model requires frequent updating of a large number of parameters. Choosing the correct batch size cannot be achieved in a one-size-fits-all manner since it is heavily dependent on the data set being used. For our study, a batch-size of two has been selected based on experimentation since it fully utilises GPU memory and, along with a scaling factor of 85 percent, produces a relatively fast training process.

In order to maintain a constant variance in gradient expectations, it is recommended to multiply the learning rate by $\sqrt{k}$ when multiplying the batch size by k [46]. As a result of extensive learning rate optimisation, a base learning rate of 0.005 has been chosen for batch size one. Thus, the final learning rate is $0.005 \cdot \sqrt{2} \approx 0.007$. This learning rate, along with a weight decay of 0.0005 and a momentum of 0.9, was used to initialise a stochastic gradient descent optimiser. The total number of epochs was set to 100. The model is saved after every epoch if the validation loss is less than the previous saved validation loss. In this manner, one can be assured that the final model, which has been trained for 100 epochs, is the one which worked best on the validation set.

The model, which has been trained purely on synthetic data, gave fairly solid results when used on original documents, as described in detail in the Evaluation section. While these results are already promising, the existing model can also be further fine-tuned using real, original documents. As annotations could not be generated this time, they had to be hand drawn, which is a very time-consuming process.

PyTorch's TorchServe [47] framework has, however, accelerated this process significantly. Employing this method allowed us to "serve" an existing trained model over the local network, where one could send images to and receive bounding-box and label prediction information. In theory, this is not much different from a simple Python script that runs a model directly to predict the layout elements of documents that are fed into it, but there are some very specific applications it can be used for. With the help of an annotation software called "BoundingBoxEditor" [48] the pre-trained model can be "served" and therefore accelerate the manual annotation process of the original documents significantly. This is due to the fact that the pre-trained model already yielded good results. Therefore, only a few adjustments and error corrections were necessary, such as correcting incorrect classifications or bounding boxes. A total of 39 original *Schematismus* documents have been manually annotated and saved in this way. Once the original *Schematismus* documents had been annotated, the annotations were utilised to fine-tune the existing model. To that purpose, the newly created data set had been divided into training and validation sets. Using the pre-trained model's weights as initial weights, these sets were trained for 100 epochs using the same parameters described earlier. Figure 12 show the training loss and validation loss for this training process.

3.3.3 Layout detection post-processing

When analysing the predictions in detail, it becomes apparent that some bounding boxes are overlapped, leading to sometimes incorrect predictions. An illustration of this phenomenon can be found in Figure 11, in which two different bounding boxes overlap on the second line.

To address this issue, a bounding box and label classification post-processing step has been developed. The process works by iterating over every predicted bounding box and then calculating the intersection over union (IoU) with every other bounding box. Essentially, the IoU represents the ratio between the overlapping area and the union area, and the closer the IoU is to 1.0, the more similar the bounding boxes are. As soon as the IoU has been calculated for every bounding box, merge candidates are identified by selecting boxes with an IoU score higher than 0.3. As a result, a maximum bounding box is calculated, which encompasses all merge candidates (including the box currently being viewed), and is used to replace the original bounding box. Considering that the merge candidates may be of different classes, the merged bounding box will be labelled with the class tag of the highest confidence score. It should be noted that bounding boxes associated with the "Curly" class will not be selected as merge candidates, as the nature of this class is to have enclosed boxes inside them. Figure 11 illustrates the overlay of the predicted boxes following the application of the post-processing step.

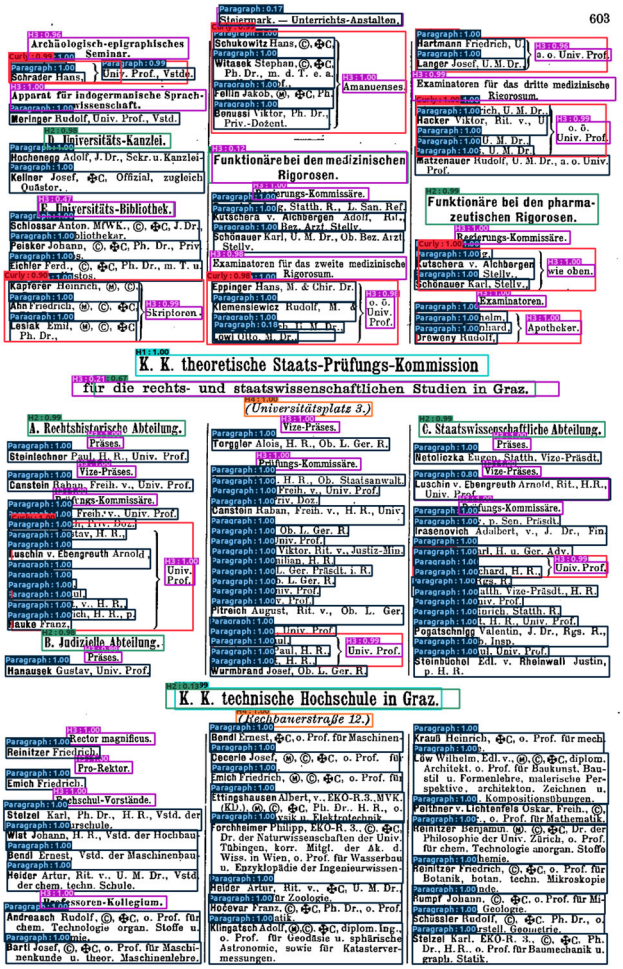


Fig. 10 This figure illustrates a randomly selected original *Schematismus* document with layout elements overlayed. Element predictions with confidence levels less than 0.1 have been omitted. The faster R-CNN model used to get these prediction results has been fine-tuned on 39 original *Schematismus* documents

Fig. 11 The figure illustrates that some bounding boxes overlap and have ambiguous label classifications, there are two overlapping bounding boxes in the middle (magenta and teal)

3.4 OCR

To extract the text within the individual elements of the predicted layout, Tesseract 5.0 was used [49]. As mentioned in Subsection 3.2, the font used in *Schematismus* documents is no longer commonly used. Despite the fact that Tesseract has been pre-trained on a number of fonts, it makes sense to utilise the custom fonts developed for the generation of

Schematismus documents to fine-tune the Tesseract optical character recognition model. In addition, due to the symbols that are used in the documents (see Fig. 4) which are unique to *Schematismus* documents, training on this font is necessary in order to recognise these symbols. To fine-tune Tesseract on such a font, images containing a single block of text rendered in the font must be generated. For this, Tesseract's built-in function "text2image" has been utilised. This function requires a large textfile of training text as one of its parameters. Despite the fact that there are existing text files for fine-tuning in German, a custom text file has been compiled using the same text generation method as described in Sect. 3.2.

In addition, a custom character mapping file called "unicharset" must be provided in order to map the *Schematismus* unique symbols to special Unicode characters. The built-in function has been used to create 50,000 images in total. Additionally, to generating individual images containing a single text block and saving them as a ".tif" file, two additional files are generated. One of them is the underlying ground truth, which is saved as a text file. As for the second file, it contains information about every character rendered within the image. It provides details about the character as well as coordinates describing its bounding box. The three files are then combined into a single ".lstmf" file, which is essential for the training process once fine-tuning begins. A note should be made regarding the fact that the German OCR model has been used as a starting point for this training process. Section 4 presents an evaluation of the fine-tuned Tesseract model, as well as additional experiments and preprocessing steps required to obtain the best results.

4 Evaluation

We start the evaluation of our model performance with an explanation of our parameter choices for the layout detection model, discussed in subsection Layout detection. Then, both the Tesseract optical character recognition algorithm and the layout detection model are evaluated in detail. As these two elements perform quite distinctly, evaluation will first take place separately, followed by an evaluation of the combined results in subsection 4.3. Furthermore, any additional pre- or post-processing steps that could be performed to further improve the results will be described.

4.1 Evaluation of the layout detection model

A set of eighteen original *Schematismus* document pages from 1910 was analysed for evaluation of the layout detection model. All the original documents are completely new to the model. It is important to note that these documents were not part of the training or validation set used to fine-tune the model. An example of how predicted bounding boxes and corresponding labels appear is shown in Figure 10. This figure illustrates the model being applied to one of eighteen selected document pages. Predicted elements feature a confidence level between 0 and 1, representing the model's certainty about the accuracy of its prediction. Boxes, which feature a confidence level below 0.1, have been omitted.

Overlapping bounding boxes were sorted out, using the post-processing step described in subsubsection 3.3.3. As a result of this step, the layout detection model can now be evaluated. For the purpose of obtaining a ground-truth of bounding-boxes with corresponding labels, the eighteen selected original documents are hand-annotated. During the drawing of the bounding boxes, extensive attention has been paid to detail, in order to reproduce a very accurate ground truth.

To measure the predicted bounding box accuracy, the intersection over union (IoU) method is again employed. For each document image in the testing-set, the bounding boxes in the ground-truth set were iterated over and the best matching prediction based on the IoU score was selected. Following the establishment of a list of all best matching predictions, the average over all IoU scores is calculated, which represents the bounding box prediction accuracy for the given page.

It should be noted that in order to distinguish between binary values in a multiclass classification problem, the metrics have to be calculated for each class individually. This is while only considering the class currently in focus to be positive (1) and all the rest as negative (0). Figure 13 illustrates this with a confusion matrix. In order to measure the performance of the classification of each layout element, four different metrics are used: accuracy, precision, recall, F_1-score (Fig. 14).

All the metrics mentioned above have been calculated for each of the eighteen documents in the test set in accordance with the ground truth. Table 4 gives a detailed overview of each metric for both fine-tuned and non-fine-tuned models. Compared to the other documents, the documents with indeces 3 and 5 performed the worst, especially in terms of accuracy. Explanations and illustrations of why these performed so poorly are provided in Sect. 5.1. Apart from that, the results for both bounding box accuracy and classification performance appear promising. Table 5 lists the final performance metrics of the fine-tuned model, averaged across all tested pages.

Additionally, Table 6 provides a statistic about the confidence with which the faster R-CNN model has predicted bounding-box and corresponding class labels for each class. There are three approaches that can be used in order to gain insight into this behavior: The first column in the table represents the average confidence level whenever anything has been detected by the model. Column two indicates the confidence level of the model's prediction, when the associated

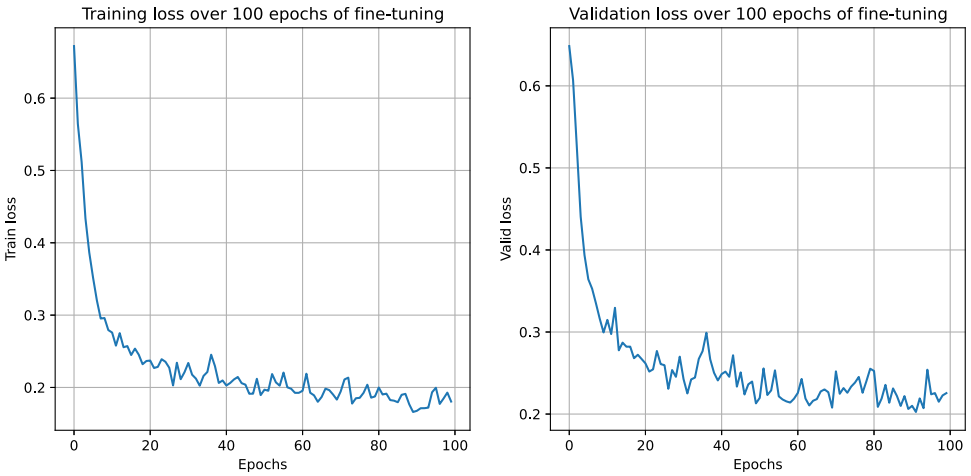


Fig. 12 This figure illustrates the training and validation losses associated with a faster R-CNN model trained on a dataset containing 39 original documents. Initial weights were derived from the weights of the existing pre-trained model

Table 4 Summary of classification accuracy, precision, recall, F_1-score, and bounding-box accuracy (based on the average IoU) for both non-fine-tuned and fine-tuned models

Document- index	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Fine-tuned																		
Accuracy	0.98	1	1	0.74	1	0.8	1	1	1	0.97	0.95	1	0.99	1	1	1	0.99	1
Precision	0.99	1	1	0.74	1	0.56	1	1	1	0.98	0.75	1	0.97	1	1	1	0.73	1
Recall	0.94	1	1	0.75	1	0.69	1	1	1	0.92	0.74	1	1	1	1	1	0.62	1
F_1	0.95	1	1	0.72	1	0.57	1	1	1	0.95	0.71	1	0.98	1	1	1	0.66	1
bbox- Accuracy	0.91	0.92	0.92	0.90	0.89	0.89	0.79	0.80	0.79	0.96	0.95	0.96	0.97	0.95	0.97	0.97	0.85	0.97
Non fine-tuned																		
Accuracy	0.81	0.89	0.67	0.81	0.93	0.50	0.97	0.99	0.97	0.83	0.78	0.66	0.59	0.69	0.52	0.54	0.88	0.57
Precision	0.51	0.75	0.38	0.49	0.65	0.50	0.50	0.50	0.50	0.47	0.47	0.64	0.71	0.65	0.55	0.66	0.41	0.63
Recall	0.54	0.71	0.32	0.51	0.53	0.38	0.49	0.49	0.49	0.44	0.43	0.45	0.56	0.55	0.50	0.44	0.45	0.51
F_1	0.52	0.71	0.35	0.49	0.56	0.42	0.49	0.50	0.49	0.44	0.45	0.52	0.62	0.59	0.48	0.53	0.43	0.55
bbox- Accuracy	0.77	0.78	0.72	0.70	0.73	0.71	0.75	0.76	0.75	0.79	0.78	0.79	0.73	0.71	0.76	0.71	0.79	0.68

Table 5 Performance measures were calculated based on the average of all 18 test documents

Performance measure	Value
Average classification accuracy	0.968
Average classification precision	0.929
Average classification recall	0.925
Average classification F_1	0.919
Average bounding-box accuracy	0.909

class was actually correct, and column three indicates the confidence level when the associated class was incorrect. Whenever no incorrect predictions have been made for a particular class, the value has been omitted. Based on the results in column three, it can be seen that the model tends to be overconfident in its predictions.

4.2 Evaluation of OCR performance

To see how Tesseract OCR performs on original *Schematismus* documents, a ground truth must be established. As this ground truth must be compiled manually by converting documents into plain text, this takes a considerable amount of time. For the evaluation process, a total of 16 original *Schematismus* pages have been manually transcribed this way. Then, in a first step, all pages were fed into a distribution of Tesseract OCR, which had not been fine-tuned to the custom font, furthermore, the pages were not preprocessed via layout detection. Tesseract was configured to use built-in page segmentation to partition the outputs of the entire pages into an easily readable and correct format. The resulting outputs did not match expected sequence. Consequently, in order to make a fair comparison between Tesseract's output and the corresponding ground-truth, the predicted outputs have been manually split and reordered to match the original

Table 6 In the Table the confidence levels of the faster R-CNN model in predicting bounding-boxes and corresponding class labels for each class are displayed

Class name	Average confidence for any prediction	Average confidence for correct prediction	Average confidence for incorrect prediction
Paragraph	0.983	0.991	0.899
BigParagraph	0.923	0.924	**0.896**
Curly	0.970	0.970	–
H1	**0.997**	**0.997**	–
H2	0.911	0.911	–
H3	0.961	0.988	0.797
H4	0.958	0.975	0.747
NameEntry	0.749	0.967	0.422

Presented are the average confidence level, the confidence level when the associated class was correct, and the confidence level when the associated class was incorrect. According to the results, the model tends to overestimate its predictions, as evidenced by the higher confidence levels associated with incorrect predictions

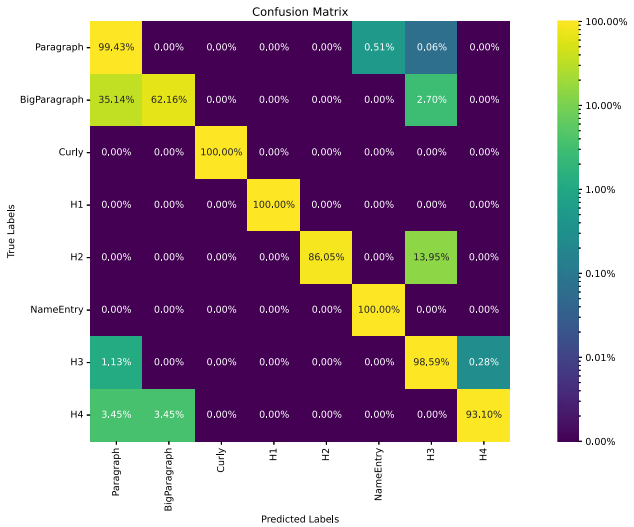


Fig. 13 This figure illustrates the confusion matrix for all predictions made on the test set. Note that the values have been normalized according to the ground truth, so that each row sums to 100 percent

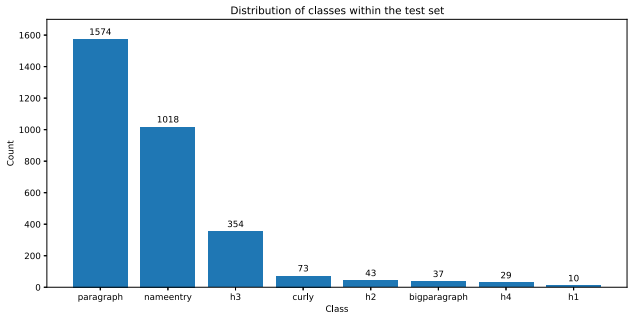


Fig. 14 Using ground truth, this figure illustrates the distribution of the various layout elements found in the test set

layout of the specific page, without altering any extracted characters or words. The constrained number of tested pages is also attributed to the time-intensive nature of this additional process. The exact same process was then repeated using a distribution of Tesseract OCR, which had been fine-tuned for the custom font. Following manual alignment, CER and WER were calculated for every block of text. The average over all pages is shown in rows one and two of Table 7. According to the results, the average CER has improved by 7.01 percentage points and the average WER by 10.94 percentage points when using the fine-tuned OCR model.

4.3 Evaluation of the OCR in combination with the layout detection model

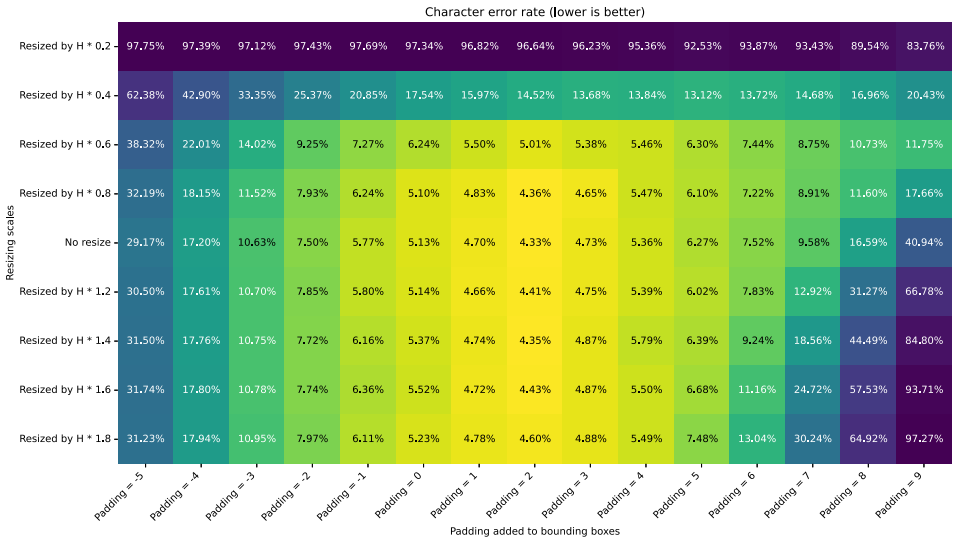
In the next step, the layout detection model was utilised to segment every structure element within the pages into individual elements. To obtain the image snippets, the original images were cropped based on the predicted bounding boxes for each page. As each element's coordinates are known in the original document, it was possible to sort them in the appropriate order, so no manual reordering was necessary. The Tesseract OCR model was then applied to each image snippet individually. Since it became evident in the prior step that the fine-tuned version performs significantly better, only this version was used. For each individual image snippet, CER and WER were calculated with the corresponding ground truth. Row three in Table 7, the averages of all predictions across all pages are presented.

According to these results, both average CER and WER have improved by another 7.87 and 6.38 percentage points respectively compared to the averages computed based on feeding the full pages into the fine-tuned Tesseract OCR algorithm. Even though we consider these improvements satisfying, we believe that OCR accuracy can be further improved. Although the layout detection model performs

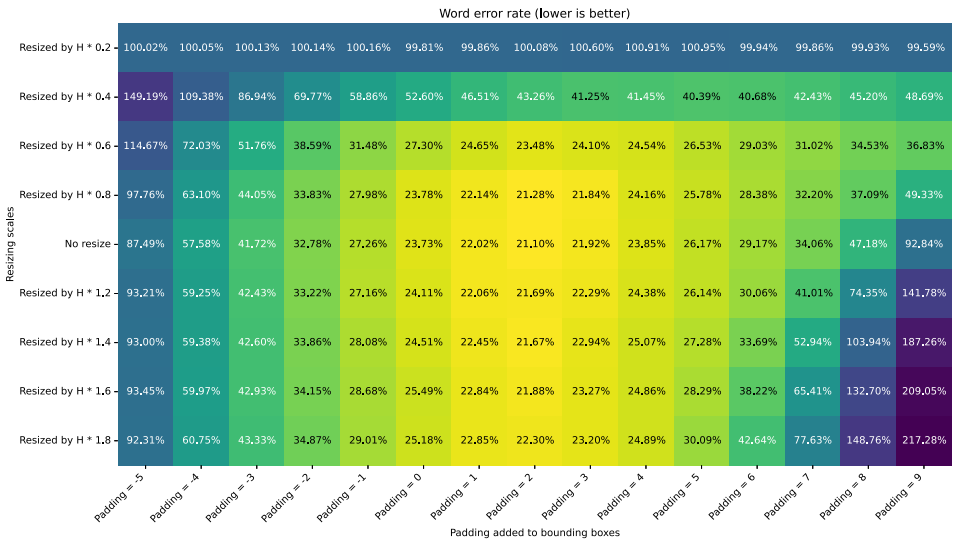
Table 7 The table presents the average CER and WER resulting from various scenarios

	Avg. CER (%)	Avg. WER (%)
Full page-without font fine-tuning	20.01	41.05
Full page-with font fine-tuning	13.00	30.11
Utilizing layout detection	5.13	23.73
Final score	4.33	21.10

The first two rows show the impact of fine-tuning Tesseract to a custom font on CER and WER on a full page. The third row presents the results when layout detection is combined with a fine-tuned Tesseract OCR model. Finally, in the last row, the results after a padding of two pixels and no resizing are applied to each snipped image, are shown



(a) average character-error-rate

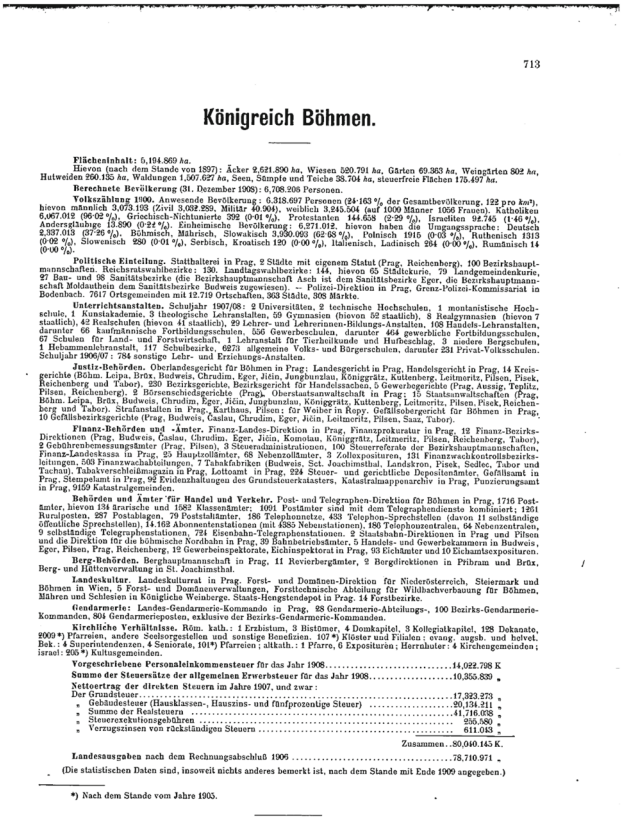


(b) average word-error-rate

Fig. 15 The figure shows the average character-error-rate and word-error-rate calculated with different levels of upscaling/downscaling and padding applied to individual image snippets

well in finding very accurate bounding boxes, sometimes characters are cropped off at the borders of the images. As a result, padding has been added around the predicted bounding box, to address this issue. As described in the official Tesseract guide on improving OCR accuracy [50], Tesseract generally works better with higher resolution images. Therefore, the individual image snippets have been resized using various scales based on the height of each image in order to find the sweet spot. The aspect ratio of the original image was maintained while upscaling (or downscaling) through linear interpolation. Figure 15 illustrate this. Interestingly, the CER and WER reach their minimum values when the original image snippet is left unscaled and padded with two pixels. This observation suggests that the resolution of the scanned pages may be sufficiently high. Row four in Table 7 shows the final average CER and WER values.

According to these results, it is possible to answer the research question, which is how much OCR accuracy can be improved by using a layout detection model as a preprocessing step to segment *Schematismus*-state documents, and feeding Tesseract individual images containing one layout structure rather than a full page. In comparison to the average CER and WER obtained on a full page using a fine-tuned Tesseract OCR model, an 8.67 percentage point improvement in the CER and 9.01 percentage point improvement in the WER were observed. In comparison with an out-of-the-box Tesseract OCR model, even higher CER and WER improvements were achieved. Here the average CER improved by a total of 15.68 percentage points and the average WER improved by a total of 19.95 percentage points.



713

Königreich Böhmen.

Flächeninhalt: 5,194.869 *ha*.
Hievon (nach dem Stande von 1897): Äcker 2,621.890 *ha*, Wiesen 529.791 *ha*, Gärten 69.363 *ha*, Weingärten 802 *ha*, Hutweiden 250.135 *ha*, Waldungen 1,507.627 *ha*, Seen, Sümpfe und Teiche 38.704 *ha*, steuerfreie Flächen 175.497 *ha*.

Berechnete Bevölkerung (31. Dezember 1908): 6,708.206 Personen.

Volkszählung 1900. Anwesende Bevölkerung: 6,318.697 Personen (24·163 % der Gesamtbevölkerung, 122 pro *km*²), hievon männlich 3,073.193 (Zivil 3,032.289, Militär 40.904), weiblich 3,245.504 (auf 1000 Männer 1056 Frauen). Katholiken 6,067.019 (96·02 %), Griechisch-Nichtunierte 292 (0·01 %), Protestanten 144.658 (2·29 %), Israeliten 92.745 (1·46 %), Andersgläubige 13.890 (0·22 %). Einheimische Bevölkerung: 6,371.012, hievon haben die Umgangssprache: Deutsch 2,337.013 (37·26 %), Böhmisch, Mährisch, Slowakisch 3,950.093 (62·68 %), Polnisch 1915 (0·03 %), Ruthenisch 1313 (0·02 %), Slowenisch 280 (0·01 %), Serbisch, Kroatisch 120 (0·00 %), Italienisch, Ladinisch 284 (0·00 %), Rumänisch 14 (0·00 %).

Politische Einteilung. Statthalterei in Prag, 2 Städte mit eigenem Statut (Prag, Reichenberg). 100 Bezirkshauptmannschaften. Reichsratswahlbezirke: 130. Landtagswahlbezirke: 144, hievon 65 Städtekurie, 79 Landgemeindenkurie, 27 Bau- und 98 Sanitätsbezirke (die Bezirkshauptmannschaft Asch ist dem Sanitätsbezirke Eger, die Bezirkshauptmannschaft Moldauthein dem Sanitätsbezirke Budweis zugewiesen). — Polizei-Direktion in Prag, Grenz-Polizei-Kommissariat in Bodenbach. 7617 Ortsgemeinden mit 12.719 Ortschaften, 363 Städte, 308 Märkte.

Unterrichtsanstalten. Schuljahr 1907/08: 2 Universitäten, 2 technische Hochschulen, 1 montanistische Hochschule, 1 Kunstakademie, 3 theologische Lehranstalten, 59 Gymnasien (hievon 52 staatlich), 8 Realgymnasien (hievon 7 staatlich), 42 Realschulen (hievon 41 staatlich), 29 Lehrer- und Lehrerinnen-Bildungs-Anstalten, 108 Handels-Lehranstalten, darunter 66 kaufmännische Fortbildungsschulen, 556 Gewerbeschulen, darunter 464 gewerbliche Fortbildungsschulen, 67 Schulen für Land- und Forstwirtschaft, 1 Lehranstalt für Tierheilkunde und Hufbeschlag, 3 niedere Bergschulen, 1 Hebammenlehranstalt, 117 Schulbezirke, 6273 allgemeine Volks- und Bürgerschulen, darunter 231 Privat-Volksschulen. Schuljahr 1906/07: 784 sonstige Lehr- und Erziehungs-Anstalten.

Justiz-Behörden. Oberlandesgericht für Böhmen in Prag; Landesgericht in Prag, Handelsgericht in Prag, 14 Kreisgerichte (Böhm. Leipa, Brüx, Budweis, Chrudim, Eger, Jičin, Jungbunzlau, Königgrätz, Kuttenberg, Leitmeritz, Pilsen, Pisek, Reichenberg und Tabor), 230 Bezirksgerichte, Bezirksgericht für Handelssachen, 5 Gewerbegerichte (Prag, Aussig, Teplitz, Pilsen, Reichenberg), 2 Börsenschiedsgerichte (Prag). Oberstaatsanwaltschaft in Prag; 15 Staatsanwaltschaften (Prag, Böhm. Leipa, Brüx, Budweis, Chrudim, Eger, Jičin, Jungbunzlau, Königgrätz, Kuttenberg, Leitmeritz, Pilsen, Pisek, Reichenberg und Tabor). Strafanstalten in Prag, Karlhaus, Pilsen; für Weiber in Ropy. Gefällsobergericht für Böhmen in Prag, 10 Gefällsbezirksgerichte (Prag, Budweis, Caslau, Chrudim, Eger, Jičin, Leitmeritz, Pilsen, Saaz, Tabor).

Finanz-Behörden und -Ämter. Finanz-Landes-Direktion in Prag, Finanzprokuratur in Prag, 12 Finanz-Bezirks-Direktionen (Prag, Budweis, Caslau, Chrudim, Eger, Jičin, Kuttenberg, Königgrätz, Leitmeritz, Pilsen, Reichenberg, Tabor), 2 Gebührenbemessungsämter (Prag, Pilsen), 3 Steueradministrationen, 100 Steuerreferate der Bezirkshauptmannschaften, Finanz-Landeskassa in Prag, 25 Hauptzollämter, 68 Nebenzollämter, 3 Zollexposituren, 131 Finanzwachkontrollsbezirksleitungen, 503 Finanzwachkontrollsabteilungen, 7 Tabakfabriken (Budweis, Sct. Joachimsthal, Landskron, Pisek, Sedlec, Tabor und Tuchau), Tabakverschleißmagazin in Prag, Lottoamt in Prag, 224 Steuer- und gerichtliche Depositenämter, Gefällsamt in Prag, Stempelamt in Prag, 92 Evidenzhaltungen des Grundsteuerkatasters, Katastralmappenarchiv in Prag, Puozierungsamt in Prag, 9159 Katastralgemeinden.

Behörden und Ämter für Handel und Verkehr. Post- und Telegraphen-Direktion für Böhmen in Prag, 1716 Postämter, hievon 134 ärarische und 1582 Klassenämter, 1691 Postämter sind mit dem Telegraphendienste kombiniert; 1951 Ruralposten, 287 Postablagen, 79 Poststallämter, 186 Telephonnetze, 433 Telephon-Sprechstellen (davon 11 selbständige öffentliche Sprechstellen), 14.162 Abonnentenstationen (mit 4383 Nebenstationen), 180 Telephoncentralen, 64 Nebenzentralen, 9 selbständige Telegraphenstationen, 724 Eisenbahn-Telegraphenstationen. 2 Staatsbahn-Direktionen in Prag und Pilsen und die Direktion für die böhmische Nordbahn in Prag, 39 Bahnbetriebsämter, 5 Handels- und Gewerbekammern in Budweis, Eger, Pilsen, Prag, Reichenberg, 19 Gewerbeinspektorate, Eichinspektorat in Prag, 93 Eichämter und 10 Eichamtsexposituren.

Berg-Behörden. Berghauptmannschaft in Prag. 11 Revierbergämter, 9 Borgdirektionen in Přibram und Brüx, Berg- und Hüttenverwaltung in St. Joachimsthal.

Landeskultur. Landeskulturrat in Prag. Forst- und Domänen-Direktion für Niederösterreich, Steiermark und Böhmen in Wien, 5 Forst- und Domänenverwaltungen, Forsttechnische Abteilung für Wildbachverbauung für Böhmen, Mähren und Schlesien in Königliche Weinberge. Staats-Hengstendepot in Prag. 14 Forstbezirke.

Gendarmerie: Landes-Gendarmerie-Kommando in Prag, 28 Gendarmerie-Abteilungs-, 100 Bezirks-Gendarmerie-Kommanden, 804 Gendarmerieposten, exklusive der Bezirks-Gendarmerie-Kommanden.

Kirchliche Verhältnisse. Röm. kath.: 1 Erzbistum, 3 Bistümer, 4 Domkapitel, 3 Kollegiatkapitel, 128 Dekanate, 2009 *) Pfarreien, andere Seelsorgestellen und sonstige Benefizien. 107 *) Klöster und Filialen: evang. augsb. und helvet. Bek.: 4 Superintendenten, 4 Seniorate, 101 *) Pfarreien; altkath.: 1 Pfarre, 6 Exposituren; Herrnhuter: 4 Kirchengemeinden; israel: 205 *) Kultusgemeinden.

Vorgeschriebene Personaleinkommensteuer für das Jahr 1908 ..... 14,022.708 K
Summe der Steuersätze der allgemeinen Erwerbsteuer für das Jahr 1908 ..... 10,355.839 „
Nettoertrag der direkten Steuern im Jahre 1907, und zwar:
Der Grundsteuer ..... 17,323.273 „
„ Gebäudesteuer (Hausklassen-, Hauszins- und fünfprozentige Steuer) ..... 20,134.211 „
„ Summe der Realsteuern ..... 41,716.038 „
„ Steuerexekutionsgebühren ..... 255.580 „
„ Verzugszinsen von rückständigen Steuern ..... 611.043 „
Zusammen ..... 80,040.145 K.
Landesausgaben nach dem Rechnungsabschluß 1905 ..... 78,710.971 „
(Die statistischen Daten sind, insoweit nichts anderes bemerkt ist, nach dem Stande mit Ende 1909 angegeben.)

*) Nach dem Stande vom Jahre 1903.

Fig. 16 This figure illustrates a document with index 5 of the evaluation set, which had the poorest layout-detection performance

5 Discussion

It is evident from the results presented in section 4 that the use of a custom-developed layout detection model to segment *Schematismus*-style documents together with a Tesseract model fine-tuned to a custom font designed to be as close to the original as possible significantly improved the quality of the extracted text. However, it should be noted that, due to the relatively small sample size of 16 pages, these results might not represent the full picture. The gain that combined layout detection and text extraction provides may be larger than metrics alone can express. That is due to two different reasons.

As each layout element is segmented by bounding boxes, it is known where the blocks are located within a document coordinate-wise. It allows the reordering of the extracted texts so that they correspond to the reading flow of the document. This is essential when extracting text from a document with columns.

The second reason is that due to the classification of the individual bounding boxes it is possible to immediately tell the class of a structure element. Thus, it is possible, for example, to extract only headings and paragraphs from a document. Moreover, this makes it easier to match paragraphs with individual headings, or to get all enclosed structure elements within curly brackets.

As for the Tesseract OCR model, the custom designed font should be improved, to make text extraction even better. Due to the fact that the current version of the font does not include certain characters such as “č” or “ň” in its unichar set, detection of these kinds of letters is not possible, resulting in errors.

The pipeline we outlined in this article is highly adaptable and could be optimised for similar tasks with relatively little effort. Particularly the process we suggest to produce synthetic training data at scale contributes largely to our capacity to adapt the layout detection model quickly. Not only does it allow to produce a significant amount of training data in relatively little time, but the use of synthetic training data also appears to significantly increase the precision of the bounding box prediction.

5.1 Error analysis

According to Table 4, documents with the indices 3 and 5 performed quite poorly compared to others in terms of accuracy. Both documents are visually very similar. As an example, the document with index 5 is shown in Fig. 16. Clearly, this document differs from the typical three-column *Schematismus*-style document. Aside from the general layout, a key difference is the indentation of each paragraph. Although these aspects were considered during the generation of synthetic documents, resulting in a separate class "BigParagraphs", the generated structures do not appear to be as similar to the originals as intended, based on layout detection results. Due to the relatively small number of examples of this type of document in the training set used for fine-tuning, we could not observe any improvement. Specifically, only two pages containing "BigParagraphs" were included in the fine-tuning training set, which appears to be too few for the model to effectively learn this class. Therefore, to improve performance on these types of *Schematismus* documents, more pages similar to those in Fig. 16 must be manually annotated and added to the fine-tuning training set.

5.2 Research questions

With regard to our first research question, we could show how OCR accuracy can be significantly improved by splitting individual document pages into their layout elements as a preprocessing step.

As for the second research question, it has been shown that fine-tuning Tesseract with a custom font results in performance improvement.

In comparison with the performance of an out-of-the-box Tesseract Model for OCR on an entire page of the *Schematismus*, the results indicated that segmenting and splitting individual document pages into their layout elements with a deep learning convolutional neural network resulted in significantly better OCR accuracy.

5.3 Outlook

The procedure we developed therefore represents a crucial step toward a significantly improved analysis of printed historical documents produced in the larger context of the long 19th century, particularly as we show how each of the two steps can be further adapted to the specific needs, requirements and challenges met by fellow researchers.

However, we expect that both, layout detection and optical character recognition, can be further optimised for even better performance on historical documents. Layout detection may benefit from increasing the training dataset, not only in terms of the number of pages but also by including a wider variety of documents. In other words, by generating documents that are visually similar to much older, in our case *Schematismus*-style documents produced in the first half of the 19th century, when a different layout was used, and including these in the training set, a more generalized and robust model may be achieved. Further, domain knowledge can be put to use to enhance text extraction. Considering that most of the printed text in *Schematismus*-style documents consists of abbreviations that are listed and described on specific pages within these documents, this information can be utilised to build a custom spell-checking algorithm to correct errors in the text extraction process. Additionally, it would be of interest to explore whether the methods used in this paper can be applied to other types of historical documents.

Our approach offers a viable solution to a number of common problems in dealing with retro-digitised historical texts in historical and humanities research contexts, yet also in industrial application. In this work, it was first shown that the breakdown of the OCR problem, and its solution in several sub-steps, is very promising. However, careful work and precise adjustment of the training data are necessary preconditions to obtain excellent performance. Future work based on our approach will tackle more diverse layouts and broader scope of documents types.

Funding Open access funding provided by Graz University of Technology.

References

1. Liu, X., Gao, F., Zhang, Q., Zhao, H.: (2019), pp. 32–39
2. Noflatscher, H.: (Böhlau, 2004), pp. 59–75
3. Bauer, V.: *Repertorium territorialer Amtskalender und Amtshandbücher im Alten Reich: Adreß- , Hof- , Staatskalender und Staatshandbücher des 18. Jahrhunderts.*, vol. vol. 2 (Klostermann, 1999)
4. ALEX. Staatshandbuch 1910: Schematismus staat (1910). https://alex.onb.ac.at/cgi-content/alex?aid=shb&datum=1910&page=597&size=45. Accessed on: 2022-10-21
5. Raphael, L.: *Die Erben von Bloch und Febvre. Annales-Geschichtsschreibung und nouvelle histoire in Frankreich 1945-1980* (Klett-Cotta, Stuttgart, 1994)
6. Jannidis, F., Kohle, H., Rehbein, M.: Digital Humanities Eine Einführung. J. B. Metzler, Stuttgart (2017). https://doi.org/10.1007/978-3-476-05446-3
7. Boros, E., Nguyen, N.K., Lejeune, G., Doucet, A.: Assessing the impact of ocr noise on multilingual event detection over digi-

tised documents. Int. J. Digit. Libr. (2022). https://doi.org/10.1007/s00799-022-00325-2
8. Wajer, M.B.W.: Internet Archive OCR Stack in 2021. Switching to Open Source Software (2021). https://ia601807.us.archive.org/35/items/merlijn-wajer-presentation/merlijn-wajer-presentation-ocr.pdf
9. Austrian National Library. Austrian Books Online (2023). https://www.onb.ac.at/en/digital-offers/austrian-books-online
10. Kettunen, K., Koistinen, M., Kervinen, J.: Ground truth ocr sample data of finnish historical newspapers and journals in data improvement validation of a re-ocring process. LIBER Quarterly **30**(1), 1–20 (2020). https://doi.org/10.18352/lq.10322
11. Staatsbibliothek, B.: Technologien & Softwareentwicklung (2023). https://www.digitale-sammlungen.de/de/technologien-und-softwareentwicklung
12. G. Markus. Issue 13: OCR. EuropeanaTech Insight is a multimedia publication about R&D developments by the EuropeanaTech Community (2019). https://pro.europeana.eu/page/issue-13-ocr
13. Ehmer, J., Mitterauer, M., Thaller, M.: Wiener Datenbank zur Europäischen Familiengeschichte (2023)
14. Becker, P., Osterkamp, J.: The Emperor's Desk (2018–2021)
15. Romberg, M.: The Viennese Court (2020–2023)
16. Popovici, V., Velková, A.: Social Mobility of Elites (2022)
17. Engl, E.: OCR-D kompakt: Ergebnisse und Stand der Forschung in der Förderinitiative. Bibliothek Forschung und Praxis **44**(2), 218–230 (2020). https://doi.org/10.1515/bfp-2020-0024
18. Martínek, J., Lenc, L., Král, P.: Building an efficient OCR system for historical documents with little training data. Neural Comput. Appl. **32**(23), 17209–17227 (2020). https://doi.org/10.1007/s00521-020-04910-x
19. Reul, C., Christ, D., Hartelt, A., Balbach, N., Wehner, M., Springmann, U., Wick, C., Grundig, C., Büttner, A., Puppe, F.: Ocr4all–an open-source tool providing a (semi-)automatic ocr workflow for historical printings (2019). https://doi.org/10.3390/app9224853. http://arxiv.org/abs/1909.04032
20. Cordell, R.: Machine Learning + Libraries. A Report on the State of the Field (2020). https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf?loclr=blogsig
21. Gasparini, A., Kautonen, H.: Understanding artificial intelligence in research libraries-extensive literature review. LIBER Quart. J. Assoc. Eur. Res. Libr. (2022). https://doi.org/10.53377/lq.10934
22. Teibenbacher, P., Kramer, D., Göderle, W.: An Inventory of Austrian Census Materials , 1857-1910. Final Report. Mosaic Working Paper **190**, 25 (2012)
23. Zechner, A., Knapp, E., Adelsberger, M.: Prices and Wages in Salzburg and Vienna, c. 1450–1850 An Introduction to the Data. Vierteljahresschrift fur Sozial und Wirtschaftsgeschichte **108**(4), 263–270 (2021). https://doi.org/10.25162/VSWG-2021-0016
24. Bavouzet, J.: in *The Habsburg Civli Service and Beyond: Bureaucracy and Civil Servants from the Vormärz to the Inter-War Years*, ed. by F. Adlgasser, F. Lindström (Verlag der Österreichischen Akademie der Wissenschaften, Vienna, 2019), pp. 167–186. https://doi.org/10.2307/j.ctvggx26b.11
25. Wang, J., Liu, C., Jin, L., Tang, G., Zhang, J., Zhang, S., Wang, Q., Wu, Y., Cai, M.: Towards robust visual information extraction in real world: New dataset and novel solution (2021). www.aaai.org
26. Douzon, T., Duffner, S., Garcia, C., Espinas, J.: Improving information extraction on business documents with specific pre-training tasks. In International Workshop on Document Analysis Systems. (Springer Science and Business Media Deutschland GmbH, 2022), pp. 111–125. https://doi.org/10.1007/978-3-031-06555-2_8
27. Carbonell, M., Fornés, A., Villegas, M., Lladós, J.: A neural model for text localization, transcription and named entity recognition in full pages. Pattern Recogn. Lett. **136**, 219–227 (2020). https://doi.org/10.1016/j.patrec.2020.05.001
28. Tarride, S., Maarand, M., Boillet, M., McGrath, J., Capel, E., V'ezina, H., Kermorvant, C.: Large-scale genealogical information extraction from handwritten Quebec parish records. Int. J. Doc. Anal. Recogn. (IJDAR) **26**(3), 255–272 (2023). https://doi.org/10.1007/s10032-023-00427-w
29. Monnier, T., Aubry, M.: in *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 91–96. (IEEE, 2020)
30. Gruber, I., Ircing, P., Neduchal,P., Hrúz, M., Hlaváč, M., Zajíc, Z., Švec, J., Bulín, M.: in *International Conference on Speech and Computer*, pp. 166–175, (Springer, 2020)
31. M. Shen, H. Lei, in *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pp. 1566–1570, (IEEE, 2015)
32. Lat, A., Jawahar, C.: in *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 3162–3167 (IEEE, 2018)
33. Grüning, T., Leifert, G., Strauß, T., Michael, J., Labahn, R.: A two-stage method for text line detection in historical documents. Int. J. Doc. Anal. Recogn. **22**(3), 285–302 (2019). https://doi.org/10.1007/s10032-019-00332-1. arXiv:1802.03345
34. Büttner, J., Martinetz, J., El-Hajj, H., Valleriani, M.: Cordeep and the sacrobosco dataset: detection of visual elements in historical documents. J. Imaging (2022). https://doi.org/10.3390/jimaging8100285
35. Binmakhashen, G.M., Mahmoud, S.A.: Document layout analysis: a comprehensive survey. ACM Comput. Surv. (2019). https://doi.org/10.1145/3355610
36. Boillet, M., Kermorvant, C., Paquet, T.: Multiple document datasets pre-training improves text line detection with deep neural networks. Proceedings-International Conference on Pattern Recognition pp. 2134–2141 (2020). https://doi.org/10.1109/ICPR48806.2021.9412447. arXiv:2012.14163
37. Xu, Y., Li, M., Cui, L., Huang,S., Wei, F., Zhou,M.: Layoutlm: Pre-training of text and layout for document image understanding. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1192–1200 (2020). https://doi.org/10.1145/3394486.3403172. http://arxiv.org/abs/1912.13318
38. Biswas, S., Riba, P., Lladós, J., Pal, U.: Beyond document object detection: instance-level segmentation of complex layouts. Int. J. Doc. Anal. Recogn. **24**(3), 269–281 (2021). https://doi.org/10.1007/s10032-021-00380-6
39. LuaTex. Luatex (2023). https://www.luatex.org/
40. Remy, P.: Name dataset. https://github.com/philipperemy/name-dataset (2021)
41. FontForge. Fontforge. https://fontforge.org/en-US/ (2023). [Accessed 09-Mar-2023]
42. Wikipedia. Liste der österreichischen Orden und Ehrenzeichen — Wikipedia, the free encyclopedia. http://de.wikipedia.org/w/index.php?title=Liste%20der%20%C3%B6sterreichischen%20Orden%20und%20Ehrenzeichen&oldid=231609032 (2023). [Online; accessed 09-March-2023]
43. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, in *Advances in Neural Information Processing Systems*, vol. 32, ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. ' Alché-Buc, E. Fox, R. Garnett (Curran Associates, Inc., 2019). https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf
44. Li, Y., Xie, S., Chen, X., Dollar, P., He, K., Girshick, R.: Benchmarking detection transfer learning with vision transformers (2021). https://doi.org/10.48550/ARXIV.2111.11429
45. Keskar, N.S., Mudigere,D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P.: On large-batch training for deep learning: Generalization

gap and sharp minima. CoRR **abs/1609.04836**. arXiv:1609.04836. (2016)
46. Krizhevsky, A.: One weird trick for parallelizing convolutional neural networks (2014). https://doi.org/10.48550/ARXIV.1404.5997
47. PyTorch. Torchserve (2023). https://pytorch.org/serve/index.html
48. Fleischhacker, M.: Boundingboxeditor. https://github.com/mfl28/BoundingBoxEditor (2023)
49. Tesseract. Tesseract (2023). https://github.com/tesseract-ocr/tesseract
50. Tesseract. Improving the quality of the output (2022). https://tesseract-ocr.github.io/tessdoc/ImproveQuality.html. Accessed on: 2022-10-21