https://doi.org/10.2352/issn.2168-3204.2022.19.1.04

Handwritten and Printed Text Identification in Historical Archival Documents

Mahsa Vafaie, FIZ Karlsruhe and Karlsruhe Institute of Technology; Karlsruhe, Germany
Oleksandra Bruns, FIZ Karlsruhe and Karlsruhe Institute of Technology; Karlsruhe, Germany
Nastasja Pilz, Landesarchiv Baden-Württemberg; Stuttgart, Germany
Jörg Waitelonis, FIZ Karlsruhe and Karlsruhe Institute of Technology; Karlsruhe, Germany
Harald Sack, FIZ Karlsruhe and Karlsruhe Institute of Technology; Karlsruhe, Germany

Abstract

Historical archival records present many challenges for OCR systems to correctly encode their content, due to visual complexity, e.g. mixed printed text and handwritten annotations, paper degradation and faded ink. This paper addresses the problem of automatic identification and separation of handwritten and printed text in historical archival documents, including the creation of an artificial pixel-level annotated dataset and the presentation of a new FCN-based model trained on historical data. Initial test results indicate 18% IoU performance improvement on recognition of printed pixels and 10% IoU performance improvement on recognition of handwritten pixels in synthesised data when compared to the state-of-the-art trained on modern documents. Furthermore, an extrinsic OCR-based evaluation on the printed layer extracted from real historical documents shows 26% performance increase.

Motivation and Problem Statement

During the last years, proper archival management has received much attention from historians and historical research communities. The two world wars, numerous revolutions and conflicts in the 20th century, as well as the ageing of documents have caused a huge loss of archival information, and thus, negatively affected the possibility to reveal knowledge about certain historical events. One of the ways to protect cultural heritage is through digitisation and transcription of archival materials, and thereby, the development of an open information society where archival content is easily retrievable, accessible and publicly available.

Accessing knowledge hidden in archival records is also one of the initial steps and a fundamental building block of the "Pilotprojekt zur Transformation der Wiedergutmachung"[1] [Pilot Project for Transformation of Reparations] (referred to as WGM in the rest of this paper). This project is centred around archival documents from the reparation process and reparation cases filed after the year 1945, in the state of Baden-Württemberg, Germany. The records contain information about one of the darkest passages in German history — first-hand knowledge about crimes towards people, persecuted and discriminated on account of their ethnicity, religion, political belief and sexual orientation during the rule of the National Socialist Regime. The content of these documents is of paramount importance both for scientific research and for the general public, including the descendants of the victims and survivors. Making this information publicly accessible and retrievable can provide historians with a great resource to explore the repressive apparatus and the interrelated Nazi policies of exclusion and destruction. Reconstructing lives of the victims of National Socialist injustice through biographical information hidden in the archival documents is also of utmost importance to the relatives of the victims, as well as the civil society.

Automatic processing of documents, in particular, the provision of transcripts of millions of scanned documents via optical character recognition (OCR) systems, speeds up and simplifies the process of information retrieval immensely. Due to the different visual appearance and structural properties of machine-printed and handwritten text, most OCR systems are specialised in recognition of either printed or handwritten text [1]. Moreover, archival records — various manuscript materials, journals with handwritten notes in the margins, forms and tables filled out by hand — consist of large amounts of mixed text, where areas of handwritten and machine-printed text are very close to each other or even overlapping (see Fig. 1). Therefore, in order to increase OCR quality, printed and handwritten parts in scanned documents have to be identified and separated before using the respective recognition system, i.e., printed or handwritten OCR [2].

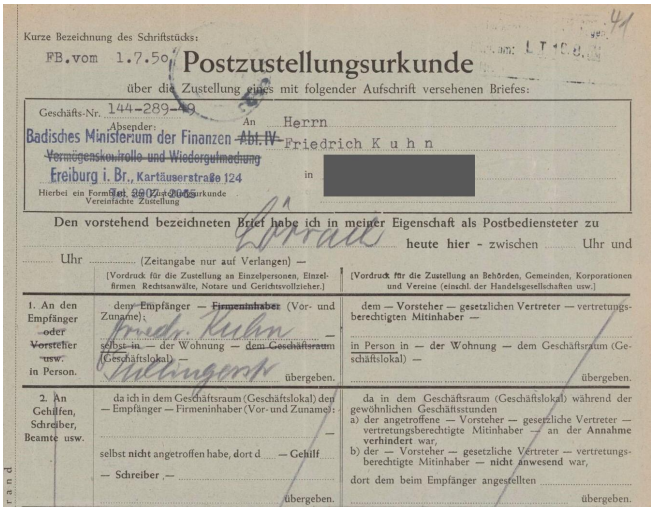
Kurze Bezeichnung des Schriftstücks:
FB. vom 1.7.50
Postzustellungsurkunde
über die Zustellung eines mit folgender Aufschrift versehenen Briefes:
Geschäfts-Nr. 144-289-49
Absender: Badisches Ministerium der Finanzen Abt. IV Vermögenskontrolle und Wiedergutmachung Freiburg i. Br., Kartäuserstraße 124
An Herrn Friedrich K u h n in [redacted]
Hierbei ein Formblatt zur Postzustellungsurkunde Vereinfachte Zustellung
Den vorstehend bezeichneten Brief habe ich in meiner Eigenschaft als Postbediensteter zu heute hier - zwischen ... Uhr und ... Uhr ... (Zeitangabe nur auf Verlangen) —
[Vordruck für die Zustellung an Einzelpersonen, Einzelfirmen Rechtsanwälte, Notare und Gerichtsvollzieher.]
[Vordruck für die Zustellung an Behörden, Gemeinden, Korporationen und Vereine (einschl. der Handelsgesellschaften usw.)]
1. An den Empfänger — oder Vorsteher usw. in Person.
dem Empfänger — Firmeninhaber (Vor- und Zuname): Friedr. Kuhn — selbst in — der Wohnung — dem Geschäftsraum (Geschäftslokal) — übergeben.
dem — Vorsteher — gesetzlichen Vertreter — vertretungsberechtigten Mitinhaber — in Person in — der Wohnung — dem Geschäftsraum (Geschäftslokal) — übergeben.
2. An Gehilfen, Schreiber, Beamte usw.
da ich in dem Geschäftsraum (Geschäftslokal) des — Empfänger — Firmeninhaber (Vor- und Zuname): — selbst nicht angetroffen habe, dort d — Gehilf — Schreiber — übergeben.
da in dem Geschäftsraum (Geschäftslokal) während der gewöhnlichen Geschäftsstunden a) der angetroffene — Vorsteher — gesetzliche Vertreter — vertretungsberechtigte Mitinhaber — an der Annahme verhindert war, b) der — Vorsteher — gesetzliche Vertreter — vertretungsberechtigte Mitinhaber — nicht anwesend war, dort dem beim Empfänger angestellten ... übergeben.

Figure 1. *A sample from "Wiedergutmachung" scanned documents, with overlapping machine printed and handwritten text*

Research has been conducted on separation of handwritten and printed text, but the conventional methods in this area have some significant limitations (e.g., [2-4]). Most of them tend to be ineffective, showing low accuracy in cases where machine-printed and handwritten parts are overlapping. Another challenge to the already existing methods and models is the special characteristics

[1] https://www.fiz-karlsruhe.de/en/forschung/wiedergutmachung

of the archival documents compared to modern documents. To the knowledge of the authors, there is no available dataset which contains historical archival records and their corresponding labels for text type identification.

An additional performance factor arises from the media representation of the records. For example, the archival records of the WGM have been digitised by scanning of either physical records (see Fig. 1) or microfilms (see Fig. 2). When microfilming, original records are scaled down and decolourized, which causes information loss and may influence the performance of the text separation system. Adding training data from different media types might result in an improvement of text type identification in historical records represented on different media types.

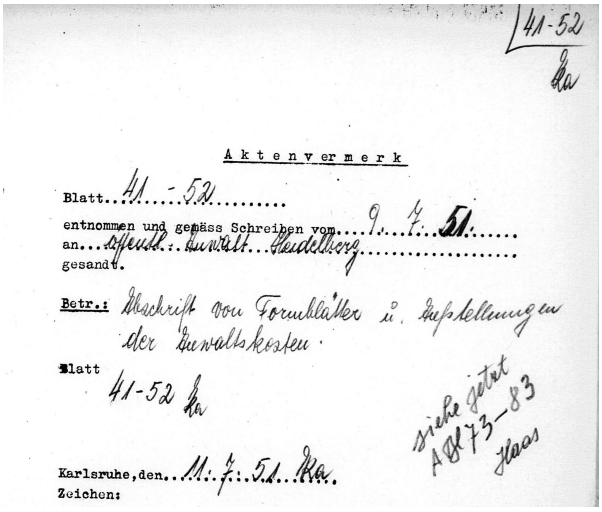

Figure 2. *A sample from "Wiedergutmachung" microfilm documents, with overlapping machine printed and handwritten text*

In this paper, the WGM-SYN, a new artificial pixel-level annotated dataset created from historical archival documents is presented. Furthermore, the WGM-MOD, a new model for identification of printed and handwritten text trained on WGM-SYN dataset is introduced. This model is tested and compared with the state-of-the-art, in two different evaluation scenarios. Both the WGM-SYN dataset and the WGM-MOD trained on this data are made publicly available and can be found on GitHub.[2]

The remainder of this paper is organised as follows: The Related Work section presents an overview of existing literature on separation of machine-printed and handwritten text and points out the gaps in this research area. Section Text Type Identification and Separation in Archival Documents describes the main contributions of the paper in more detail. First, the procedure for creating the WGM-SYN dataset is explained and then, the model architecture and experiment settings are described. In the Results section, the two evaluation scenarios and the achieved results are presented and discussed. Finally, Conclusion and Future Work summarises the achievements of this paper and highlights the next steps.

Related Work

Traditional approaches for text type identification and separation include algorithms based on Hidden Markov Models (HMMs) and Support Vector Machines (SVMs) that differ mostly on the level of separation. For example, in [4] the separation is conducted on the line level, and in [2] and [3] on the word level. These approaches are predicated on the visual impressions and structural features of text components. However, they do not cover cases where printed and handwritten texts overlap.

Recent solutions for the text segmentation task at the level of pixels imply methods based on deep neural networks [5-9]. Providing pixel-level labels, these methods allow separation of printed and handwritten components, even in case of overlappings. In [8], the first pixel-level separation method based on end-to-end learning of a CNN was proposed. To alleviate the class imbalance problem that occurs as the number of background pixels is much larger than the foreground, a new loss function considering class frequencies and sample difficulties was introduced. To compensate for the lack of an appropriate pixel-level annotated dataset, the authors proposed a data synthesis method that provides pixel-level annotations and overlappings of different text types. In [9] a lightweight Fully Convolutional Network (FCN-light) was introduced, which aims at solving the problem of printed and handwritten text identification in the overlapping parts of scanned documents. For training this model, modern documents that contain a mix of machine-printed and handwritten text were used. However, only 9.84% of the documents contained overlapping text. FCN-light results in 0.85 total mean Intersection over Union (IoU), when tested on modern documents with simple layouts, but underperforms when applied on images with different formats and layouts [9]. IoU is the standard performance measure that is commonly used for any image segmentation problem [10].

Training a network for the pixel-wise separation task requires data with pixel-level annotations. IAM [11] and CVL [12] are two datasets, which provide XML metadata that can be used to create pixel-level annotations for document images. Using this XML data, pixel-level annotated masks containing information on whether the text is handwritten or machine-printed can be created. Both IAM and CVL datasets contain a block of machine-printed text, followed by the handwritten version of the same text. For the machine-printed texts the same font and size were used over all pages. On the other hand, the parts generated by hand were written by a variety of writers. Even though these two datasets contain a mix of machine-printed and handwritten text, they still lack parts where the two text types are overlapping. On the other hand, the simple layout and clean background of the documents in these datasets are not representative of historical archival records.

Crowdsourcing transcription of archival materials has recently become a solution to the problem of information extraction from challenging archival materials [13-15], the quality control of automatic information extraction [16], and the need for greater public engagement with archives. However, due to data privacy issues, public access to some archival records, including the records of the WGM is restricted. Thus, transcription and information extraction in these documents cannot benefit from crowdsourcing efforts without data anonimisation.

[2] https://github.com/ISE-FIZKarlsruhe/Wiedergutmachung

Text Type Identification and Separation in Archival Documents

In this section, the main contributions of the paper are presented. First, the process of synthesising an artificial dataset using historical archival documents is described. Following the dataset description, the model architecture and experiment settings are presented.

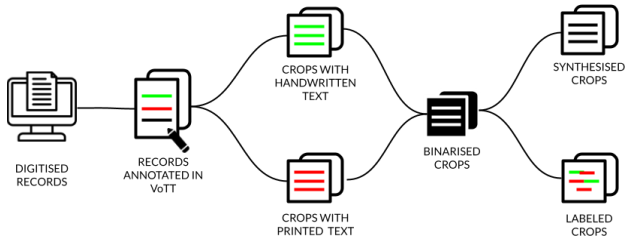

Figure 3. *The workflow of the dataset synthesis process*

WGM-SYN dataset. To solve the task of pixel-wise identification of handwritten and printed text and due to the lack of pixel-level annotated historical data, an artificial dataset, called WGM-SYN, based on the pipeline proposed by [8] is synthesised. The basis of this dataset is a subset of historical archival records from the "Pilotprojekt zur Wiedergutmachung". The subset consists of scans of 150 original archival documents in colour and 153 black-and-white microfilms, with different layouts such as forms filled out by hand, typewritten certificates, testimonies and declarations. The scanned documents are manually annotated using VOTT[3], an open-source image annotation tool. The regions containing only one type of text, i.e., printed or handwritten, are extracted, resulting in 117 images of handwritten and 320 images of printed text extracted from microfilm scans, and 202 images of handwritten and 447 images of printed text extracted from colour scans. In order to remove microfilm noise created by lighting conditions and background artefacts which could result in inaccuracies in the labels, all images are binarised using the document enhancement method proposed at [17]. After binarisation, machine-printed and handwritten images are merged together to create black and white artificial crops with mixed and overlapping printed and handwritten text. Ground truth labels for these crops are created by channel-wise concatenation of the same binarised images. The data synthesis pipeline results in 3,764 different crops of size 256×256, containing handwritten and printed text, and their corresponding pixel-level annotations with three colour-coded tags — green for handwritten, red for printed and blue for background. Figure 4 shows three samples from this synthesised dataset and their ground truth annotations.

WGM-MOD model architecture. Following the architecture of [9], the WGM-MOD model architecture is a variation of the FCN-8 architecture [5], with 294,952 parameters to learn. This model has been trained on 3,335 crops from WGM-SYN for 50 epochs. The loss function is weighted categorical cross entropy with pre-set values for the weights (w=[0.4, 0.5, 0.1] for printed, handwritten and background respectively). Since there is a variable number of samples from each class, the weighted loss function is chosen to account for the class imbalance problem.

The classification pipeline also includes a dense Conditional Random Field (CRF) in the post-processing step, at which the label assigned to a pixel at a particular location could change based on the labels of its adjacent pixels. CRFs are a type of graphical probabilistic model used for labelling sequential data with dependencies between samples [18]. There are multiple types of CRF models such as Linear CRFs, Grid CRFs, Skip-Chain CRF and Dense CRFs [19]. In dense or fully connected CRFs, every node is connected to n-1 nodes, i.e, all nodes in the image are connected to every other node. Due to the fully connected structure, this is the best possible CRF to be applied to an image segmentation process. Applying CRFs, label agreement between similar pixels is maximised by assigning pixels with similar features to the same predicted class [20].

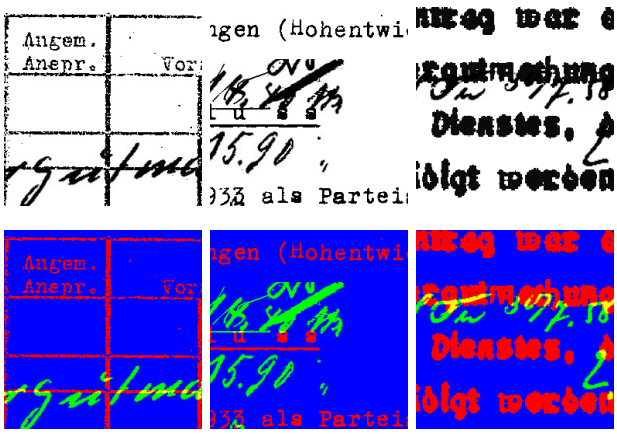

Figure 4. *The artificially synthesised image crops and their corresponding pixel-level annotations with three labels - blue for background, green for handwritten and red for printed text*

Intrinsic and Extrinsic Model Evaluation

In this section, the performance of the WGM-MOD is compared to the baseline model, FCN-light, in two different evaluation scenarios. In an intrinsic evaluation, the Intersection over Union (IoU) is used as a measure for performance evaluation, using an unseen subset of synthesised data for testing. Furthermore, a subsequent OCR task is conducted to observe the impact of text separation with the new model on OCR performance.

IoU evaluation. The performance of the two models, FCN-light and WGM-MOD are evaluated on 429 crops from WGM-SYN's test set. The results of the evaluation are shown in Table 1. The obtained results show a considerable improvement in the identification of printed and handwritten text using the model trained with WGM-SYN. Figure 5 presents the results of pixel-wise classification of sample crops from the test set using the two models, as well as the ground truth.

The relatively low IoU scores compared to the 0.85 total mean IoU reported at [9] can be attributed to the complex nature of the synthetic data and the great amount of overlappings in WGM-SYN in comparison with text overlappings in the data used in [9].

The CRF post-processing step does not improve the performance of any of the models, but still, proves useful in

[3] https://github.com/microsoft/VoTT

homogenising clusters of pixels, and therefore, improving the performance of the subsequent OCR task on the separated text layers, as discussed in the next paragraph.

Table 1. Results achieved by FCN-light and WGM-MOD with and without CRF post-processing

	FCN-light	FCN-light +CRF	WGM-MOD	WGM-MOD +CRF
Mean printed IoU	0.24	0.11	**0.42**	0.41
Mean handwritten IoU	0.26	0.23	**0.36**	0.32
Mean background IoU	0.72	0.72	**0.74**	**0.74**
Total mean IoU	0.41	0.36	**0.50**	0.49

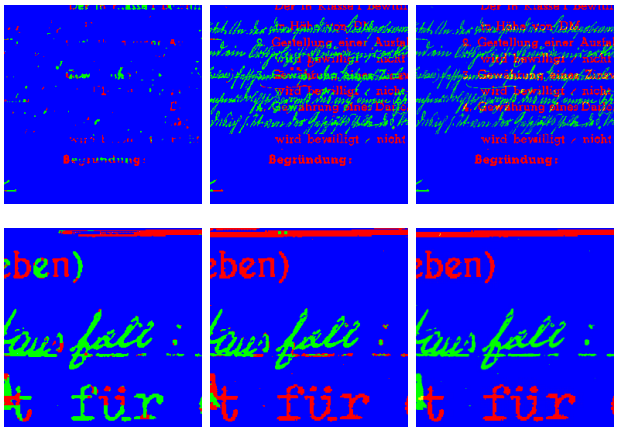

Figure 5. *Results of pixel-wise classification of two synthesised crops with FCN-light (left) and WGM-MOD (middle), and the ground truth (right)*

OCR evaluation. To measure the performance of the models on real archival documents, an extrinsic OCR evaluation has been performed on 30 historical documents from the "Wiedergutmachung" project. The ground truth for these documents have been created, including tags for the machine-printed and handwritten parts of the documents, by archivists at the State Archives of Baden-Württemberg. An extrinsic evaluation scenario is required since creation of pixel-level labels for original documents is unrealistic and labour-intensive. Therefore, instead of an IoU evaluation with real documents, the OCR performance is measured on the resulting separated machine-printed layers of the real documents.

After separation of machine-printed layers using FCN-light and WGM-MOD, OCR transcripts have been made for these layers using Transkribus[4]. Finally, these transcripts have been compared against the ground truth transcripts, with PRImA Text Eval 1.5[5] tool and using Flexible Character Accuracy as a reading-order-independent OCR performance measure [7].

The results of the OCR evaluation for the machine-printed pixels separated from the documents using FCN-light and WGM-MOD can be seen in Table 2. The significantly increased OCR performance on the separated machine-printed layers with WGM-MOD indicates a more accurate identification of the machine-printed and handwritten pixels by this model.

To observe the reported performance improvement on the real archival documents, the results of the separation task performed on the document crop in Figure 2 by the two models are provided in Figure 6.

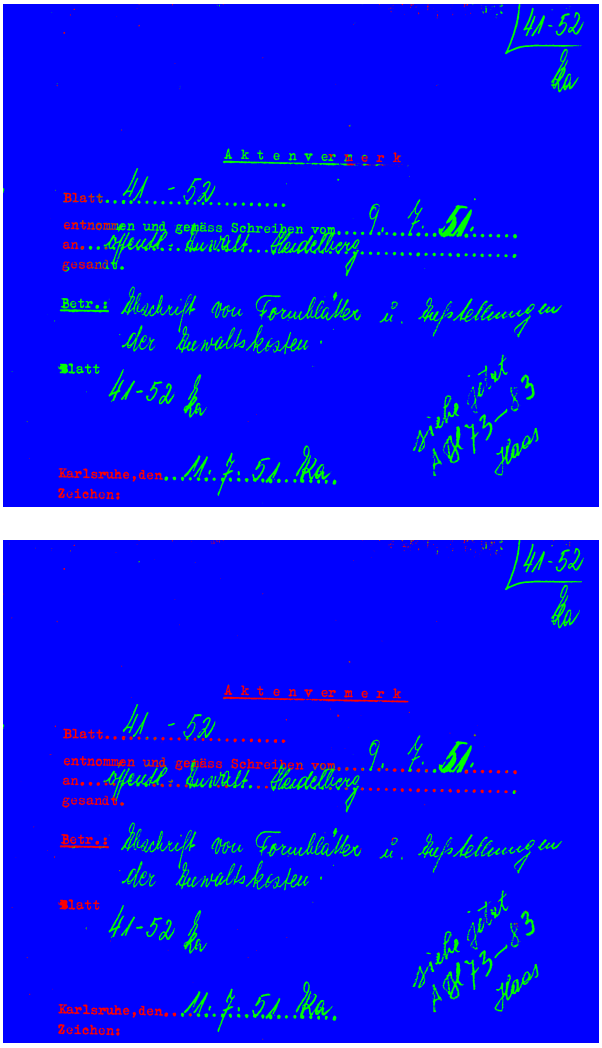

Figure 6. *Results of the separation task performed on the document crop in Figure 2 by the the FCN-light model (top) and the WGM model (bottom) with the CRF post-processing step*

Table 2. OCR performance on printed-pixel-separated layers of 30 Wiedergutmachung documents using FCN-light and WGM-MOD, with and without CRF post-processing

	FCN-light	FCN-light +CRF	WGM-MOD	WGM-MOD +CRF
Mean OCR	0.588	0.465	0.672	**0.723**

[4] https://readcoop.eu/transkribus
[5] https://www.primaresearch.org/tools/PerformanceEvaluation

Conclusions and Future Work

The results of the experiments on artificial and real documents show that creation of the WGM-SYN dataset and combining it with the FCN-light architecture, increases the performance of the state-of-the-art system for separation of machine-printed and handwritten text in historical archival documents. The performance improvement is even more significant in the downstream task of Optical Character Recognition (OCR), which is the goal in many of the projects with archival historical documents as input. As can be seen in Figure 2, the peculiarities of historical documents such as tinted ink and old fonts, e.g. typewriter fonts, create challenges for a model trained on modern documents, but can be handled by a model trained on historical data.

As part of future work, the amount of training data and its variety will be increased to include more font and handwriting styles. Using the data synthesis method, a larger amount of data can be augmented by resizing, translating and rotating the text images, to an extent that doesn't make the artificial data look unrealistic. On the other hand, in order to increase the amount of ground truth data for the evaluation process, crowdsourcing tools and frameworks will be considered, for whole or cropped documents which do not include sensitive information to comply with data privacy agreements.

One further step in recognition and separation of text types includes recognition of stamps, since they cannot be placed in the two classes of handwritten and printed text. Information in stamps is particularly interesting in WGM, since they indicate dates and offices involved in the processing of reparation cases. It has been noticed that the current model has difficulties classifying stamp pixels, with a tendency to recognise round stamps as handwritten text and stamps in a rectangular frame as printed text (see Fig. 7). This can be explained by the similarity of the round stamps with curves to handwritten text and vice versa. However, in order to increase the accuracy of the separation task, a stamp detection module will be added to the separation pipeline.

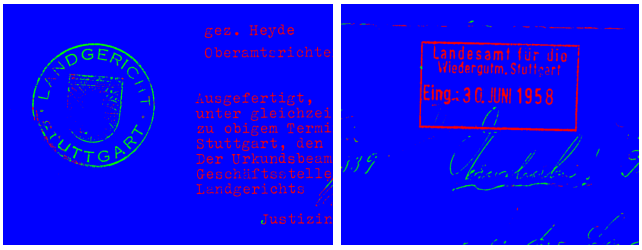

Figure 7. *Two stamps identified as handwritten (right) and printed (left) text by the WGM-MOD*

With regard to the model architecture, hyper-parameter optimisation for both the network and the CRF post-processing module will be performed. Moreover, replacing the weighted categorical cross entropy with focal loss [21] could also enhance the model performance.

Identification and separation of text types is a prerequisite for better OCR performance. Improving the OCR quality of various old fonts and handwritings is the next step in the information extraction pipeline from historical archival documents. In an endeavour to open up archives and increase accessibility and explorability of archives, it is also important to further identify and map persons, organisations, locations, and events to external resources and authority files [22]. Ultimately, this effort will give previously hidden insights about one of the darkest passages of human history and reveal stories that may otherwise be forgotten.

Acknowledgement

This work is funded by the Deutsche Forschungsgemeinschaft with grant number 396707386.

References

[1] N. Islam, Z. Islam, and N. Noor, "A survey on optical character recognition systems," arXiv preprint arXiv:1710.05703, 2017.

[2] B. M Garlapati and S. R. Chalamala, "A system for handwritten and printed text classification," in UKSim-AMSS 19th Int. Conference on Computer Modelling & Simulation, Cambridge, UK, 2017.

[3] J. K. Guo and M. Y. Ma, "Separating handwritten material from machine printed text using hidden markov models," in 6th Int. Conference on Document Analysis and Recognition, Seattle, Washington, USA, 2001.

[4] E. Kavallieratou, S. Stamatatos, and H. Antonopoulou, "Machine-printed from handwritten text discrimination," in 9th Int. Workshop on Frontiers in Handwriting Recognition, Tokyo, Japan, 2004.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015.

[6] O. Ronneberger, P. Fischer, and T. Brox, U-net: "Convolutional networks for biomedical image segmentation," in Int. Conference on Medical Image Computing and Computer-assisted Intervention, Munich, Germany, 2015.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in 26th Conference on Advances in Neural Information Processing Systems, Nevada, USA, 2013.

[8] J. Jo, H. I. Koo, J. W. Soh, and N. I. Cho, "Handwritten text segmentation via end-to-end learning of convolutional neural networks," Jour. Multimedia Tools and Applications, vol. 79, no. 43-44, pp. 32137–32150, 2020.

[9] N. Dutly, F. Slimane, and R. Ingold, "Phti-ws: A printed and handwritten text identification web service based on FCN and CRF post-processing." in Int. Conference on Document Analysis and Recognition Workshops (ICDARW), Sydney, Australia, 2019.

[10] M. A. Rahman, Y. Wang. "Optimizing intersection-over-union in deep neural networks for image segmentation," in Int. Symposium on Visual Computing. Springer, Cham, pp. 234-244, 2016.

[11] U. V. Marti, and H. Bunke, "The IAM-database: an English sentence database for offline handwriting recognition," Int. Jour. on Document Analysis and Recognition, vol. 5, no. 1, pp. 39-46, 2002.

[12] F. Kleber et al. "Cvl-database: An off-line database for writer retrieval, writer identification and word spotting," in 12th Int. Conference on Document Analysis and Recognition, Washington, DC, USA , 2013.

[13] A. Fornés, J. Lladós, J. Mas, J. M. Pujades, and A. Cabré, A, "A bimodal crowdsourcing platform for demographic historical manuscripts," in the First Int. Conference on Digital Access to Textual Cultural Heritage, Madrid, Spain, 2014.

[14] M. Terras, "Crowdsourcing in the Digital Humanities" in A new Companion to Digital Humanities, edited by S. Schreibman, R. Siemens, and J. Unsworth, John Wiley & Sons, 2016.

[15] A. G. Noll, "Crowdsourcing transcriptions of archival materials," in Graduate History Conference, Boston, USA, 2013.

[16] D. L. Hansen, P. J. Schone, D. Corey, M. Reid, and J. Gehring, "Quality control mechanisms for crowdsourcing: peer review, arbitration, & expertise at familysearch indexing," in Proceedings of the 16th ACM Conference on Computer-supported Cooperative Work, San Antonio, Texas, USA, 2013.

[17] N. Kligler, S. Katz, and A. Tal. "Document enhancement using visibility detection," in the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018.

[18] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[19] A. Dhawan, P. Bodani, and V. Garg, "Post Processing of Image Segmentation using Conditional Random Fields," in 6th Int. Conference on Computing for Sustainable Global Development (INDIACom), India, 2019.

[20] M. T. Teichmann and R. Cipolla, "Convolutional CRFs for semantic segmentation," arXiv preprint arXiv:1805.04777, 2018.

[21] L. Tsung-Yi, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in the IEEE Int. Conference on Computer Vision (ICCV), Venice, Italy, 2017.

[22] M. Vafaie, O. Bruns, N. Pilz, D. Dessi, and H. Sack, "Modelling Archival Hierarchies in Practice: Key Aspects and Lessons Learned," in 6th Int. Workshop on Computational History, HistoInformatics, online, 2021.

Author Biography

Mahsa Vafaie received her MSc. in Language & Communication Technologies as a joint degree from Saarland University and University of the Basque Country (2017). Currently, she is a junior researcher in the Information Service Engineering group at FIZ Karlsruhe, and a PhD student at KIT. Her research is focused on machine learning and knowledge representation for Digital Cultural Heritage.

Oleksandra Bruns received her M.A. in Computational Linguistics at Bielefeld University in 2020. She is a PhD student for Information Service Engineering at FIZ Karlsruhe and KIT. Her research work focuses on the areas of Semantic Web technologies, machine learning and NLP in the Cultural Heritage domain.

Nastasja Pilz received her MSc. degree in modern history, political science and translation studies from the University of Heidelberg. Since 2013, she has been working at the Landesarchiv Baden-Württemberg and was in charge of various projects focused on the archival role in the reappraisal of national socialist and postwar injustice. Currently, she and her team are responsible for the development of a digital portal providing access to all documents of German Wiedergutmachung of national socialist injustice.

Jörg Waitelonis received his PhD in Computer Science at KIT in 2018. Now he is a senior researcher in the Information Service Engineering group at FIZ Karlsruhe and CEO of yovisto GmbH. His research focuses on information retrieval and semantic search.

Harald Sack received his PhD in Computer Science at the University of Trier in 2002. He is full professor for Information Service Engineering at KIT and heads the ISE department at FIZ Karlsruhe. His research focus is knowledge representations, machine learning and the semantic analysis of unstructured data in the application context of innovative information systems.