DHQ: Digital Humanities Quarterly
Preview
2024
Volume 18 Number 1

Cross-codex Learning for Reliable Scribe Identification in Medieval Manuscripts

Julius Weißmann <weissmann_dot_julius_at_gmail_dot_com>, Media and Digital Technologies, St. Pölten University of Applied Sciences https://orcid.org/0000-0002-7086-941X
Markus Seidl <markus_dot_seidl_at_fhstp_dot_ac_dot_at>, Media and Digital Technologies, St. Pölten University of Applied Sciences https://orcid.org/0000-0002-7966-3602
Anya Dietrich <a_dot_dietrich_at_med_dot_uni-frankfurt_dot_de>, MEG Unit, Brain Imaging Center https://orcid.org/0000-0002-8564-1055
Martin Haltrich <m_dot_haltrich_at_stift-klosterneuburg_dot_at>, Library, Klosterneuburg Abbey https://orcid.org/0000-0002-2585-7188

Abstract

Historic scribe identification is a substantial task for obtaining information about the past. Uniform script styles, such as the Carolingian minuscule, make it a difficult task for classification to focus on meaningful features. Therefore, we demonstrate in this paper the importance of cross-codex training data for CNN based text-independent off-line scribe identification, to overcome codex dependent overfitting. We report three main findings: First, we found that preprocessing with masked grayscale images instead of RGB images clearly increased the F1-score of the classification results. Second, we trained different neural networks on our complex data, validating time and accuracy differences in order to define the most reliable network architecture. With AlexNet, the network with the best trade-off between F1-score and time, we achieved for individual classes F1-scores of up to 0,96 on line level and up to 1.0 on page level in classification. Third, we could replicate the finding that the CNN output can be further improved by implementing a reject option, giving more stable results. We present the results on our large scale open source dataset – the Codex Claustroneoburgensis database (CCl-DB) – containing a significant number of writings from different scribes in several codices. We demonstrate for the first time on a dataset with such a variety of codices that paleographic decisions can be reproduced automatically and precisely with CNNs. This gives manifold new and fast possibilities for paleographers to gain insights into unlabeled material, but also to develop further hypotheses.

1 Introduction

In recent years, there has been a notable surge in the digitization of historical documents, allowing for machine processing and providing unprecedented opportunities for researchers, including palaeographers. A pivotal challenge lies in the identification of scribes, recognizing the hands responsible for manuscript creation. However, this task is inherently time-intensive for researchers like palaeographers and codicologists, constraining its scope. 1

Simultaneously, the differentiation and recognition of scribal hands represent an essential methodology. This capability facilitates the determination of dates, specific scribal hands, or the geographic origins, enabling the formulation of insights into the trajectories of medieval scribes across diverse monasteries and scriptoria. Such information contributes significantly to our understanding of dating, place of origin, and insights into the scribes themselves and their organizational structures [Kluge 2019], [Landau 2004], [Powitz 2007]. 2

In European scriptoria the script *Carolingian minuscule* was used until the second half of 12th century for writing and copying books (codices). This Latin script was the first standardized script in the medieval period. Although codices 3 typically lack specific notations regarding which scribe wrote particular sections, the identification of scribes across various codices is invaluable for understanding scriptoria organization and the movement of codices and scribes between monasteries.

The historic auxiliary science of paleography addresses, among other aspects, the identification of scribes based on distinct, scribe-typical features in their writing. However, given the sheer volume of codex pages, this task remains tedious and time-consuming, demanding a high level of domain expertise. Automation through machines holds promise for expediting and enhancing the assignment of scribal hands, thereby relieving palaeographers from laborious and repetitive tasks, making the process faster and more comprehensive. 4

For instance, the library of Stift Klosterneurburg, amassed over the past 900 years, boasting nearly 300,000 copies. This remarkable collection, dating back to the 12th century, exemplifies the potential of automated techniques in efficiently identifying scribes within the vast repository of codex pages [Haltich 2014]. 5

Consequently, only a limited amount of medieval codices has been investigated with a focus on scribe identification. The automation utilizing pattern recognition and machine learning allows for larger amounts of material. However, to the best of our knowledge data of most automation-based approaches are either limited to even only one [De Stefano et al 2011], [Cilia et al 2020a], [Cilia et al 2020b] or a few different codices or include material from way larger time periods (e.g. [Fiel et al 2017], [Chammas et al 2020]). 6

To overcome these shortcomings, we investigate automated scribe identification on a large dataset we compiled: The CCI-DB [Seidl and Haltrich 2014]. This dataset contains 51 individual codices with a total amount of 17071 pages. These codices originate from the library of Klosterneuburg Abbey and were written in the late 12th century in Carolingian minuscule [Schneider 2014], [Bischoff 2004]. The scarcity of information about the scribes underscores the significant value of this new dataset and its associated processing capabilities. 7

We are aiming to answer two central questions: 8

1. Can the scribe assignments coming from decades of work by paleographic experts [Haidinger 1983], [Haidinger 1991], [Lackner 2012] be successfully modelled and predicted?
2. If so, can we use the models to predict scribes for codex pages that have unclear scribe assignments or no scribe assignments at all?

A substantial potential data specific risk seen in work by others that could render our work useless is that we could model not only script specific features but also book specific features such as the parchment and page margins. To mitigate this risk, we identified scribes that have been found in at least 3 codices. The subset we use contains 25200[1] random lines uniformly distributed over 7 scribes in 31 different historic codices, in order to train the models to recognize the scribe without codex specific features (see Figure 1).

In the last decade, convolutional neural networks[2] (CNNs) have proven to efficiently classify writers in modern and historic context and other tasks such as segmentation [Oliveira et al 2018], optical character recognition [Islam et al 2016], and writer identification [Xing and Qiao 2016], [De Stefano et al 2011]. 9

For our classification model we compared several general purpose object and concept detection CNNs as well as specific architectures for scribe identification (see Table 3). Surprisingly, the classic AlexNet [Krizhevsky et al 2012] provided the best trade-off between F1-score and time. We show that we can distinguish the scribes described by paleographic experts and even identify potential wrong scribe assignments. Furthermore, in combination with the reject option introduced by Cilia et al. (2020a) we demonstrate that we can reliably predict the scribes for codices with unclear or missing scribe assignments. 10

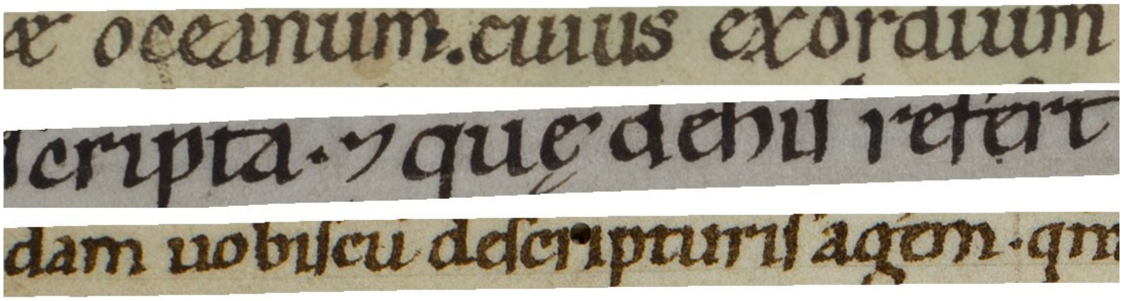


Figure 1. Examples from the CCl-DB [Seidl and Haltrich 2014] of three lines of different codices from one scribe. The ink and parchment appearance differs, although it's written from the same scribe (class A 30) [Haidinger 1991], [Lackner 2012]. Top: CCl 206, middle: CCl 197, bottom: CCl 217.

In this paper, we focus on three major issues: 11

- the importance of cross-codex based training data for automatic scribe identification
- the feasibility of training a model based on scribe assignments by the paleographers Haidinger (2010, 1983, 1991) and Lackner (2012)
- the necessity of exploiting the confidence in scribe predictions to reveal uncertainties in the dataset.

The latter is a central requirement, as there is no objective ground truth for scribe assignments for the medieval codices we are using. Our contribution in this paper is threefold: 12

Firstly, we demonstrate the necessity of book independent training of scribe models, which has been neglected in other studies. Secondly, we demonstrate that contrary to the results of Xing and Qiao (2016) standard architectures are sufficiently accurate to reliably identify medieval scribes in a classification pipeline. Thirdly, our work consequently facilitates comprehensive and convincing studies on large datasets and allows new insights into the historic monastic life and the relationships between the monasteries.[3] 13

The paper is structured as follows: after an overview about related work in Section 2 we outline the dataset in Section 3. In Section 4 we explain the applied methods for scribe identification. We show and discuss the results in Section 5 and finally conclude the paper in Section 6. 14

2 Related work

Computer-aided historic handwritten document analysis includes segmentation, text recognition, dating and writer identification as well as verification. Segmentation usually separates the written or drawn content from the carrier material (such as parchment or paper) [Oliveira et al 2018], [Tensmeyer et al 2017]. Based on this, a possible next step is handwritten text recognition (HTR) [Chammas et al 2018]. Either the segments or the written content alone or a combination of both modalities allow further investigations like dating, writer verification [Shaikh et al 2020], [Dey et al 2017] and identification [Chammas et al 2020], [Xing and Qiao 2016], [Fiel et al 2017]. 15

Different processes for scribe identification can be used, such as text-dependent and text-independent. Text-dependent [Said et al 1998] methods allow identifying the writer on particular characters or words, whereas text-independent [Yang et al 2016] approaches can be applied on new unseen content. Two kinds of handwritten text patterns can be used: on-line and offline writer identification. On-line [Yang et al 2016] systems work with time series of the formation process, while off-line [Xing and Qiao 2016] solutions are limited to the images of the written text document. 16

Writer identification is a strong topic in document analysis and therefore much discussed. In the last years, a variety of solutions have been provided. These methods can be grouped into codebook-based and codebook-free methods. Codebook-based refer to a codebook that serves as a background model. This model is based on statistical features like in [Maaten and Postma 2005]. The codebook-free methods include for example the width of ink traces, which was used from Brink et al. (2021) to predict the writer in medieval and modern handwritten documents, or the hinge feature 17 provided by [He and Schomaker 2014] in order to identify writers in handwritten English text in the IAM-dataset [Marti and Bunke 2002]. Further, there have been strong results in using the Scale-Invariant Feature Transform (SIFT) for writer identification [Xiong et al 2015], [Fiel and Sablatnig 2012], [Wu et al 2014].

In recent years, the number of Deep Learning[4] (DL) based studies in document analysis increased drastically [Chammas et al 2020], [Cilia et al 2020b], [Xing and Qiao 2016]. As mentioned in the introduction, the interest in using such techniques is due to their ability to provide powerful state-of-the-art solutions in an efficient and reliable way. During the training, the DL model learns the best fitting features for the classification. Therefore, no handcrafted features are required. For example, Fiel and Sablatnig (2015) demonstrated the strong performance of DL by employing CNNs for scribe identification on modern datasets. Xing and Qiao (2016) performed a writer identification on the modern IAM and the HWDB1.1 dataset. They developed a special multi-stream CNN architecture based on AlexNet and outperformed previous approaches on handcrafted features. Cilia et al. (2020a) demonstrated a comparison between deep learning and handcrafted features on the Avila Bible that is written in *Carolingian minuscule*. These classical features, as Cilia et al. call them, are handcrafted features, that have been developed in cooperation with paleographers. The results of their studies emphasize the effectiveness of deep learning features in contrast to the handcrafted features. 18

In our research, we investigate *scribe* identification in contrast to *writer* identification, since the specific individual of the writing is generally not known for our material. Scribe identification is discussed in a medieval context only in a very limited range, like at the International Conference on Document Analysis and Recognition (ICDAR) competitions [Fiel et al 2017] or in conjunction with the Avila bible [Cilia et al 2020a]. However, the aforementioned datasets are of limited use for our goals. The Avila bible is literally only one codex and consequently does not allow cross-codex scribe identification. The datasets used in the historical ICDAR competitions [Fiel et al 2017], [Christlein et al 2019] span time periods of many centuries, and hence include different scripts and carrier materials. Thus, scribe identification in the large amounts of medieval codices in Europe’s libraries is still a challenge and our approach allows novel insights as it focuses on a wide range of codices of several scribes in the short period of the late 12th century. 19

3 Dataset

We perform experiments on a subset (see Table 1) of seven scribes provided by the CCI-DB [Seidl and Haltrich 2014]. We selected the scribes which have contributed to as many books as possible to allow cross-codex evaluation. Samples in the dataset were handwritten on parchment in *Carolingian minuscule* (see Figure 1 and Figure 2) on one- and two-column pages. These codices have been written down in the scriptorium of Klosterneuburg in the last third of the 12th century. The data is provided by the Scribe ID AI[5] project and has been labelled by paleographic experts based on the activity of the paleographers Haidinger (2010, 1983, 1991) and Lackner (2012). 52 labelled codices are provided in the CCI-DB. This database enables new possibilities within document analysis and especially in handwriting recognition. 20

Class	Training codices (Test A and Test B)			Separate codices (Test B only)		
	Lines per codex	#Codices	Codices	Lines per codex	#Codices	Codices
A 30	450	8	30, 31, 197, 206, 226, 246, 256, 257	1500	2	32217
A 259	1200	3	209, 259, 949	1754	1	706
B 259	720	5	259, 622, 706, 212, 671	1439	1	246
A 20	1200	3	21, 28, 39	3000	1	20
B 20	900	4	22, 195, 216, 764	119	1	20
A 215	1200	3	215, 219, 703	3000	1	245
A 258	1800	2	258, 707	3000	1	203

Table 1. Subset of the CCl-DB. The dataset for our experiments is a subset of the seven common scribes of the CCl-DB [Seidl and Haltrich 2014]. We generated a dataset, with two groups of codices – the "Training Codices" and the "Separate Codices". From the "Training Codices" we choose for every class 3600 random lines, that are uniformly distributed over all books. In total these are 25200 lines that are separated into training, validation and test data (Test A) according to the ratio of 60%, 20% and 20% respectively. The separate books are used for an extra test set Test B. There are up to 3000 lines of one class, depending on the codex-size.

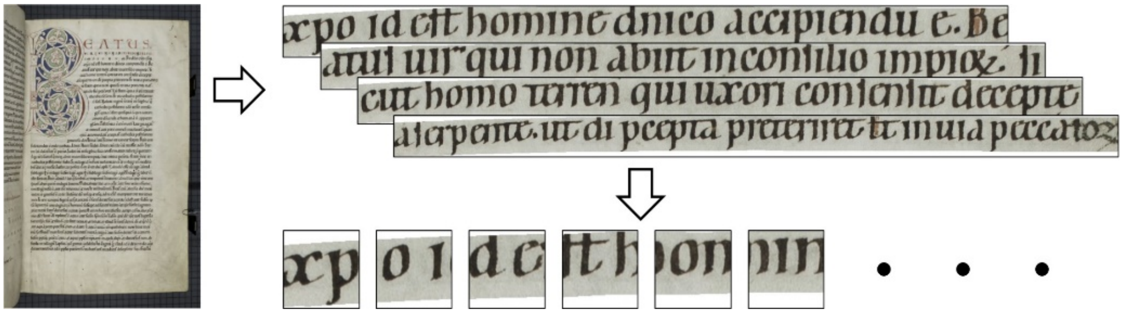


Figure 2. Image data of the CCl-DB [Seidl and Haltrich 2014] (CCl 20, S. 3r, hand A 20 [Haidinger 1983]). The CCl-DB provides the codices page by page (left) and on line-level (right-top). The line-level images are produced automatically by the segmentation of Transkribus [Kahle et al 2017]. The neural network classification works with squared images, therefore we cropped the line images into squares and resized them into the network input size.

To the best of our knowledge, there is no comparable database available that provides the workings of many medieval scribes in various codices and in such a short period of time. 21

4 Methods for Scribe Identification

This section will introduce the pipeline of our line-based scribe identification approach. The pipeline is grouped into three main parts (see Figure 3). In the first part – the image preprocessing – we focus on providing the following network input images that are reduced to their scribe specific information. The second part presents the neural network classification approach. Here we compare different CNNs as image classifiers for scribe identification. Finally, the third part covers the post-processing which generates the final score based on line- and page-level. This is where we introduce the computation of the line score and the reject option – a method to improve the prediction. All parts will be detailed in the following. 22

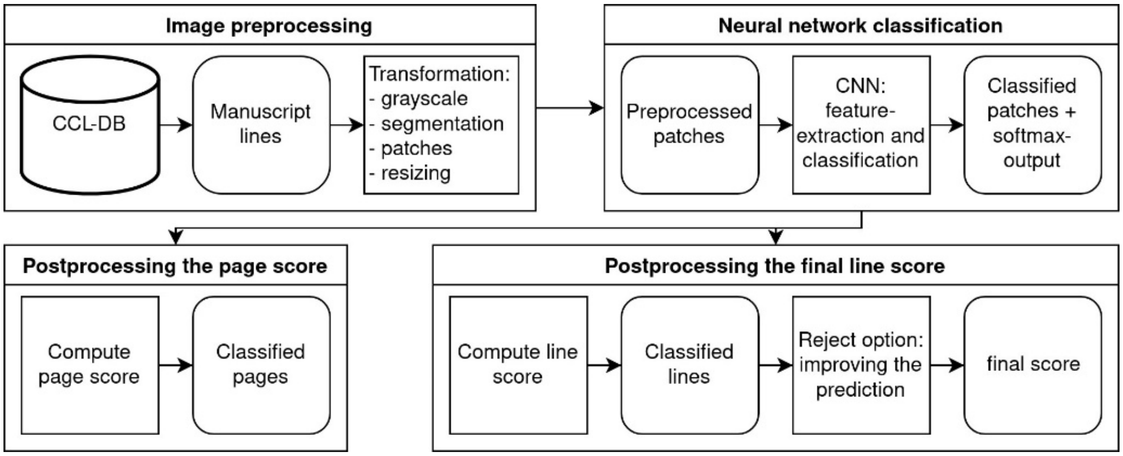


Image preprocessing
CCL-DB
Manuscript lines
Transformation:
- grayscale
- segmentation
- patches
- resizing
Neural network classification
Preprocessed patches
CNN: feature-extraction and classification
Classified patches + softmax-output
Postprocessing the page score
Compute page score
Classified pages
Postprocessing the final line score
Compute line score
Classified lines
Reject option: improving the prediction
final score

Figure 3. Overall procedure of the proposed scribe recognition system. The rounded boxes symbolize data, whereas the angular boxes show processes.

4.1 Image preprocessing

The dataset contains not only the codex pages, but also the extracted lines. The line data is usually correctly extracted from the pages, however there are some small snippets with no noticeable content in the dataset. As image lines are generally of wide aspect ratio, we use a simple heuristic ($\mathrm{width}/\mathrm{height} \leq 5$) to skip these uninformative snippets already in the step of preprocessing. 23

For studying the optimal input image data we generated grayscale images and masked grayscale images additionally to the RGB data (see Figure 4). When we masked[6] the images we followed the example of De Stefano et al. (2018). We applied their fixed threshold value, equal to 135, to separate the ink and the parchment of the grayscale images (see Figure 4). In related work [Cilia et al 2020a], [Fiel et al 2017], [Fiel and Sablatnig 2015], [Cloppet et al 2016], binarization is often applied to text images. However, since our masking does not work reliably enough to produce meaningful binarization, we omit this step in this work. 24

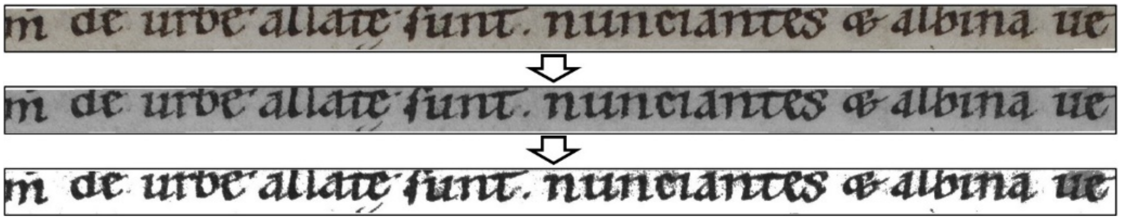


Figure 4. Example for the preprocessing (CCl 212, S. 1r, hand B 259 [Lackner 2012]). The lines of the CCl-DB [Seidl and Haltrich 2014] are provided as RGB images (top). From these, we converted the images to grayscale (middle). The masked grayscaled images are produced by removing the background from the ink.

As already mentioned, the lines of the CCl-DB are of different aspect ratio, but the networks we implemented are working with a fixed input size. In order to handle the different lengths of the lines, we followed the patch scanning strategy of Xing and Qiao (2016). First, the images have been resized in height to the specific network input image height, while maintaining the aspect ratio of the line. Afterwards, we cropped the lines from left to right into patches (see Figure 2). This sliding window comprises the network specific input image width. Due to the large dataset (see Table 1), there is no need for data augmentation.[7] Hence, we generated patches with no overlap. Only one overlap occurs at the last image of each line, as the last patch of the line is generated by positioning the sliding window at the end of the line. Finally, we scaled and normalized the patches. 25

4.2 Patch level classification

Xing and Qiao (2016) achieved high identification accuracies with their customized CNN's on the line level data of the 26

IAM [Marti and Bunke 2002] dataset. They optimized AlexNet to the task of writer identification on the IAM dataset and denoted the architecture Half DeepWriter. Next, they developed the DeepWriter architecture, which is an improvement of Half DeepWriter that enables the computation of two sequential image patches in a single pass with a multi-stream architecture. Xing and Qiao showed that DeepWriter produces the best results on the line level data of the IAM dataset. Therefore, we implemented these three auspicious architectures as per description of Xing and Qiao (2016), when we tested the potential of preprocessed images on our data (see Table 1).

27 Additionally, we compared several other general-purpose object and concept detection architectures in our study to find the best one suited to our specific data. For this purpose, we used models provided by Torchvision [Marcel and Rodriguez 2010] (see Table 3) and only adapted the input layer to the grayscale images and the output layer to the seven classes.

28 As shown in different studies, pre-training and fine-tuning a CNN can lead to better results [Xing and Qiao 2016] , [Studer et al 2019]. Xing and Qiao (2016) demonstrated this on the IAM [Marti and Bunke 2002] and the HWDB [Liu et al 2011] dataset. Therefore, we trained the described models on the IAM dataset, fine-tuned them on our data and compared the results with the from scratch trained weights.

29 The models have been trained with a batch size of 32 over 10 epochs with a learning rate of $1 * 10^{-5}$ on ADAM. The error was calculated with the cross entropy. Over all ten epochs, the instance of the model that performed best on the validation data has been saved for the next steps of the experiments.

4.3 Postprocessing

30 We pursue a patch, line and page level classification, but as already described, the network classification is on patch level. To compute the line and page score, we follow the example of Xing and Qiao (2016). They calculate the final score vector f_i for the j^{th} writer, of all patches N of one line: $f_i = \frac{1}{N} \sum_{i=1}^{N} f_{ij}$. This averaged Softmax output serves as the basis for the final step of the post-processing – the reject option.

31 Cilia et al. (2020a) proposed the reject option to generate more reliable results for the writer identification on the Avila bible. They showed that sometimes it is better to withdraw a precarious decision than accepting all predictions. In such a case, they reject the prediction with the reject option.

32 We used the line score as a probability distribution to check the probabilities for all writers. As Cilia et al. explained, the error-reject curve shows the impact of the reject rate to the wrong predictions and allows finding the optimal threshold for rejecting a prediction. Our reject rate is given by the line score of one writer. The reject rate corresponds to the wrong predictions.

5 Results

33 The purpose of this study was to train a model that classifies reliable and efficiently scribes in cross-codex data. We want to enable the automatic continuation of the work of paleography experts following their example, in order to allow research in large scale. In this study we would like to find a model that is not only reliable but also fast in processing, as it is the basis for research on active learning.

34 In Table 2 we show the importance of cross-codex test data and the risk of overfitting.

35 In the evaluation of our scribe classification pipeline, we found:

1. Image preprocessing plays a key role in cross-codex scribe identification. In comparison to RGB images, masked grayscale images roughly doubled the F1-score in the classification task.
2. Further, we showed that AlexNet provides very fast, and among the most reliable predictions in classifying the scribes of our data set.
3. Contrary to expectations, pre-training the network on the IAM database leads to worse results, which is why we omitted this step.

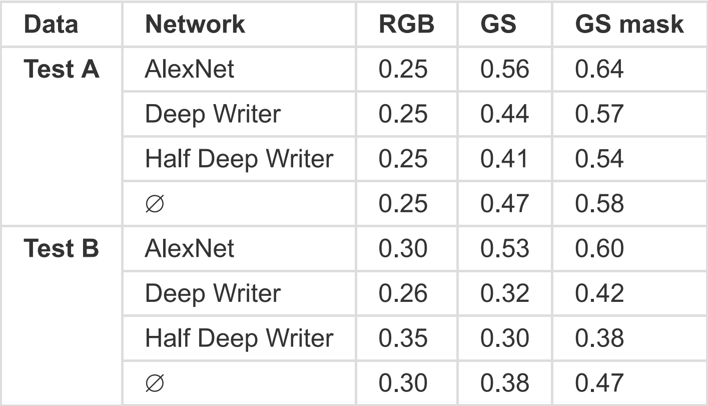
Applying the best fitting trained model, it turned out that it is very effective and even indicates incorrect data. 36

Moreover, we introduce the reject option on our dataset in order to get rid of precarious classifications and found that it underlines the results. 37

Finally, we deployed the pipeline to processes open paleographic topics. 38

5.1 Cross-codex data

39 The CCI-DB provides handwritings of several scribes in different codices. Therefore, we compared two test sets (see Table 2) to check if the networks tend to learn codex-specific features. For this experiment we trained the architectures AlexNet, Half Deep Writer and DeepWriter. Referring to Xing and Qiao (2016) these networks are suitable for handwriting identification. We found, that the test set Test B which contains test samples from books which have not been used for training (see Table 1) is more comprehensive than Test A because all three trained networks performed better on Test A whether the input images have been RGB grayscale or masked. We conclude, that the networks learned codex-specific features in Test A. Therefore, we used the test set Test B for further experiments to obtain more reliable results.



Data	Network	RGB	GS	GS mask
Test A	AlexNet	0.25	0.56	0.64
	Deep Writer	0.25	0.44	0.57
	Half Deep Writer	0.25	0.41	0.54
	∅	0.25	0.47	0.58
Test B	AlexNet	0.30	0.53	0.60
	Deep Writer	0.26	0.32	0.42
	Half Deep Writer	0.35	0.30	0.38
	∅	0.30	0.38	0.47

Table 2. F1-score for image preprocessing on patch level. Preprocessed input images and their impact on the classification of the test data. The experiment was performed on the three different networks AlexNet, Half DeepWriter and Deepwriter. As a result, the averaged F1-score is given. We provide two F1-scores of separate test sets (see Table 1).

5.2 Classification pipeline

40 To find the best type of input image for the CNN classification, we preprocessed the dataset in three different ways. We compared RGB images with grayscale and masked grayscale images and found that masked grayscale images produce the best F1-score in the test data (see Table 2). Consequently, the following experiments are based on this powerful image preprocessing. The masking has proven to be effective enough even though we used a simple threshold-based algorithm that sometimes does not reliably distinguish between ink and parchment. Hence, we could replace it in further studies by a better learning-based solution.

Network	F1 Test B	Time	Image (height, width)
Densenet*	0.61	503	224, 224
AlexnetNet	**0.60**	**115**	227, 227
ResNet18*	0.56	164	224, 224
Inception v3*	0.55	683	299, 299
VGG*	0.53	610	224, 224
SqueezeNet*	0.50	135	224, 224
MNASNet*	0.43	196	224, 224
DeepWriter	0.42	66	113, 226
Half DeepWriter	0.38	62	113, 113

Table 3. Network performance on patch level. Nine different CNN architectures are trained on the data of Table 1 to compare their performance on patch level. The weighted averaged F1-Score we use, is measured on Test B (see Table 1) and rounded to two decimal places. The training-time is given in rounded minutes. The models marked by * originate from the torchvision library.

As there are different networks available for image classification, we compare in Table 3 nine powerful architectures. AlexNet achieves with an F1-score of 0.60 on patch level and 115 minutes training time the best trade-off between F1-score and time. Only DenseNet achieved a small improvement of 0.01 in comparison to AlexNet but with a training time of 503 minutes it is much less time efficient. Even though latest state-of-the-art models performed better, with a growing number of parameters, the processing time would be impractical for our purpose, and as already shown (see Table 2) data-centric approaches such as masked grayscale images are more influential. Given these F1 and training time results, we claim AlexNet to be best suited for our purposes and thus used it for all further experiments. 41

To understand whether pre-training is beneficial, we follow the example of Xing and Qiao (2016). Accordingly, we pre-trained AlexNet on 301 writers of the IAM dataset and fine-tuned the model on the CCI-DB data. The pre-trained and fine-tuned model generates an F1-score of 0.58 whereas the from scratch trained model outperformed this result with an F1-score of 0.60 on page level. Because of the lower F1-score of the pre-trained and finetuned model model, we assume that there are not enough shared features between the CCI-DB and the IAM dataset. However, as shown by Studer et al. (2019), models in general benefit from pre-training. We assume, that larger and more comprehensive datasets than the IAM handwriting database could improve our model. 42

5.3 Automatic paleographic classification

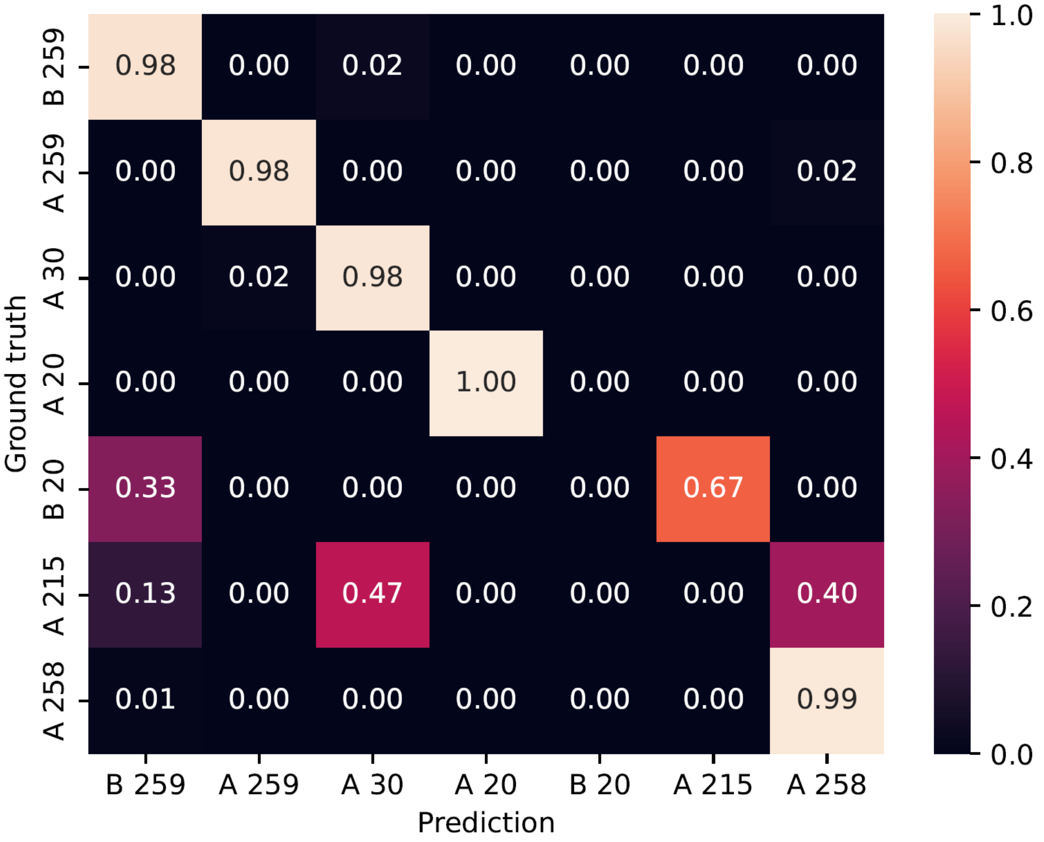
To figure out, which reliability values can be reached by the trained AlexNet, we evaluated the data of Test B. The main experiments are performed on line level (see Figure 7), but we also provide test results on patch and page level (see Table 4). Furthermore, we show the confusion matrix of the same test data on page level in Figure 5. We observe, that the line level classification generally reinforces the patch level results and the page level classification generally reinforces the line level results. 43

Another particularity in Table 4 and Figure 5 are the dichotomous scores. Five of seven classes are predicted well, such as F1-scores up to 1.0 on page level and 0.96 on line level in the case of A20 (see Table 4). Only the two classes B 20 and A 215 seem to be less precisely predicted on the test data. We observe low F1-Scores of 0.0 on page level as well as 0.24 and 0.06 on line level respectively. However, the low predictions for the two classes B 20 and A 215 strengthen the hypothesis of a powerful model as investigations of our paleographic partners revealed the labelling of these classes to be most probably incorrect. The paleographers actually labeled this test data as several not defined classes. Thus, the classification of the two classes proposed by the network might indeed correspond to the correct scribes and could therefore give new insight. However, confirmation by future research using our approach would be needed. 44

Scribe	F1-p	F1-l	F1-pg	#-p	#-l	#-pg
B 259	0.60	0.76	0.85	31309	1340	42
A 259	0.79	0.94	0.98	34004	1672	52
A 30	0.62	0.76	0.74	56173	2805	66
A 20	0.90	0.96	1.00	71565	2814	72
B 20	0.05	0.24	0.00	1957	111	3
A 215	0.07	0.06	0.00	65689	2897	92
A 258	0.68	0.79	0.83	67520	2893	96

Table 4. Results of Test B. F1-score on Test B(see Table 1). The F1-score is measured on patch- (p) and line- (l) and page-level (pg). The weighted averaged test data are random lines of unseen books, as these lines are of various length, they result in a different number of patches. Due to the image preprocessing the total of samples is slightly lower than in Table 1.

45



Ground truth \ Prediction	B 259	A 259	A 30	A 20	B 20	A 215	A 258
B 259	0.98	0.00	0.02	0.00	0.00	0.00	0.00
A 259	0.00	0.98	0.00	0.00	0.00	0.00	0.02
A 30	0.00	0.02	0.98	0.00	0.00	0.00	0.00
A 20	0.00	0.00	0.00	1.00	0.00	0.00	0.00
B 20	0.33	0.00	0.00	0.00	0.00	0.67	0.00
A 215	0.13	0.00	0.47	0.00	0.00	0.00	0.40
A 258	0.01	0.00	0.00	0.00	0.00	0.00	0.99

Figure 5. Confusion matrix of the data of Test B on page level (see Table 3).

5.4 Reject Option

46 To investigate whether implementing a reject option improves results we tested the five classes (B 259, A 259, A 30, A 20, A 258) on it. These are the classes shown previously without conflicts in the test data of Test B. Figure 6 shows that increasing the reject rate minimizes the error. The reject curves of Figure 6 drop quickly, indicating a strong influence of the threshold. As Cilia et al. (2020a) explained, the reject option is therefore suitable for the scribe identification on the CCl-DB. Therefore, we choose a reject option based on the threshold of 40 %, as it ensures low error with high sample rate. As class B 20 and A 215 are difficult to evaluate from the test data, the threshold is adapted to 60%.

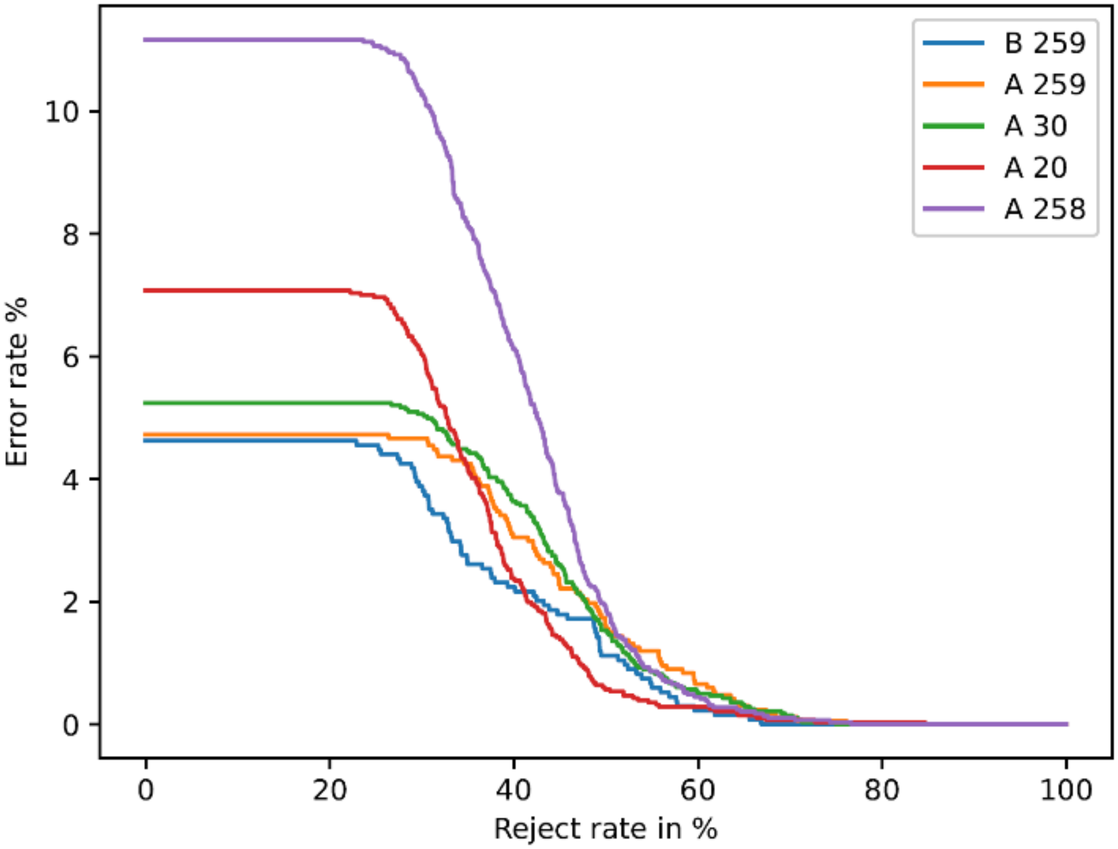
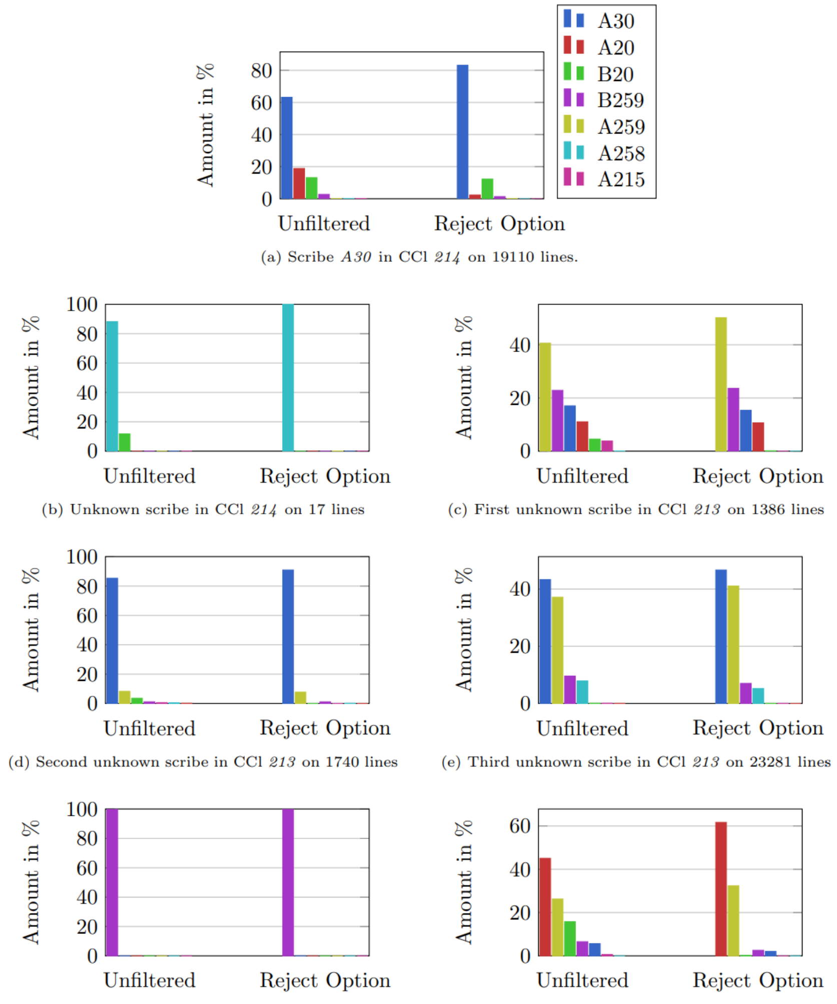


Figure 6. Error-reject curves for five different scribes on the data of Test B on line level (see Table 1).

5.5 Into the wild: Using our model on data with unknown scribes

The central aim of our approach is to contribute new insights for paleography. Therefore, we examined sections from the codices that the experts limited to one scribe, although they could not determine the exact individual. As shown in Figure 7 the trained model contributes meaningful classifications for these parts. The first plot can be considered as reference, this is the part of CCl 214 written by A 30. In this example, the model recognizes class A 30 as main class. The remaining six plots can be differentiated into two groups. On the one hand there are plots b, d and f which give significant results that refer to one scribe class B 20, A 30 and A 215 respectively. On the other hand, there are plots like in c, e and g that produce diffuse classification, not focused on one class. We conclude that these less significant predictions are caused either by scribes the model hasn't learned yet or by more than one scribe. 47



(a) Scribe *A30* in CCl *214* on 19110 lines.

(b) Unknown scribe in CCl *214* on 17 lines

(c) First unknown scribe in CCl *213* on 1386 lines

(d) Second unknown scribe in CCl *213* on 1740 lines

(e) Third unknown scribe in CCl *213* on 23281 lines

(f) Fourth unknown scribe in CCl *213* on 25 lines

(g) Fifth unknown scribe in CCl *213* on 899 lines

Figure 7. Scribe identification on line level and with reject option. All figures are based on the parts of one scribe. These new codices are labeled with one given (a) and six unknown scribes (b-g).

6 Conclusion

48 In this paper, we want to study the question of how to train a reliable and efficient model that allows cross-codex scribe identification in the strongly standardized medieval Carolingian minuscule of the CCl-DB. To this aim, we first figured out the risk of codex specific overfitting and showed the importance of cross-codex data to overcome this issue. We also found, that the reduction of RGB-images to grayscale masked images helps the network to focus on scribe specific features and leads to significantly better results.

49 After comparing several networks, AlexNet was used in our pipeline to generate a classification on patch, line and page level. Finally, we improved the final score by implementing the reject option.

50 One of the limitations of the proposed method is the basic segmentation, which is challenging on the historic parchment. This limitation leads to a natural direction of future work, focusing on improving the segmentation method that also allows binarization. The outcomes of these investigations currently form the foundation for advancing automated scribe identification. The recognition system described in this paper has been integrated into the backend of an active learning application, while concurrently, we are collaboratively developing an intuitively accessible application with continuous annotations in close cooperation with paleographers. The inclusion of a visual interface will empower experts to scrutinize and refine predictions, facilitating an iterative process of retraining our model and validating it against new scribe hypotheses. These findings aim to unlock novel possibilities and analytical tools, fostering a more profound comprehension of diverse medieval scriptoria. Going forward, this research endeavors to bring researchers closer to

addressing open questions regarding the organizational aspects of scriptoria in the high medieval monasteries of (Lower) Austria, with additional evidence and interpretations serving as valuable support.

Funding

This work has received funding from the Lower Austrian FTI-Call 2018 under grant agreement No FTI18-004 (project Active Machine Learning for Automatic Scribe Identification in 12th Century Manuscripts). Moreover, the work was supported by Erasmus+ from the German Academic Exchange Service (DAAD). 51

Acknowledgments

We would like to thank the team of the Klosterneuburg abbey library, as well as the team of the institute of Creative\Media/Technologies of the St. Polten University of Applied Sciences for their help. 52

Notes

[1] A typical single column page contains 31 or 32 lines. The vast majority of our books is in single column layout, hence we can roughly estimate that the 25200 lines correspond to 800 pages.

[2] Recently, artificial neural networks have demonstrated remarkable efficiency in handling unstructured data, including images, text, and speech. Convolutional Neural Networks (CNNs), as a specialized architecture, have proven highly beneficial in image processing by adeptly learning spatial hierarchies of features.

[3] The code for the experiments is publicly accessible: https://gitlab.rlp.net/studiengang-digitale-methodik/abschlussarbeiten/ma-repo/-/tree/main/experiments

[4] Deep learning is a subset of machine learning that involves neural networks with multiple layers (deep neural networks) to model and process complex patterns in data.

[5] See: https://research.fhstp.ac.at/en/projects/scribe-id-ai

[6] Image masking is a technique of selectively concealing or revealing specific portions of an image. In this example, masking is done by removing the parchment/background in the image, to focus the model's attention on the features of the handwriting.

[7] Data augmentation can be used to enlarge the dataset. In image classification it involves applying various transformations (such as rotation, flipping, and scaling) to the existing training dataset to create additional diverse samples, enhancing the model's ability to generalize to different variations of the input images.

Works Cited

Bischoff 2004 Bischoff, B. (2004) *Paläographie des römischen Altertums und des abendländischen Mittelalters*, E. Schmidt, Berlin.

Brink et al 2012 Brink, A., Smit, J., Bulacu, M. and Schomaker, L. (2012) “Writer identification using directional ink-trace width measurements”, *Pattern Recognition* 45(1), 162–171. Available at: https://www.sciencedirect.com/science/article/pii/S0031320311002810.

Chammas et al 2018 Chammas, E., Mokbel, C. and Likforman-Sulem, L. (2018) “Handwriting recognition of historical documents with few labeled data”, in *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, IEEE Computer Society, Los Alamitos, CA, USA, pp. 43–48. Available at: https://doi.ieeecomputersociety.org/10.1109/DAS.2018.15.

Chammas et al 2020 Chammas, M., Makhoul, A. and DEMERJIAN, J. (2020) “Writer identification for historical handwritten documents using a single feature extraction method”, in *19th International Conference on Machine Learning and Applications (ICMLA 2020)*, Miami (on line), United States. Available at: https://hal.archives-ouvertes.fr/hal-03017586.

Christlein et al 2019 Christlein, V., Nicolaou, A., Seuret, M., Stutzmann, D. and Maier, A. (2019) “Icdar 2019 competition on image retrieval for historical handwritten documents”, in *2019 International Conference on Document Analysis and Recognition*. IEEE, pp. 1505–1509.

Cilia et al 2020a Cilia, N. D., De Stefano, C., Fontanella, F., Marrocco, C., Molinara, M. and Freca, A. S. d. (2020a), “An experimental comparison between deep learning and classical machine learning approaches for writer identification in medieval documents”, *Journal of Imaging* 6(9). Available at: https://www.mdpi.com/2313-433X/6/9/89.

Cilia et al 2020b Cilia, N., De Stefano, C., Fontanella, F., Marrocco, C., Molinara, M. and Freca, A. (2020b), “An end-to-end deep learning system for medieval writer identification”, *Pattern Recognition Letters* 129, 137–143. Available at: https://www.sciencedirect.com/science/article/pii/S0167865519303460.

Cloppet et al 2016 Cloppet, F., Eglin, V., Kieu, V. C., Stutzmann, D. and Vincent, N. (2016) “Icfhr2016 competition on the classification of medieval handwritings in latin script”, in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 590–595.

De Stefano et al 2011 De Stefano, C., Fontanella, F., Maniaci, M. and Scotto di Freca, A. (2011) “A method for scribe distinction in medieval manuscripts using page layout features”, in *Image Analysis and Processing–ICIAP 2011: 16th International Conference Proceedings*, pp. 393–402.

De Stefano et al 2018 De Stefano, C., Maniaci, M., Fontanella, F. and Scotto di Freca, A. (2018) “Layout measures for writer identification in mediaeval documents”, *Measurement* 127, 443–452. Available at: https://www.sciencedirect.com/science/article/pii/S0263224118305359.

Dey et al 2017 Dey, S., Dutta, A., Toledo, J., Ghosh, S., Llados, J. and Pal, U. (2017) “Signet: Convolutional siamese network for writer independent offline signature verification”, *CoRR* abs/1707.02131. Available at: http://arxiv.org/abs/1707.02131.

Fiel and Sablatnig 2012 Fiel, S. and Sablatnig, R. (2012) “Writer retrieval and writer identification using local features”, *Proceedings - 10th IAPR International Workshop on Document Analysis Systems, DAS 2012*.

Fiel and Sablatnig 2015 Fiel, S. and Sablatnig, R. (2015) *Writer identification and retrieval using a convolutional neural network*, in G. Azzopardi and N. Petkov, eds, *Computer Analysis of Images and Patterns*, Springer, Springer International Publishing, Cham, pp. 26–37.

Fiel et al 2017 Fiel, S., Kleber, F., Diem, M., Christlein, V., Louloudis, G., Nikos, S. and Gatos, B. (2017) “Icdar 2017 competition on historical document writer identification (historical-wi)”, in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 01, IEEE, pp. 1377– 1382. Available at: https://ieeexplore.ieee.org/abstract/document/8270156.

Haidinger 1983 Haidinger, A. (1983) *Katalog der Handschriften des Augustiner Chorherrenstiftes Klosterneuburg*, Vol. 2 of *Veröffentlichungen der Kommission für Schrift- und Buchwesen des Mittelalters*, Wien.

Haidinger 1991 Haidinger, A. (1991) *Katalog der Handschriften des Augustiner Chorherrenstiftes Klosterneuburg*, Vol. 2 of *Veröffentlichungen der Kommission für Schrift- und Buchwesen des Mittelalters*, Wien.

Haidinger 2010 Haidinger, A. (2010) “manuscripta.at – ein webportal zu mittelalterlichen handschriften in österreichischen bibliotheken”, *Schriften der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare (VÖB)* pp. 53–61. Available at: https://manuscripta.at/.

Haltich 2014 Haltich M. (2014) “Die Stiftsbibliothek”, in: *Das Stift Klosterneuburg : wo sich Himmel und Erde begegnen*. Hrsg. von Wolfgang Huber. Doessel: Janos Stekovics Verlag, 2014, p. 216–229.

He and Schomaker 2014 He, S. and Schomaker, L. (2014) “Delta-n hinge: Rotation-invariant features for writer identification”, in *2014 22nd International Conference on Pattern Recognition*, pp. 2023–2028.

Islam et al 2016 Islam, N., Islam, Z. and Noor, N. (2016) “A survey on optical character recognition system”, *ITB Journal of Information and Communication Technology*.

Kahle et al 2017 Kahle, P., Colutto, S., Hackl, G. and Mühlberger, G. (2017) “Transkribus - a service platform for transcription, recognition and retrieval of historical documents”, in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 04, pp. 19–24.

Kluge 2019 Kluge M. (2019) *Handschriften des Mittelalters: Grundwissen Kodikologie und Paläographie*. Thorbecke Jan Verlag. Available at: https://books.google.de/books?id=unFHwgEACAAJ.

Krizhevsky et al 2012 Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012) “Imagenet classification with deep convolutional neural networks”, in F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, eds, *Advances in Neural Information Processing Systems*, Vol. 25, Curran Associates, Inc. Available at: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68cPaper.pdf.

Lackner 2012 Lackner, F. (2012) *Katalog der Handschriften des Augustiner Chorherrenstiftes Klosterneuburg*, Vol. 2 of

Veröffentlichungen der Kommission für Schrift- und Buchwesen des Mittelalters, Wien.

Landau 2004 Landau P. (2004) “Die Lex Baiuvariorum. Entstehungszeit, Entstehungsort und Charakter von Bayerns ältester Rechts- und Geschichtsquelle”; *vorgetragen in der Gesamtsitzung* vom 6. Juni 2003. München. Available at: http://publikationen.badw.de/de/019366060.

Liu et al 2011 Liu, C.-L., Yin, F., Wang, D.-H. and Wang, Q.-F. (2011) “Casia online and offline chinese handwriting databases”, pp. 37 – 41.

Maaten and Postma 2005 Maaten, L. V. D. and Postma, E. (2005) “Improving automatic writer identification”, in *PROC. OF 17TH BELGIUM-NETHERLANDS CONFERENCE ON ARTIFICIAL INTELLIGENCE (BNAIC 2005*, pp. 260–266.

Marcel and Rodriguez 2010 Marcel, S. and Rodriguez, Y. (2010) “Torchvision the machine-vision package of torch”, in *Proceedings of the 18th ACM International Conference on Multimedia*, MM ’10, Association for Computing Machinery, New York, NY, USA, p. 1485–1488. Available at: https://doi.org/10.1145/1873951.1874254.

Marti and Bunke 2002 Marti, U. and Bunke, H. (2002) “The iam-database: an english sentence database for offline handwriting recognition”, *International Journal on Document Analysis and Recognition* 5(1), 39–46.

Oliveira et al 2018 Oliveira, S. A., Seguin, B. and Kaplan, F. (2018) “dhsegment: A generic deeplearning approach for document segmentation”, *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)* pp. 7–12.

Powitz 2007 Powitz G. (2007) “Was vermag Paläographie?”, in: Urkundensprachen im germanisch-romanischen Grenzgebiet: Beiträge zum Kolloquium am 5./6. Oktober 1995 in Trier, hrsg. von K. Gärtner und G. Holtus (Trierer historische Forschungen, 35), p. 223–251.

Said et al 1998 Said, H., Baker, K. and Tan, T. (1998) “Personal identification based on handwriting”, *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No.98EX170)* 2, 1761–1764.

Schneider 2014 Schneider, K. (2014) *Paläographie und Handschriftenkunde für Germanisten*, De Gruyter, Berlin/Boston.

Seidl and Haltrich 2014 Seidl, M., Haltrich, M. (2014) “Codex claustroneoburgensis-datenbank (ccl-db)”. Available at: https://phaidra.fhstp.ac.at/view/o:4631.

Shaikh et al 2020 Shaikh, M. A., Duan, T., Chauhan, M. and Srihari, S. N. (2020) “Attention based writer independent verification”, *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)* pp. 373–379.

Studer et al 2019 Studer, L., Alberti, M., Pondenkandath, V., Goktepe, P., Kolonko, T., Fischer, A., Liwicki, M. and Ingold, R. (2019) “A comprehensive study of imagenet pre-training for historical document image analysis”, *2019 International Conference on Document Analysis and Recognition (ICDAR)* pp. 720–725.

Tensmeyer et al 2017 Tensmeyer, C., Davis, B., Wigington, C., Lee, I. and Barrett, B. (2017) “Pagenet: Page boundary extraction in historical handwritten documents”, in *Proceedings of the 4th International Workshop on Historical Document Imaging and Processing*, HIP2017, Association for Computing Machinery, New York, NY, USA, p. 59–64. Available at: https://doi.org/10.1145/3151509.3151522.

Wu et al 2014 Wu, X., Tang, Y. and Bu, W. (2014) “Offline text-independent writer identification based on scale invariant feature transform”, *Information Forensics and Security, IEEE Transactions on 9*, pp. 526–536.

Xing and Qiao 2016 Xing, L. and Qiao, Y. (2016) “Deepwriter: A Multi-stream Deep CNN for Text-Independent Writer Identification”, in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, IEEE Computer Society, Los Alamitos, CA, USA, pp. 584–589. Available at: https://doi.ieeecomputersociety.org/10.1109/ICFHR.2016.0112.

Xiong et al 2015 Xiong, Y.-J., Wen, Y., Wang, P. S. P. and Lu, Y. (2015) “Text-independent writer identification using sift descriptor and contour-directional feature”, in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 91–95.

Yang et al 2016 Yang, W., Jin, L. and Liu, M. (2016) “Deepwriterid: An end-to-end online text-independent writer identification system”, *IEEE Intelligent Systems* 31(2), 45–53. Available at: https://doi.org/10.1109/MIS.2016.22.