https://doi.org/10.2352/issn.2168-3204.2023.20.1.7

Improvements in Handwritten and Printed Text Separation in Historical Archival Documents

Mahsa Vafaie, FIZ Karlsruhe and Karlsruhe Institute of Technology; Karlsruhe, Germany
Jörg Waitelonis, FIZ Karlsruhe and Karlsruhe Institute of Technology; Karlsruhe, Germany
Harald Sack, FIZ Karlsruhe and Karlsruhe Institute of Technology; Karlsruhe, Germany

Abstract

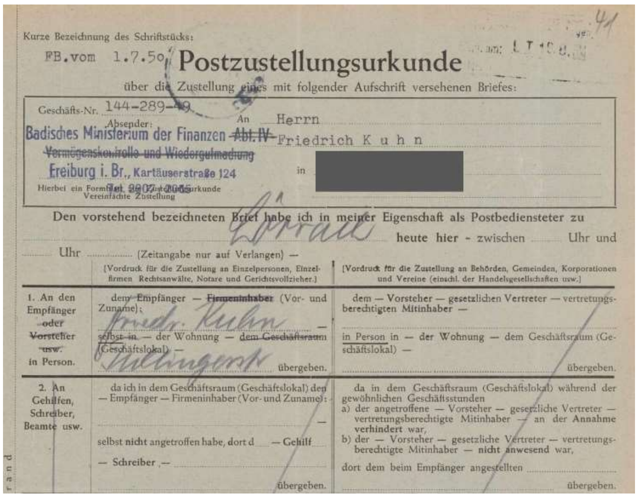
The presence of handwritten text and annotations combined with typewritten and machine-printed text in historical archival records make them visually complex, posing challenges for OCR systems in accurately transcribing their content. This paper is an extension of [1], reporting on improvements in the separation of handwritten text from machine-printed text (including typewriters), by the use of FCN-based models trained on datasets created from different data synthesis pipelines. Results show a significant increase of about 20% in the intrinsic evaluation on artificial test sets, and 8% improvement in the extrinsic evaluation on a subsequent OCR task on real archival documents.

Motivation and Problem Statement

Historical archival documents contain important information about certain historical events, places and people. Historians and historical research communities can benefit greatly from proper archival management, as well as information extraction pipelines that transform unstructured data stacked in archival institutes, into knowledge. Moreover, archival documents have always been at risk of deterioration and destruction as a result of ageing, war and other phenomena. Digitisation and transcription of archival material is one way to safeguard cultural heritage, enabling the establishment of an open information society, wherein archival content is readily retrievable, reachable, and accessible to the public.

Automated document processing, specifically the generation of transcripts for scanned documents through optical character recognition (OCR) systems, significantly expedites and streamlines the information retrieval process. However, to date, the quality of OCR systems is highly dependent on several factors, such as text and font type. Due to the distinct visual characteristics and structural attributes of machine-printed and handwritten text, OCR systems are typically designed to recognise either printed or handwritten text but not both [2]. In addition, archival documents, such as manuscripts, journals with handwritten annotations, and hand-filled forms and tables, often contain a significant amount of mixed text, where handwritten and machine-printed text are in close proximity or even overlapping. To improve OCR accuracy, it is beneficial to distinguish between printed and handwritten sections of text images and use the appropriate recognition system, i.e., printed OCR or handwritten OCR, accordingly [3].

Algorithms such as Hidden Markov Models (HMMs) and Support Vector Machines (SVMs) are commonly used for identifying and separating different types of text [3-5]. However, these traditional methods have limitations when dealing with cases where printed and handwritten text are combined and overlap. More recent solutions that involve deep neural networks, perform pixel-level labelling of text images [6-10]. These models are able to differentiate between printed and handwritten components, even



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
Hierbei ein Formblatt zur Zustellungsurkunde
Vereinfachte Zustellung
An Herrn
Friedrich K u h n
in [redacted]

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) den — 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*

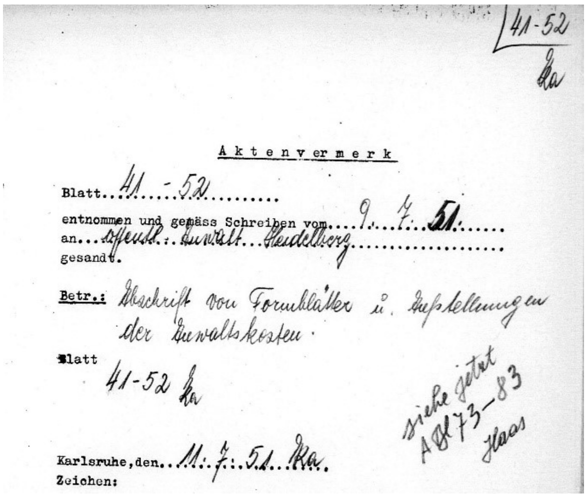
when they are in close proximity or overlapping. However, the majority of these methods are not very effective when it comes to accurately identifying text type in historical archival documents. The unique characteristics of archival documents pose a challenge for existing methods and models, unlike modern documents. As far as the authors know, there is currently no dataset available that includes historical archival documents along with their corresponding labels for identifying text types.

An additional performance factor in the separation of machine-printed and handwritten text in archival records arises from the media representation of the records. For example, archival records can be digitised by scanning either the original records (see Fig. 1) or microfilmed records (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.

In [1], we addressed the problem of automatic identification of handwritten and printed text in historical archival documents, in the context of the "Pilotprojekt zur Transformation der Wiedergutmachung"[1] [Pilot Project for Transformation of Reparations]. This project is focused on accessing knowledge hidden in historical records related to the claims of compensation and compensation proceedings that were submitted after 1945, in

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

the state of Baden-Württemberg, Germany. The records hold crucial information pertaining to one of the darkest periods in German history, offering first-hand accounts of atrocities committed against individuals who were persecuted and discriminated against on the basis of their ethnicity, religion, political beliefs, and sexual orientation during the National Socialist regime. Our contributions in [1] included creation of WGM-SYN, an artificial pixel-level ground-truth dataset made from historical data, and WGM-MOD, a model trained on this dataset. The data synthesis pipeline in [1] is initiated by denoising and binarising homogenous text regions extracted from original archival documents. This initial step generates an artificial dataset that has been cleansed of noise and document background texture, resulting in information loss when compared to original archival documents. In order to make artificial datasets that look more similar to original historical documents, new data synthesis pipelines are proposed and implemented.


41-52
Ka

Aktenvermerk

Blatt..41.-.52.........

entnommen und gemäss Schreiben vom....9.7.51.....

an...öffentl. Anwalt Heidelberg...................

gesandt.

Betr.: Abschrift von Formblätter u. Aufstellungen der Anwaltskosten.

Blatt 41-52 Ka

siehe jetzt A 8173-83 Haas

Karlsruhe,den..11.7.51.Ka.

Zeichen:

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

In this paper, which is an extension of [1], three different data synthesis pipelines for creation of artificial pixel-level annotated datasets are presented. These datasets are created from historical archival documents. Furthermore, different models for identification of printed and handwritten text trained on these datasets are introduced. These models are tested and compared with the baselines model from [1], in two different evaluation scenarios. All the artificial datasets and models trained on these datasets are publicly available on GitHub[2].

Related Work

Algorithms based on Hidden Markov Models (HMMs) and Support Vector Machines (SVMs) are traditionally used for text type identification and separation. These approaches differ mostly on the level of separation, with some operating at the line level and others at the word level. For example, in [5] the separation is conducted on the line level, and in [3] and [4] on the word level. These approaches rely on the visual and structural characteristics of text components. However, they do not address cases where printed and handwritten texts are intermixed or overlapping.

Recent solutions for text separation at the pixel level use deep neural networks [6-10]. These methods use pixel-level labels to separate printed and handwritten components, even when they overlap. The first method for pixel-level separation using end-to-end learning of a CNN was proposed in [9]. To address the problem of class imbalance, where the number of background pixels is much larger than foreground pixels, the authors introduced a new loss function that considers class frequencies and sample difficulties. To compensate for the lack of an appropriate dataset with pixel-level annotations, the authors also proposed a data synthesis method that creates pixel-level annotations and overlappings of different text types. The authors of [10] presented a lightweight Fully Convolutional Network (FCN-light) designed to recognise printed and handwritten text in overlapping areas of scanned documents. The model was trained on modern documents that contained a mix of machine-printed and handwritten text, but only a small percentage of those documents (9.84%) had regions with overlapping text. When tested on modern documents with simple layouts, FCN-light achieved a total mean Intersection over Union (IoU) score of 0.85. IoU is the standard performance measure that is commonly used for image segmentation problems [11]. However, the model's performance was weaker when applied to images with different formats and layouts [10].

To train a network for the pixel-wise separation, datasets with pixel-level annotations are required. The IAM [12] and CVL [13] datasets provide XML metadata that can be used to create pixel-level annotations for document images. These annotations can indicate whether the text is handwritten or machine-printed. Both datasets contain a block of machine-printed text followed by a handwritten version of the same text. The machine-printed texts are consistent in font and size across all pages, while the handwritten parts are written by different individuals. However, these datasets do not include parts where machine-printed and handwritten text overlap. Additionally, the documents in these datasets have simple layouts and clean backgrounds, which do not reflect the complexity of historical archival records.

Our previous work in [1] aims to solve the issue of identifying and separating handwritten and machine-printed text in historical archival documents. To achieve this goal, an artificial pixel-level annotated dataset is created from archival documents, and an FCN-based model is trained on this dataset. The preliminary test results indicate that the proposed model outperforms the state-of-the-art model trained on modern documents, in two different evaluation scenarios, on artificial data and actual historical archival documents. These findings suggest that the proposed approach of synthesising artificial data based on archival documents can significantly enhance the recognition accuracy of historical archival documents and contribute to improving the accessibility and explorability of historical records. Modifications to the data synthesis pipeline and the results of separation task using models trained on datasets created with the modified pipelines are described in the remainder of this paper.

Data Synthesis Pipelines for Text Type Identification in Archival Documents

In this section, the main contributions of this paper are presented. First, the different pipelines for synthesising artificial datasets using historical archival documents are described. These datasets are

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

derived from a subset of historical records from the "Pilotprojekt zur Wiedergutmachung". This subset includes 150 archival documents scanned in colour and 153 archival documents represented on black-and-white microfilms. These documents contain various formats, such as hand-filled forms, typewritten certificates, testimonials, and statements. The digitised documents are manually labelled using an open-source image annotation tool called VOTT[3]. In the annotation process, areas of the documents that contain only one type of text (machine-printed or handwritten) are selected, resulting in 117 handwritten and 320 printed text images from microfilm scans, and 202 handwritten and 447 printed text images from colour scans. In this section, following the description of the three data synthesis pipelines, the different datasets, model architecture and experimental settings are presented.

Data Synthesis. Conventional methods for the separation of handwritten and printed text have some significant limitations (e.g., [3-5]). Most of the already existing methods and models tend to show low accuracy in cases where machine-printed and handwritten text regions are overlapping. Moreover, the special characteristics of the archival documents compared to modern documents pose a challenge to existing systems. To the knowledge of the authors, there is no manually annotated dataset which contains historical archival records with overlapping machine printed and handwritten text and their corresponding labels. In an attempt to address this problem, [1] introduces an artificial pixel-level annotated dataset, called WGM-SYN, created from historical archival records from the reparation process and reparation cases filed after the fall of the Nazi Regime. In addition to this baseline system (Pipeline 1), two other pipelines are introduced which create artificial data more similar to original archival documents.

Pipeline 1. As presented in [1], this pipeline is based on the work by [8]. In this pipeline, regions containing only one type of text, i.e., printed or handwritten, are extracted from original archival records. To prevent inaccuracies in the labels caused by microfilm noise and background artefacts, all images are binarised using the document enhancement technique described in [14]. This first step results in "clean" homogenous crops. The handwritten crops are further processed (translated, rotated and resized) to increase the variety and in the final step merged with machine-printed crops to create black and white artificial crops with overlapping text types and their corresponding ground truth labels.

Pipeline 2. As opposed to Pipeline 1, this pipeline does not involve denoising in the initial stage. By using the extracted crops as they are, the final artificial crops look more similar to original documents with the same kind of noise and background artefacts that appear on the input homogenous crops. In this pipeline the ground truth labels are created from denoised images for accuracy and noise reduction.

Pipeline 3. The artificially synthesised data in pipeline 2 is binarised with threshold Sauvola to create another set of data without the information in the colour channels.

Datasets and FCN-based Model Training

Datasets. In order to create artificial datasets with overlapping machine-printed and handwritten text, each of the three pipelines explained above are used in combination with text region images extracted from microfilms and scanned documents (photos). This dataset division based on the representation medium (e.g., paper, microfilm, parchment, etc.) of archival records allows us to study the effects of representation medium on text separation in microfilms and scanned documents. The resulting datasets are then combined to create a mixed dataset. The dataset created from the combination of pipeline 1 and a mixture of microfilms and scanned documents (*bin_mix* in Table 1) is the same as WGM-SYN from [1]. Table 1 shows the titles for each of the synthesised datasets and their sizes. Each dataset is further divided into training (80%), validation (10%) and test (10%) sets.

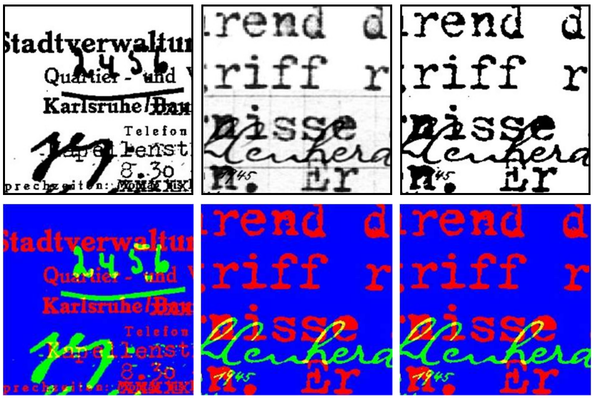


Figure 3. *Sample crops from each pipeline (top. synthesised crops, bottom. labelled crops; left: pipeline 1, middle: pipeline 2, right: pipeline 3)*

Model Architecture. Following [10], a variation of the FCN-8 architecture [6] is used for training, with 294,952 parameters to learn. The default number of epochs for each run is 50, and to avoid the problem of overfitting to validation data, early stopping has been implemented with a patience of 5 on validation loss. 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 pixels 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.

Table 1. Datasets based on data synthesis pipelines and input crops; # shows the number of synthesised crops in each dataset

	Microfilm (#)	Photo (#)	Mixed (#)
Pipeline 1	*bin_mf* (2,307)	*bin_ph* (1,953)	*bin_mix* (4,260)
Pipeline 2	*unbin_mf* (1,850)	*unbin_ph* (1,430)	*unbin_mix* (3,280)
Pipeline 3	*bin2_mf* (1,850)	*bin2_ph* (1,430)	*bin2_mix* (3,280)

The classification pipeline includes a dense Conditional Random Field (CRF) module in the post-processing stage. This allows the label assigned to a pixel in a particular location to be adapted, using probabilistic inference to estimate the most likely label for each pixel, given the observed image data and contextual information from neighbouring pixels. In fully connected CRFs, each node is connected to all other nodes in the image. By applying CRFs, pixels with similar features are assigned to the same predicted class, maximising label agreement between similar pixels [15].

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

Table 2. Top. Results of the intrinsic evaluation measured by IoU, achieved by different models
Bottom. Mean OCR performance measured by Flexible Character Accuracy on printed pixel-separated layers of archival documents

		Models trained with microfilm crops			Models trained with photo crops			Models trained with mixed crops		
		Pipeline 1 *bin_mf*	Pipeline 2 *unbin_mf*	Pipeline 3 *bin2_mf*	Pipeline 1 *bin_ph*	Pipeline 2 *unbin_ph*	Pipeline 3 *bin2_ph*	Pipeline 1 *bin_mix [1]*	Pipeline 2 *unbin_mix*	Pipeline 3 *bin2_mix*
Total Mean IoU		52%	71%	71%	51%	69%	70%	50%	71%	70%
Mean OCR	30 Scanned documents	67.24%	73.77%	71.40%	63.44%	73.54%	**80.06%**	72.34%	76.25%	79.34%
	50 Microfilm documents	73.43%	**83.25%**	76.26%	48.15%	56.18%	67.91%	64.37%	68.23%	70.25%

Intrinsic and Extrinsic Evaluation

In this section, the performances of the different models are presented in two different evaluation scenarios. *IoU (Intersection over Union)*, as an evaluation metric commonly used in computer vision, is the measure for intrinsic evaluation, tested on an unseen subset of each synthesised dataset. Moreover, a subsequent *OCR* task is carried out, and the results are provided to examine the impact of different text separation models on the recognition of machine-printed text in archival documents. Since creation of pixel-level labels for original documents is unrealistic and labour-intensive, this extrinsic evaluation scenario seems necessary.

IoU Evaluation. The performances of the different models are evaluated using artificially synthesised test sets from the same dataset the model was trained on, with corresponding ground truth labels and Intersection over Union (IoU) as the evaluation metric. IoU measures the overlap between the predicted segmentation mask and the ground truth mask. It is calculated by dividing the area of the intersection between the two masks by the area of their union. The IoU score ranges from 0 to 1, with higher scores indicating better performance. A score of 1 means that the predicted and ground truth masks perfectly overlap, while a score of 0 means that there is no overlap at all. The results of this evaluation are shown in the top part of Table 2. The IoU is calculated for all three different classes, i.e., machine-printed, handwritten and background pixels, and the average over these three classes is shown in the table as Total mean IoU. The obtained results indicate a significant improvement of about 18-21% when recognising printed and handwritten text using models trained on datasets created with the new data synthesis pipelines (i.e., unbin and bin2 datasets).

OCR Evaluation. The performance of various models on real archival documents was evaluated using an extrinsic OCR assessment, since creating pixel-level labels for original documents is impractical and labour-intensive. In this case, the quality of the OCR is measured on the resulting separated machine-printed layers of the actual documents, rather than through an IoU evaluation. This allows us to observe how well the text separation models perform on real historical archival documents.

OCR transcripts for these layers are created using the open-source OCR engine Transkribus print 0.3[4] with a CER of 1.60% on validation set. The transcripts are then compared against the ground truth transcripts, with PRImA Text Eval 1.5[5] tool and Flexible Character Accuracy as a reading-order-independent OCR evaluation metric [8]. The accuracy of reading order depends on how well the complex layouts of archival documents are analysed and detected, and thus, it is rather irrelevant to our goal in the text separation task. The evaluation is performed on a set of 30 historical documents from the "Wiedergutmachung" project available in digital form as scanned documents and a set of 50 documents from the same project available in a different media representation, as microfilms. The same documents were used in [1] for OCR evaluation. The ground truth transcripts for these documents have been created by archivists at the State Archives of Baden-Württemberg. The second part of Table 2 shows the results of the OCR evaluation for the machine-printed layers separated from these documents using different models.

Conclusions and Future Work

As mentioned previously, WGM-SYN from [1], is the same dataset called *bin_mix* here and the performance of WGM-MOD, which is the baseline model trained on WGM-SYN, is the same as the model trained on *bin_mix* in this work. The results of the experiments presented in Table 2, show that the use of datasets created with pipelines 2 and 3 increase the performance of the models, in both evaluation scenarios. The highest flexible character accuracy on scanned documents is achieved by training a model on crops generated from scanned documents. Table 2 shows that the model trained on *unbin_ph* increases the performance of the OCR task by 8% compared to the baseline model, trained on *bin_mix*. Similarly, the highest OCR accuracy on microfilm documents is achieved by the model trained on crops generated from documents on the same representation medium, i.e., microfilms. Figure 4 shows the results of pixel-wise separation of text types on a scanned document using two models, *unbin_mf* (right), trained on unbinarised microfilm images and *unbin_ph* (left), trained on unbinarised scanned document images. It is easily noticeable on this figure that the model trained on scanned documents performs better when applied on an actual document of the same type. The difference is quite significant in identification of text type in the middle of the document, where printed text is misclassified as handwritten text with *unbin_mf*, but correctly classified with *unbin_ph*.

Expansion of the training datasets by adding more font and handwriting styles, and using data augmentation techniques to generate larger datasets can improve the performance of the models. To improve the recognition and separation of different types of text, a stamp recognition module should be included as well. The current

[4] https://readcoop.eu/transkribus

[5] https://www.primaresearch.org/tools/PerformanceEvaluation

Figure 4. *The results of pixel-wise separation of a scanned archival document with two different models unbin_ph (left) and unbin_mf (right)*

models struggle to correctly classify stamp pixels, mistaking round stamps for handwritten text and stamps in a rectangular frame for printed text [see Fig. 5]. This can be explained by the similarity of the round stamps with curves to handwritten text and vice versa. To enhance the accuracy of the separation task, a stamp detection module will be incorporated into the separation pipeline.

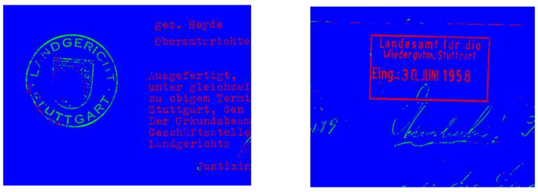


Figure 5. *Two stamps identified as handwritten (left) and printed (right) text*

With regard to the model architecture, replacing the weighted categorical cross entropy with focal loss [16] could also enhance the model performance.

To increase the performance of currently available OCR systems, it is advantageous to identify and distinguish between different types of text. The next step in extracting information from historical archival documents is to improve the quality of OCR for various old fonts and handwritings. To make archives more accessible and to facilitate exploration of archival documents, it is also crucial to identify and map individuals, organisations, locations, and events to external resources and authority files [17]. Ultimately, these efforts will reveal previously unknown information about human history and bring to light stories that might otherwise be forgotten.

References

[1] M. Vafaie, et al. "Handwritten and printed text identification in historical archival documents," in Archiving Conference, Society for Imaging Science and Technology, 2022.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.
[9] 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.
[10] 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.
[11] 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.
[12] 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.
[13] 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.
[14] 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.
[15] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.
[16] L. Tsung-Yi, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in the IEEE Int. Conference on ComputerVision (ICCV), Venice, Italy, 2017.
[17] 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.

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 system