Handwritten and printed text separation for historical documents

Bachelor Thesis

by

Anastasia Prikhodina

Degree Course: Industrial Engineering and Management B.Sc.

Institute of Applied Informatics and Formal Description Methods (AIFB)

KIT Department of Economics and Management

Reviewer:	Prof. Dr. Harald Sack
Second Reviewer:	Prof. Dr. J. Marius Zöllner
Advisors:	M.Sc. Oleksandra Bruns, M.Sc. Mahsa Vafaie
Submitted:	14 Oktober 2021

Zusammenfassung

Historische Dokumente stellen für optische Zeichenerkennungssysteme (OCR) eine große Herausforderung dar, insbesondere Dokumente von schlechter Qualität, die handschriftliche Anmerkungen, Stempel, Unterschriften und historische Schriftarten enthalten. Da die meisten OCR-Systeme entweder maschinell gedruckte oder handschriftliche Texte erkennen, müssen im Falle von gedruckten und handschriftlichen Komponenten vor dem Einsatz des jeweiligen Erkennungssystems diese getrennt werden. Diese Thesis befasst sich mit dem Problem der Segmentierung von Handschriften und Druckschriften in historischen lateinischen Textdokumenten. Um das Problem des Mangels an Daten, bei welchen handschriftliche und maschinell gedruckte Bestandteile vorkommen, die sich auf derselben Seite befinden oder sogar überlappen, sowie deren pixelweisen Annotationen zu lösen, wurde die in [12] vorgeschlagene Datensynthesemethode angewandt und neue Datensätze erzeugt. Diese neu erstellten Bilder und ihre Beschriftungen auf Pixelebene wurden zum Training des in [5] vorgestellten Fully Convolutional Model (FCN) verwendet. Das neu trainierte Modell hat bessere Ergebnisse bei der Trennung von maschinell gedrucktem und handgeschriebenem Text in historischen Dokumenten gezeigt.

Abstract

Historical documents present many challenges for Optical Character Recognition Systems (OCR), especially documents of poor quality containing handwritten annotations, stamps, signatures, and historical fonts. As most OCRs recognize either machine-printed or handwritten texts, printed and handwritten parts have to be separated before using the respective recognition system. This thesis addresses the problem of segmentation of handwritings and printings in historical Latin text documents. To alleviate the problem of lack of data containing handwritten and machine-printed components located on the same page or even overlapping each other as well as their pixel-wise annotations, the data synthesis method proposed in [12] was applied and new datasets were generated. The newly created images and their pixel-level labels were used to train Fully Convolutional Model (FCN) introduced in [5]. The newly trained model has shown better results in the separation of machine-printed and handwritten text in historical documents.

Contents

List of Figures

List of Tables

1 Introduction

In order to avoid repeating mistakes in the future, we have to be able to learn from the past. Historical records are a major source of knowledge, documenting an important part of our past and consequently, our cultural identity is hidden behind the lines of historical documents. For that reason, they help us develop a better understanding of how things work in the world.

Automatic understanding of historical documents remains an active area of research in computer science. In recent years, many libraries all around the world have published such documents, making them readily available to the general public [33, 6]. However, raw document images and scans published by online digital libraries are of the greatest benefit if there is a textual transcription available, in particular for handwritten[1] components [27]. The transcriptions serve mainly the information retrieval purpose. Thus, the invaluable information the historical documents contain must be extracted.

To make the retrieval of required information faster and to make it possible to reference documents much more easily, the documents need to be completely searchable for keywords or whole phrases. To that end, we need data to be machine-understandable. That greatly facilitates data management and reduces limitations in maintenance and accessibility [15]. As a result, the textual content of millions of scanned documents can be browsed with search masks and strings and become available to the public. In this way, with the help of contemporary tools such as various translation software programs, the historical documents are also readily accessible for non-native speakers of the language in which the documents are. Besides the accessibility in different languages, digital libraries allow multiple and simultaneous access to documents as well as access on demand. All that speeds up and simplifies information retrieval enormously.

1.1 Motivation

This thesis addresses the problem of handwritten and machine-printed text separation in the context of historical documents. There are records in historical documents containing mixed text, where text areas of handwritten and machine-printed components are very close to or even overlapping each other. Examples of such records are archival documents, manuscript materials, journals with handwritten notes in the margins, forms and tables filled out by hand, etc. However, the information retrieval, in this case, remains a challenging problem as the handwritten scripts are overly unstructured and hard to understand due to a variety of reasons including irregularities in writing, ligatures, abbreviations, historical fonts, historical spelling variants, paper degradation, displaced

[1] In this work terms “handwritten” and “hand-generated” are used interchangeably.

characters, blotches, faint texts and bleed-through from the following page [37]. To retrieve this information manually, an extremely high amount of human labor is needed, which makes the retrieving process highly expensive and time-consuming. As a matter of fact, there has been some research on the classification of handwritten and printed text, but the conventional methods in this area have some significant limitations [12]. Most of them tend to be ineffective showing low accuracy in cases where these machine-printed and handwritten parts are overlapping each other or are located on the same page. Such cases seem closest to reality and therefore are of most significance.

In fact, the stage of document segmentation is an essential pre-processing stage in an Optical Character Recognition (OCR) system, which is used to read and transcribe scanned documents. As input data, the system takes scanned images and converts them into digitized text [25]. Since most OCR systems recognize either machine-printed or handwritten text [11], documents that contain mixed text have to be segmented into printed and handwritten parts first for using the respective OCR system [7].

The contribution of this thesis is manifold. First, it includes the generation of new labeled image data containing many overlappings synthesized from handwritings and printings of Latin texts (English and German) for training. Secondly, data collection and analysis of the existing datasets are part of the contribution as well. Furthermore, it introduces new models built upon the FCN [5] for identification of printed and handwritten parts in images and trained with the newly synthesized data. Finally, the models are evaluated and compared with each other as well as with the original model [5].

This thesis is structured as follows. To begin with, some topic-specific concepts used in this work are provided in Chapter 2. Chapter 3 introduces the OCR of historical documents, and it briefly summarizes conventional methods used for the text separation and its major stages. In Chapter 4 data collection process, as well as model architecture, is demonstrated. The results of model training are shown in Chapter 5. Finally, Chapter 6 discusses these results as well as summarizes and concludes the outcome of the thesis.

2 Background and Preliminaries

In this chapter, some essential concepts related to this work are described. The following paragraphs are intended to give a brief introduction to the theoretical background terms, which will be used later in this thesis in the data pre-processing, classification, and post-processing stages.

2.1 Ground Truth

Ground truth is the reality one wants a model to predict. In other words, it is what one sets as the target value for the training and validation data. In some cases, ground truth can be wrong, if some samples are incorrectly annotated. As a matter of fact, the performance of the model is directly dependent on the quality of the ground truth used to train and test the model with [8].

2.2 Loss Function

Loss function is mainly used to correct the model's predictions and set the parameters so that the model improves itself. To put it differently, the machines learn using a loss function, as this metric shows how "far" an estimated value is from the target value: the greater this deviation, the higher the loss value. In fact, there are several most commonly used loss functions, and the choice of the proper loss function for a given problem depends essentially upon computational issues [32].

Within the scope of this work, *weighted categorical cross-entropy* is used as a loss function. The standard weighted categorical cross-entropy loss is given by:

$$J_{wcce} = -\frac{1}{M}\sum_{k=1}^{K}\sum_{m=1}^{M} w_k \times y_m^k \times \log\left(h_\theta\left(x_m, k\right)\right)$$

where

M number of training examples

K number of classes

w_k weight for class k

y_m^k target label for training example m for class k

x_m input for training example m

h_θ model with neural network weights θ [10].

Since in this thesis a distinction is made between machine-printed text, handwritten text, and background, there is a three-class classification problem. In the case of a multi-

class classification task, the categorical cross-entropy loss function can be weighted by class, increasing or decreasing the relative penalty of a probabilistic false negative for an individual class [10].

2.3 CRFs

Conditional Random Fields (CRFs) is a framework for building probabilistic models to segment and label sequence data presented in [19]. CRFs are often applied to the task of pattern recognition and sequence prediction and are widely used as a post-processing technique for image segmentation to achieve better clarity in a segmented image [3].

There are multiple types of CRF models such as Linear CRFs, Grid CRFs, Skip-Chain CRF and Dense CRFs [3]. Basically, CRF takes into account the "neighbouring" samples. As such, Linear CRFs are most commonly used for Natural Language Processing (NLP) tasks while Grid CRFs have been applied widely for pattern recognition problems.

In dense or fully connected CRFs, every node is connected to n-1 nodes, meaning all nodes in the image are connected to every other node as depicted in Figure 1, which was taken from [3]. 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 maximized by assigning pixels with similar features the same prediction [38]. As a matter of fact, with so many relationships comes the computational complexity, which implies that a lot of time is required to compute all these relationships [3].

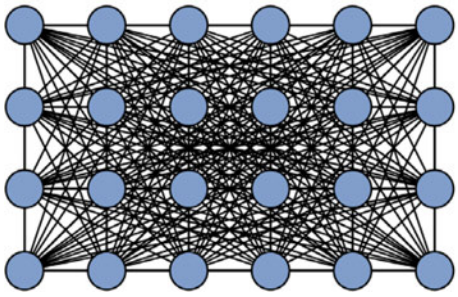


Figure 1: Illustration of fully connected CRFs [3].

3 Related Work

3.1 OCR and hOCR

While printed text recognition is considered a solved problem, converting handwriting into machine-encoded text still remains a challenging problem [4]. In [11], Islam et al. stated that based on the type of input, the OCR systems can be categorized as handwriting recognition and machine-printed character recognition. An OCR system depends mainly on the extraction of features and classification of these features, hence handwritten OCR (hOCR) is considered as a subfield of OCR [25]. That is due to the challenges that hOCR systems encounter such as a variety of individual handwritings, including different styles and alphabets, variability of strokes from person to person, inconsistent width and height of characters, rotation to the right and/or left, etc. In addition, while text in printed documents is usually located along horizontal lines, text written by hand does not necessarily sit in straight lines. Also, handwritings are often a mix of touching and not touching characters, making the distance between neighboring components variable (see Figure 2).

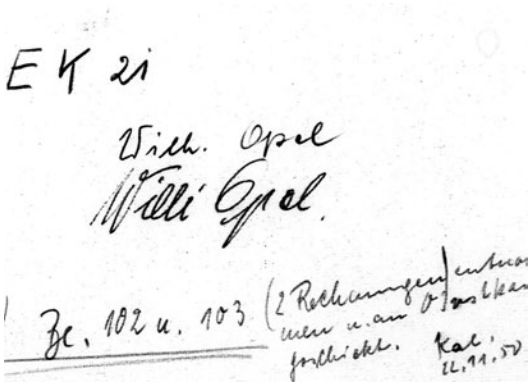


Figure 2: A crop of a handwritten region in a historical document illustrating inconsistent width and height of characters and text rotation.

Moreover, some handwritten scripts such as the ones in historical documents, which are the center of our interest, have features that make the task even more challenging. In fact, working with historical records, one often has to deal with images of poor quality due to paper degradation over time and images of uneven illumination, whilst image quality is critical for handwriting recognition [39]. Furthermore, these images have complex backgrounds arising from paper textures (e.g., curved edges near book bindings) and from paper degradation, which generates noise and increases processing time [37]. Also, cursive handwriting where neighboring characters are usually connected, makes separation and recognition of symbols harder.

OCR of historical printed documents is a challenging task as well. Similarly, as in

the case of handwritten historical records, problems such as lack of quality in scanned document images, degraded state of paper resulting in excessive noise and black spots, historical fonts and spellings still remain.

As has been said, due to different visual appearance and structural properties of machine-printed and handwritten text, separation of mixed text is required for using respective efficient OCR system [7]. Thus, before being passed on to a suitable OCR, this text has to be identified either as printed or handwritten and separated.

Character recognition in historical documents is not a new problem and has been a research topic in a number of studies. In [1], Breuel et al. used LSTM-based OCR for printed English text and historical German Fraktur text for the first time. Their approach without any language modelling yielded low error rates compared to other leading OCR engines such as OCRopus[2] and Tesseract[3]. In [37], Springmann et al. focused on recognizing historical fonts and historical spelling variants in document images of Latin texts. With their experiments for the standard systems, which are ABBYYFineReader[4], Tesseract and OCRopus, they showed that high character recognition accuracies (up to 80.57% for ABBYYFineReader, 78.77% for Tesseract and 81.66% for OCRopus) might be expected also for 16th century records [37]. OCR of historical archival documents in Finnish was explored by Drobac in [4]. The main problem of such records is that they are mainly written in Gothic (Fraktur) font, which is harder to recognize as Fraktur documents contain many touching characters and ligatures [4]. A character accuracy of 93.50% with OCRopus software as well as accuracy of over 94% with additional post-processing was achieved in [4]. In [2], Chammas et al. presented a deep convolutional recurrent neural network (CRNN) system for handwriting recognition in multilingual Latin historical documents. In fact, most hOCR systems work at the line-level by transforming the text-line image into a sequence of feature vectors. The novelty of the proposed method is that it also works with a small amount of manually segmented and labeled text-lines where there are only transcriptions at the paragraph-level available with no text-line information.

3.2 Synthesis of Ground Truth

As mentioned before, separation of text regions is required for using a respective OCR system. For that purpose, text components have to be identified first (e.g. as *handwritten* or *printed*) by a classifier. These annotations are called ground truth - the target value for the training and test data. In the context of this work, ground truth gives information on whether a component is known to be handwritten or printed in reality. In fact, ground

[2]Available at: *https://github.com/ocropus/ocropy*
[3]Available at: *https://github.com/tesseract-ocr*
[4]Available at: *https://www.abbyy.com/*

truth creation is crucial for all steps of an automatic document image processing pipeline such as layout analysis, text recognition, word spotting, writer identification, as well as document understanding [8]. To accelerate the ground-truthing process, Fischer et al. proposed a semi-automatic ground truth creation method for recognition of handwritten components in historical documents [6]. Once features are extracted, Fischer et al. suggested to train HMM in order to find optimal word boundaries for a given transcription. On the other hand, a proposed method in [8] for creating ground truth for historical manuscripts is based on document graphs and a pen-based scribbling interaction. In [26], Nafchi et al. presented the PhaseGT tool, which also aims to reduce the manual ground-truthing effort. In cases where a dataset provides XML data containing information such as bounding boxes of every single word, line, or even a paragraph, ground truth can be created using this metadata. In this manner, Dutly et al.[5] created ground truth for *IAM* dataset. Moreover, ground truth can be created with specific software. In recent years, several on-line available tools for ground truth creation such as VoTT by Microsoft[5] and Amazon SageMaker Ground Truth by Amazon[6] have been published.

3.3 Binarization

Before distinguishing between handwritten and printed components in images, pre-processing steps such as binarization can be applied to enhance the quality of the images [11]. As Tensmeyer et al. mentioned in [39], binarization is a critical pre-processing step in many applications and helps facilitate other document processing tasks such as character recognition. As a matter of fact, the quality of the binarization can significantly affect system performance [39]. Indeed, there is less noise in binary images, since binarization can serve as a noise removal process, which increases document readability [39]. Moreover, binary images take less disk space, which is critical when working with big datasets of images.

In general, historical document images are more difficult to binarize than modern scanned documents [39] because of degraded condition and since because many historical documents are digitized with cameras, some of which produced images have uneven illumination due to bad lighting or because the page is not fat (e.g., curved edges near bookbindings). In addition, typical features of historical document images such as complex backgrounds, stains, creases, border noise, faint text, and multiple text colors only add to the complexity of historical document binarization [39]. In fact, there is no single binarization algorithm that performs equally well on every image [20]. Therefore, each binarization method aims to address some particular problem.

The classical thresholding algorithms include *Otsu* [29], *Niblack* [28], *Sauvola* [34]

[5] Available at: *https://github.com/Microsoft/VoTT/*

[6] Available at: *https://aws.amazon.com/de/sagemaker/groundtruth/*

and *Wolf* [40] binarizations [39]. In *Otsu*'s global thresholding, which remains popular as Tensmeyer and Martinez claim in [39], a value of the threshold is not chosen but is determined automatically. Otsu's approach effectively handles images with uniform background and has no parameters to tune, but often fails in the presence of images with non-uniform background.

In contrast, the other three methods fall under the category of a local threshold considering threshold parameters over small regions [36]. *Niblack* is a simple local adaptive threshold based on window mean and standard deviation [39]. To solve the problem with background-only windows of *Niblack*, the method was improved to *Sauvola* binarization, whereas *Wolf* is an extension of *Sauvola* with global normalization [39].

3.4 Classification

As previously stated, text in images has to be categorized (e.g. *printed* or *handwritten*) before being passed on to an OCR system [11]. The text classification problem can be addressed using various techniques that aim at differentiating between handwritten and printed text. Traditional approaches in this field include algorithms based on Hidden Markov Models (HMMs). In [33] Sánchez et al. wrote that handwritten text recognition borrows concepts and methods from the field of Automatic Speech Recognition (ASR), since to some extent, the transcription of handwritten text images is comparable with the task of recognizing continuous speech in an audio file. Because there were efficient techniques for line detection, existing methods for transcription of handwritten text also worked on a line-level [33]. To discriminate hand-generated and machine-printed annotations in document images, Guo et al. [9] applied HMM that segments text on a word-level. In [30], Peng et al. used G-means based classification and a Markov Random Field (MRF) relabeling procedure. Thus, the whole document modeled as an MRF is divided into three classes: printed, handwritten, and overlapped texts. Subsequently, Peng et al. applied an MRF-based classification approach to assign the overlapping cases to the remaining classes (machine-printed and hand-generated) using pixel-level separation.

In [35], Shetty et al. proposed to use Conditional Random Fields (CRFs) in document segmentation and compared their approach to other methods such as Naive Bayes and Neural Networks. A two-level classification approach using Support Vector Machine (SVM) classifier for text localization in documents is described in [13]. Similarly, SVM was applied in [7], where Garlapati et al. described an approach to classify machine-printed and handwritten components at world-level by their visual impression and structural features. In the proposed method the model was trained and tested on IAM[7] dataset since it contains both components.

[7] Available at: *https://fki.tic.heia-fr.ch/databases/iam-handwriting-database*

Recent solutions for the image segmentation task on a pixel-level imply methods based on deep neural networks [18, 12, 31, 22, 5]. Yielding pixel-wise labels, these methods allow the distinction between handwritten and machine-printed components in case of overlappings. Especially Convolutional Neural Networks(CNNs) are being used extensively for the recognition of optical characters. This is partially due to the availability of large datasets and partially due to the fact that CNN-based architectures are well suited for recognition tasks where input is an image [25]. In [12], Jo et al. propose the first pixel-level separation method based on end-to-end learning of a CNN. 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 is introduced. To solve the problem of the lack of an appropriate dataset, the authors develop a data synthesis method that provides pixel-level annotations and many overlappings.

In [5], Dutly et al. introduced a lightweight Fully Convolutional Network (FCN) model, which solves the problem of printed and handwritten text identification on a pixel-level by combining the model with a Conditional Random Field (CRF) for post-processing. As Dutly et al. illustrated in [5], the number of parameters used by the lightweight architecture, which is inspired by the FCN-8 architecture [22], is significantly smaller than the U-Net - another well known fully convolutional architecture [31] also used for the image segmentation task. In [5], Dutly et al. used a pixel-based classification method as classification methods based on other separation levels such as word-, line-, region-based proved to be ineffective at distinguishing handwritten text which overlaps printed parts.

3.5 Post-processing

After classification, post-processing can be performed to achieve better clarity in segmented images [3] and therefore to improve the accuracy of OCR results [11]. For this purpose, Hidden Markov Models (HMMs) have been well understood and widely applied as a post-correction technique in image segmentation [19, 9]. In the last few years, Conditional Random Fields (CRFs) are often used for image segmentation problems. [21, 17, 3]. Alternatively, graphical models like Markov Random Fields (MRFs) are popular in the process of image segmentation as a post-processing technique [14].

Dense CRFs can be applied for processing segmented images to get higher accuracy images [3] and therefore for attaining better clarity in segmented images [17]. Krähenbühl and Koltun have demonstrated in [17] that dense pixel-level connectivity considerably improves segmentation and labeling accuracy in segmented images and thus leads to significantly more accurate pixel-level classification performance.

4 Data Collection and Model

As noted above, due to historical fonts, historical spellings, paper degradation state, image quality, faded ink, and overlappings (see Figure 3), historical records are especially complicated documents for OCR.

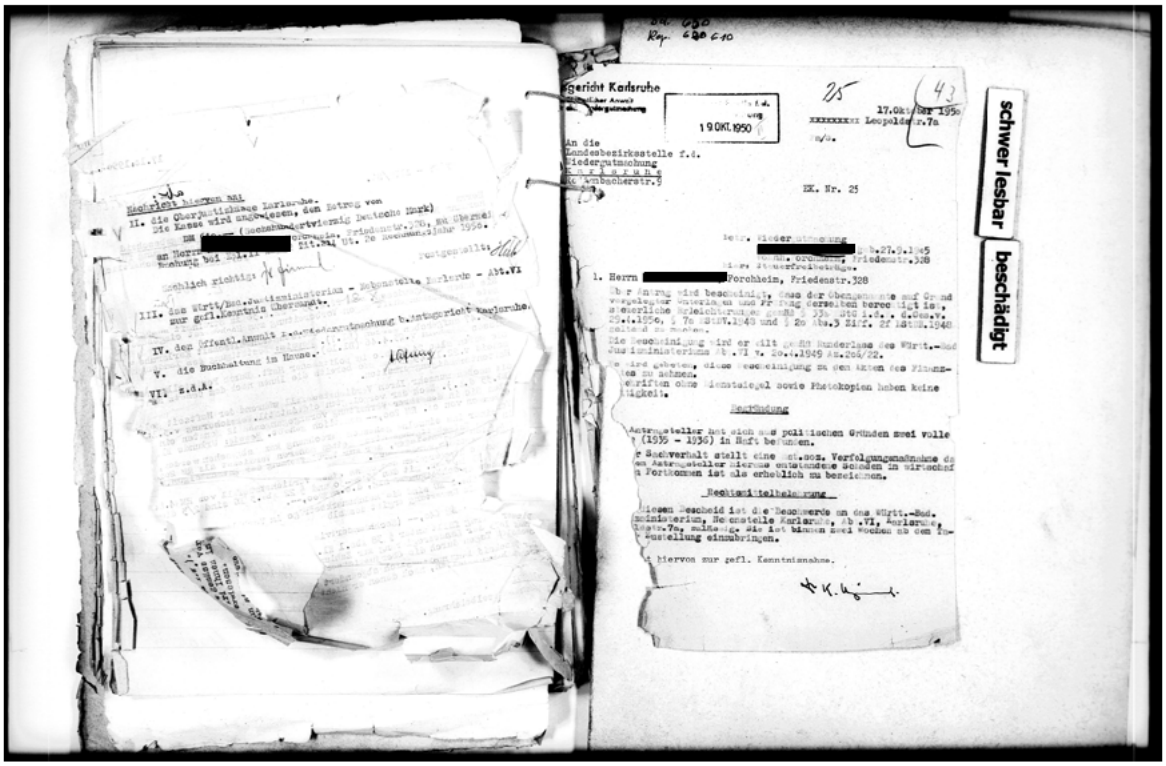


Figure 3: Illustration of challenging aspects of historical documents: poor quality of page image, overlappings, faded ink, and degraded spots (*wgm* dataset).

Since the focus of this thesis lies on cases where machine-printed and handwritten components are overlapping each other, data with similar features and properties is considered. Having crops only with printed or handwritten texts, a new dataset that serves our purpose can be created by simply overlaying these two components. In this chapter, the existing labeled datasets with mixed fonts are collected and reviewed against their suitability for the handwritten and printed text separation task in noisy historical documents. Additionally, new datasets that fulfill the training goals are generated.

4.1 *IAM* Database

The *IAM* is a handwritten database of the English language. In total, 657 writers contributed samples of their handwriting to the dataset, which consists of 1539 forms of English text with corresponding XML metadata. Besides, *IAM* contains isolated text lines and word images. The Figure 4 below illustrates an *IAM* form, an isolated text line

and some isolated words.

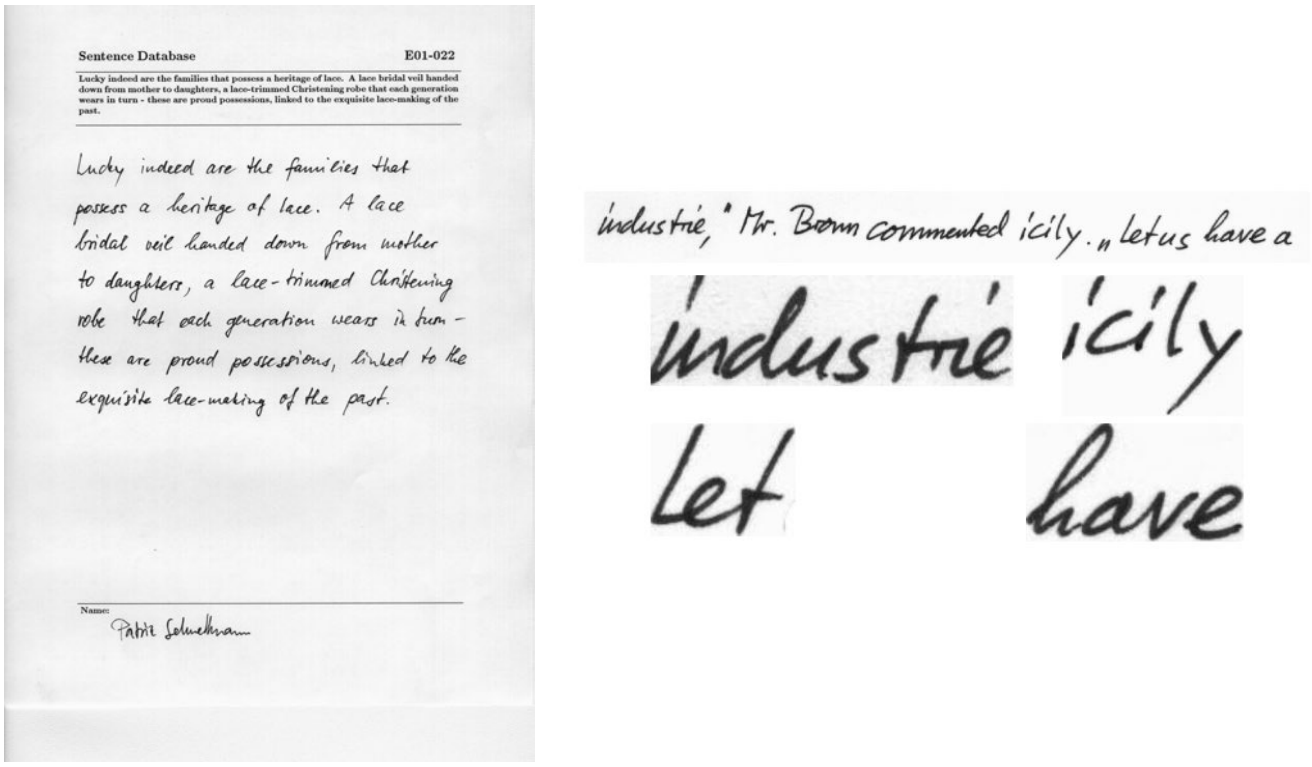


Figure 4: Samples of an *IAM* form, a text line, and some extracted words.

IAM dataset is one of a few datasets, which provides XML metadata that can be used to create pixel-level annotations of images or ground truth. Using this XML data, pixel-level annotated masks containing information on whether the text is handwritten or machine-printed can be created. For segmentation of *IAM* forms, in [5] Dutly et al. used fixed y values. In practice, segmentation of printed and handwritten parts using these fixed coordinates was not successful over all *IAM* forms. There were some images, where a few text lines were misclassified above or under the separation line between printed and handwritten parts, since a position of the separation line is not fixed (see Figure 5).

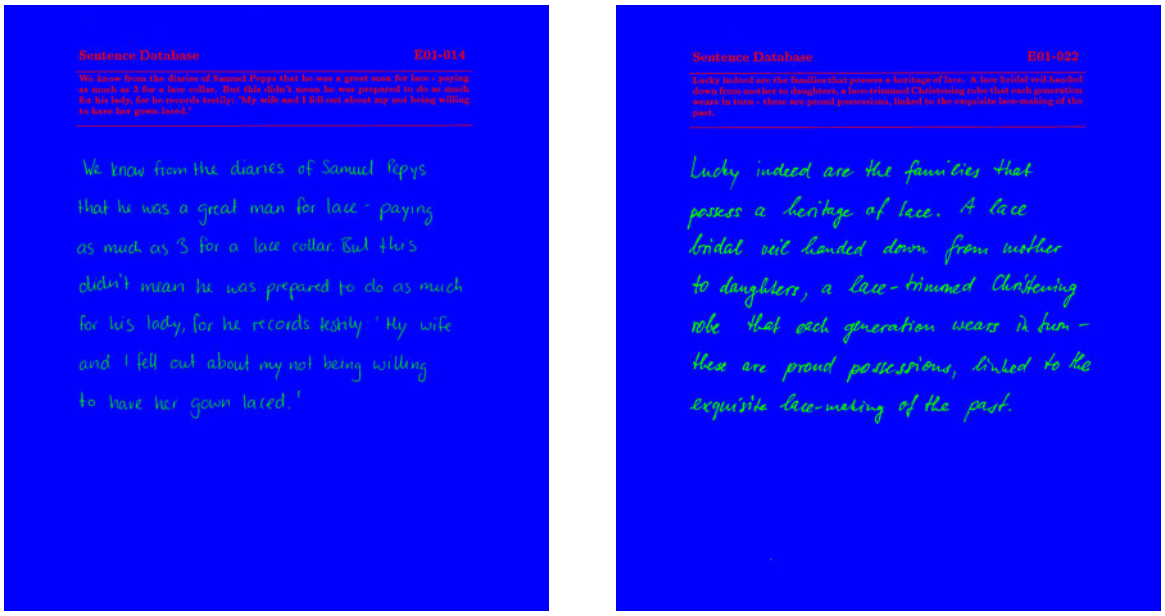
Thus, the *y-lower* value was adjusted as well as the *y-upper* value was expanded as in some forms the lower part of the handwritten text was truncated as illustrated in Figure 7. Since there is a footer containing printed and handwritten regions after the main handwritten part in the *IAM* forms, the *y-upper* value was expanded to the footer. The experimentally determined *y-upper* value is 2785 instead of 2215. As the printed line in the footer is of the same font and size as the other printed part in the top, it is not of much use and can be omitted. The same also applies to the handwritten line in the footer. The corresponding script[8] and the pixel-wise annotations[9] of the *IAM* can be found in

[8]https://github.com/anaprikho/printed-hw-segmentation/blob/dev/dataset_generation/iam2segmentation/iam2segmentation.py

[9]https://github.com/anaprikho/printed-hw-segmentation/tree/master/datasets/IAM

Figure 5: Some failed segmentation results of handwritten and printed regions by creating pixel-wise annotations of *IAM* forms.

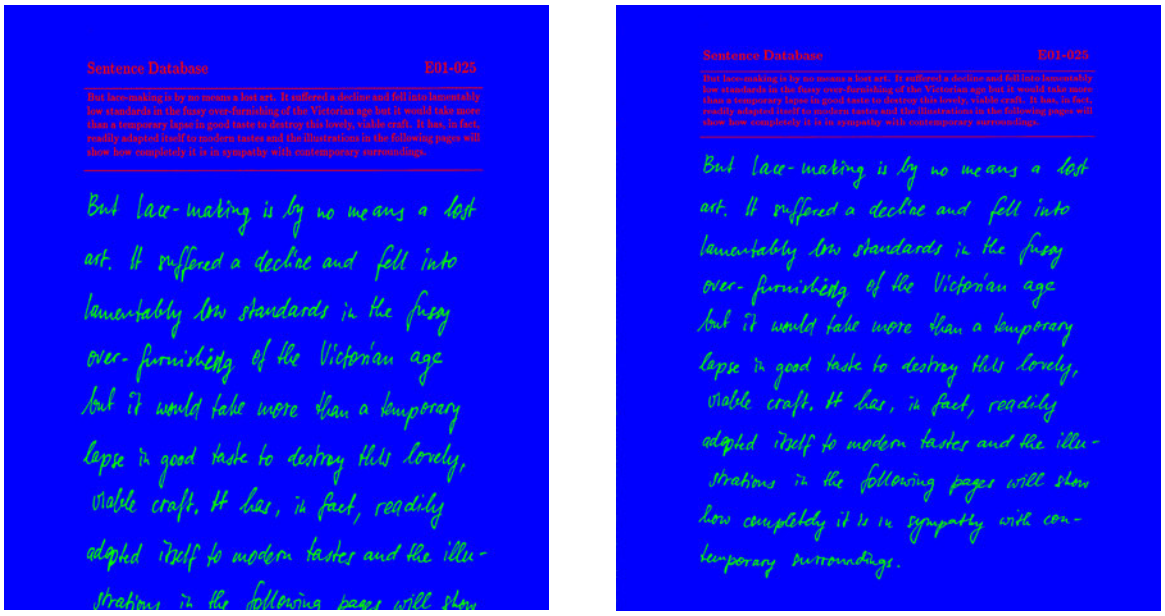
the git repository.



(a) y-lo and y-up fixed

(b) y-lo and y-up fixed

Figure 6: Some segmentation results of handwritten and printed regions after adjusting *y-upper* and *y-lower* values by creating pixel-wise annotations of *IAM* forms.



(a) y-lo = 2215 (b) y-lo = 2785

Figure 7: A truncated lower part of a handwritten region in an *IAM* form.

4.2 *CVL* and *jottueset* Databases

To generate a new dataset with the help of the data synthesis method we need two types of input data: scripts written by hand and machine-printed scripts. For that purpose, the *CVL*[10] and the *jottueset*[11] databases were collected.

4.2.1 *CVL* Database

Kleber et al. presented the *CVL* database in [16]. It contains handwritten lines, handwritten words, and whole pages of mixed script. The CVL database consists of document images with cursive handwritten German and English text, which has been chosen from literary works. Within the context of the data synthesis method, large text blocks such as paragraphs rather than single words or lines are of most interest.

In fact, there is a couple of machine-printed sentences on top of each page followed by a handwritten part, which completely replicate the printed text above. The printed text block is placed between two horizontal separation lines. 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, since in total 310 writers participated in the dataset, 27 of which wrote 7 texts and 283 writers had to write 5 texts. Compared to the machine-printed sentences, these texts show the diversities of handwriting. The

[10] https://cvl.tuwien.ac.at/research/cvl-databases/an-off-line-database-for-writer-retrieval-writer-identification-and-word-spotting

[11] https://drive.google.com/file/d/1Q4kDiJts-yi9IhsYT6ku5Y4WNhwagnPJ/view

ink color and its intensity also vary from author to author. Actually, for each page there is a cropped version only with handwritten text available. Because of this, there is no need to retrieve the metadata stored in the XML files to find out information on text locations to crop the handwritten parts out. In other words, no pre-processing of this data is necessary.

In this manner, 1604 crops containing only handwritten text were collected, which can be subsequently used for generation of a new dataset using the data synthesis method proposed in [12].

4.2.2 *jottueset* Database

To generate a new dataset using the data synthesis method [12], handwritten and machine-printed crops are required. Since there are already *CVL* handwritten crops collected that can be used for data synthesis, machine-printed crops are needed. For this purpose, the *jottueset* database can be used. This dataset includes 141 scanned in black and white questionnaire documents, which mainly contain forms and tables. It is important to emphasize that the questionnaires are not filled in, which means there are only machine-printed components of a few different print fonts, some of them are also bold or italic. As a matter of fact, the *jottueset* dataset is characterized by relatively low diversity compared to the *CVL*.

Since there is no ground truth readily available for these scanned documents, their pixel-level annotations were created[12]. The Figure 8 illustrates an annotated sample of the *jottueset* dataset, where red and blue denote machine-printed components and background, respectively.

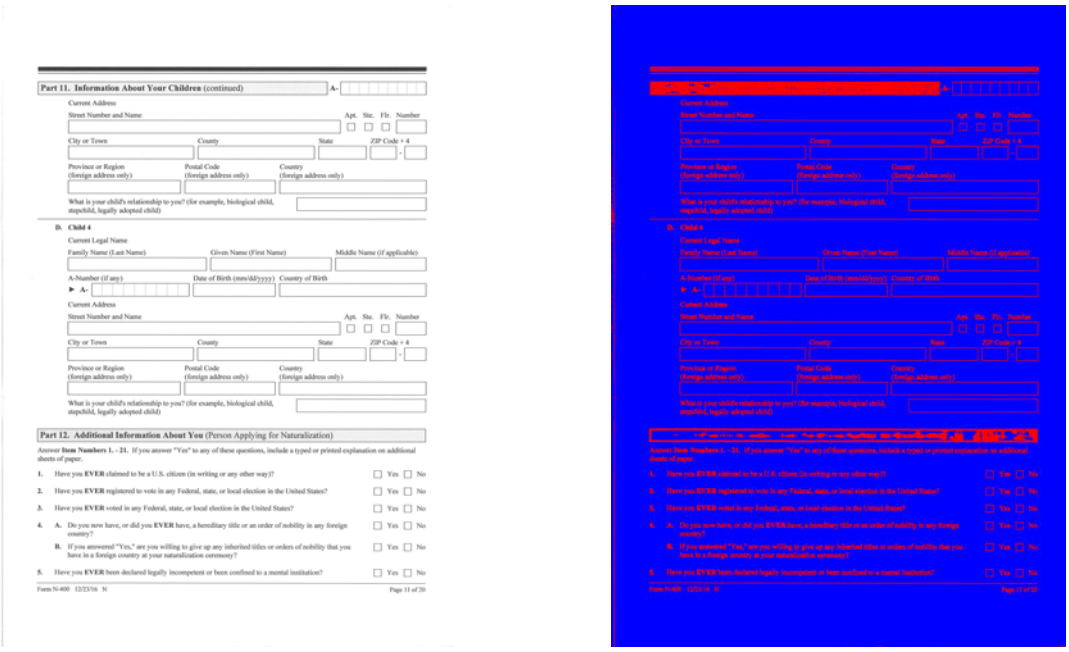
In total, 141 pixel-wise annotations[13] for *jottueset* images containing only machine-printed parts were generated, which can be subsequently used for data synthesis.

4.2.3 Data Synthesis

As already mentioned before, there is a lack of datasets containing both handwritten and machine-printed components overlapping each other. Most existing datasets, however, do not provide pixel-level annotations (ground truth), which are required for training and accuracy calculation of segmentation results performed by a model later. As a solution, a new dataset using a data synthesis method introduced in [12] can be created by practically overlaying handwritten crops on printed components. The Figure 10 illustrates the data

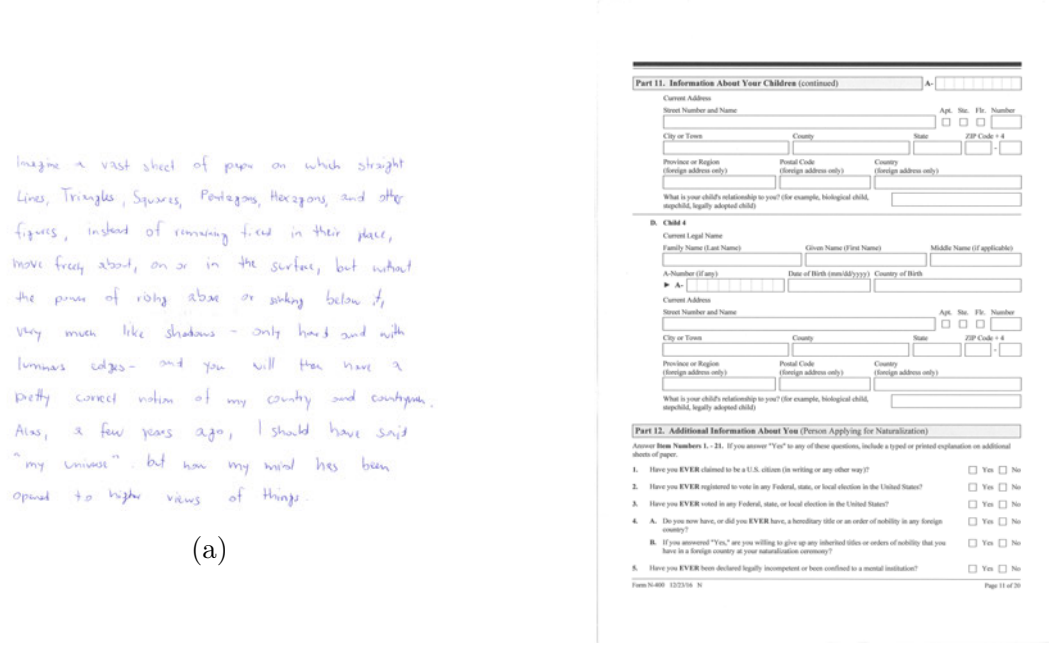
[12]https://github.com/anaprikho/printed-hw-segmentation/blob/dev/dataset_generation /printedsci2segmentation/gen_data.py

[13]https://github.com/anaprikho/printed-hw-segmentation/tree/master/datasets/jottueset_gt



(a) (b)

Figure 8: **(a)**: an original sample of the *jottueset* data; **(b)**: its pixel-level annotation.



(a) (b)

Figure 9: Examples of samples used for the data synthesis: **(a)**: a cropped handwritten region of an *CVL* page; **(b)**: a sample from the *jottueset* database.

synthesis process. In order to reflect the diversity of real documents, some randomized transformations such as resizing and translation were applied to handwritten components.

This method allows to generate realistic pixel-level annotated training samples hav-

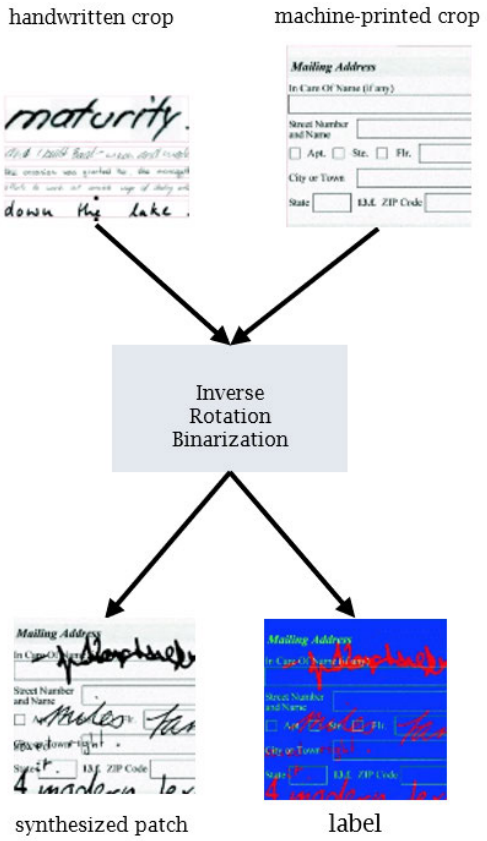


Figure 10: Pipeline of the data synthesis method.

ing many overlappings of printed and handwritten characters. However, the authors argued that simply overlaying of handwritten and machine-printed components is not enough as such approach leads to undesirable synthesizing results such as block artifacts (illustrated in [12]). For more realistic images Jo et al. multiplied images with their binary masks to extract only handwritten pixels.

To synthesize data containing many overlappings, this data generation method was applied to the *CVL* crops and the *jottueset* scanned questionnaire documents (see Figure 9). In doing so, new patches and their corresponding pixel-level annotations were created. However, after executing the method with unchanged parameters, there were many newly created patches containing an excessive number of handwritten components merging into whole bunches of "ink" spots. Thus, some parameters were adjusted[14], since keeping them unchanged led to unsatisfied results. The number of sentences was reduced from 100 to 4 as well as an array of scales were extended from [0.7, 1., 1.5] to [0.7, 1., 1,5, 2., 2.3, 2.7, 3.5]. Regarding the image size, it was set to 256*256 pixels. A list of the parameters used for the generation of new patches is shown in Table 1. Blue, red, and green denote background, machine-printed text, and handwritten text pixels, respectively.

[14]https://github.com/anaprikho/HTSNet/blob/master/data_generation.py

Yellow are overlapping areas. As a result, altogether 16,548 images were synthesized[15]. Table 2 represents an overview of the numbers of samples in each database as well as of newly synthesized patches. Examples of some data synthesis results are represented in Figure 11, where the first row shows synthesized patches and the second row indicates corresponding pixel-level annotations.

	cvl-jottueset
NUM_SENTENCE	4
SCALES	[0.7, 1., 1.5, 2., 2.3, 2.7, 3.5]
PATCH_SIZE	256
STRIDE_SCALE	1.5
MAX_ROTATION	0

Table 1: List of parameters used in the data synthesizing method with the *CVL* handwritten crops and *jottueset* printed documents.

	handwritten	machine-printed
cvl handwritten crops	1604	-
jottueset forms	-	141
synthesized patches (cvl-jottueset)	**16548**	

Table 2: Overview of the number of crops in *CVL* and *jottueset* databases and of the synthesized patches.

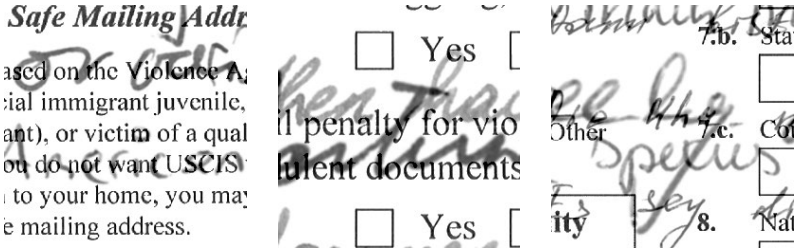
4.3 *wgm* Database

4.3.1 *"Wiedergutmachung"* Project

Another data we collected is a subset of the data provided by the Landesarchiv Baden-Württemberg (LABW) within the scope of the *"Wiedergutmachung"*[16] (*wgm* for short) project. In fact, the aim of the project is to present the history of reparations in Germany as an essential aspect of Germany's postwar and democratic period more clearly than it has been done so far. Thus, the digitization and subsequent enrichment of corresponding catalog data with the help of machine-generated metadata will be tested as part of the *"Wiedergutmachung"* project. This implies the creation of a digital library system by bringing and storing the digitized associate knowledge together. Consequently, the centralized access to the historical information can be made possible in the Archive Portal-D,

[15] https://github.com/anaprikho/printed-hw-segmentation/tree/master/datasets/cvl_jottueset

[16] https://www.fiz-karlsruhe.de/de/forschung/wiedergutmachung
https://www.landesarchiv-bw.de/de/landesarchiv/projekte/projekt-zur-wiedergutmachung/71002



(a) patches

(b) labels

Figure 11: Some results of data synthesis using *CVL* handwritten crops and images of *jottueset* printed documents : **(a)** synthesized patches **(b)** corresponding pixel-level annotations of the patches.

an online information system that allows sector-specific access to the archival information and digitized archives (such as documents and photographs) of the German Digital Library. Without a doubt, this project will make a vital contribution to the research of the changes in German society after 1945.

wgm	number of images
Microfilm	153
color photots	150

Table 3: Number of images in the Microfilm and color photos subsets (*wgm*).

The collected *wgm* data consists of document scans including archival records such as forms filled in by hand, typewritten certificates, testimonies, and declarations dating back to the 20th century and contains plenty of personal information such as names, birth dates, places of residence, occupation, etc. In fact, there are two types of media representation of archival records in the *wgm* dataset: Microfilm and color photos. As shown in Table 3, there are 153 images in the Microfilm and 150 images in the color photos subset in total. The Microfilm includes black and white scans of typewritten documents with a small number of handwritten notes in the main text as well as in the margins, underlinings, checkmarks, digits, and signatures (see Figure 12). In contrast to the Microfilm, there are many typewritten forms filled in by hand in the second subset and scans are in full color (see Figure 13).

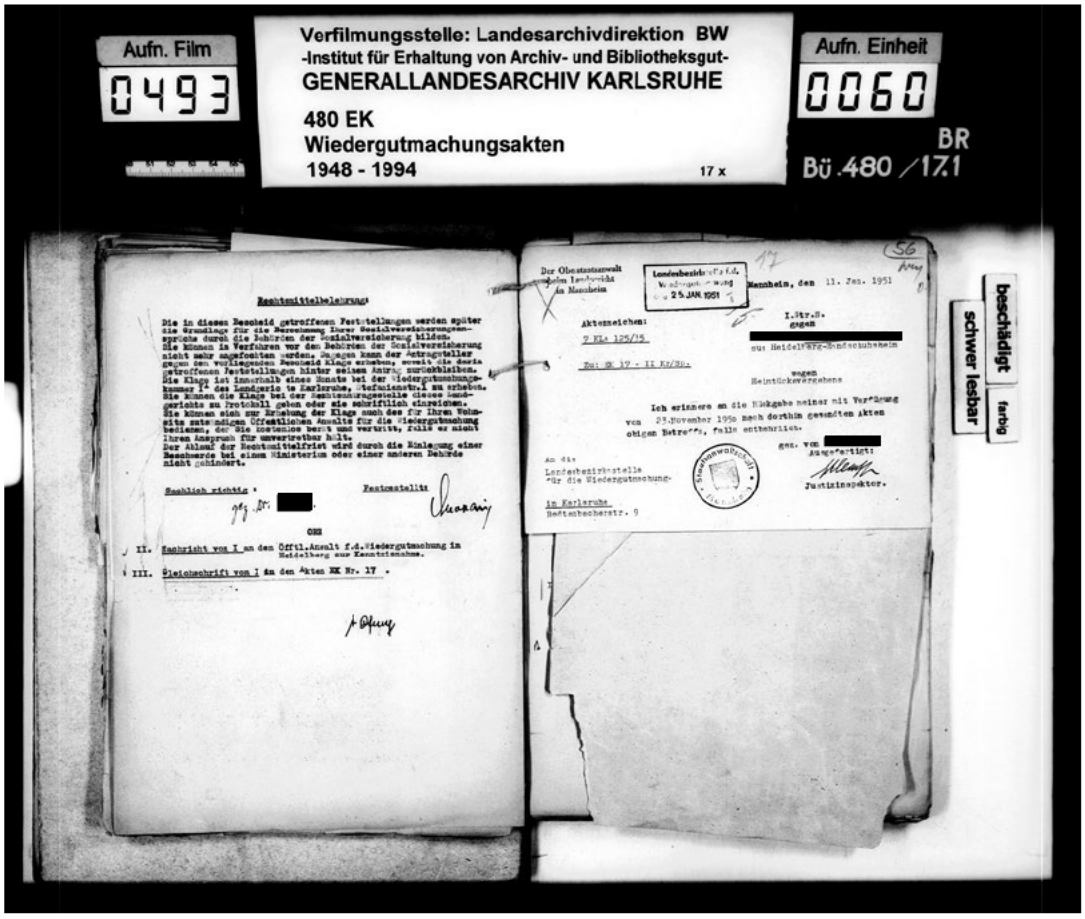


Figure 12: A sample of the Microfilm subset (*wgm*).

4.3.2 Data Labeling

Since the goal of the thesis is to separate a printed text from a handwritten text in challenging historical documents (such as scans of historical records represented in Figures 12 and 13), the creation of a labeled dataset similar to the real documents is required for model training and testing. For the generation of realistic pixel-annotated samples containing printed and handwritten text and overlaps, the previously mentioned data synthesis method [12] was applied to this data as well.

Since there are multiple examples of old-style typewriter and handwriting fonts as well as old German spelling in the *wgm* documents, use can be made of both printed and hand-generated parts of these documents. To extract the areas of hand- and typewritings from the document scans and create crops from them, the VoTT annotation and labeling tool for images was used. Images from a local file on a computer can be used as input data. After loading the images, tags such as *signature* (for signatures), *handwritten* (for handwritings), *machineprinted* (for printed text) and *mixed* (for overlappings and text regions where handwritten and printed components are close to each other) were create. Then, the images were labeled appropriately. Labeled images can be exported as JSON

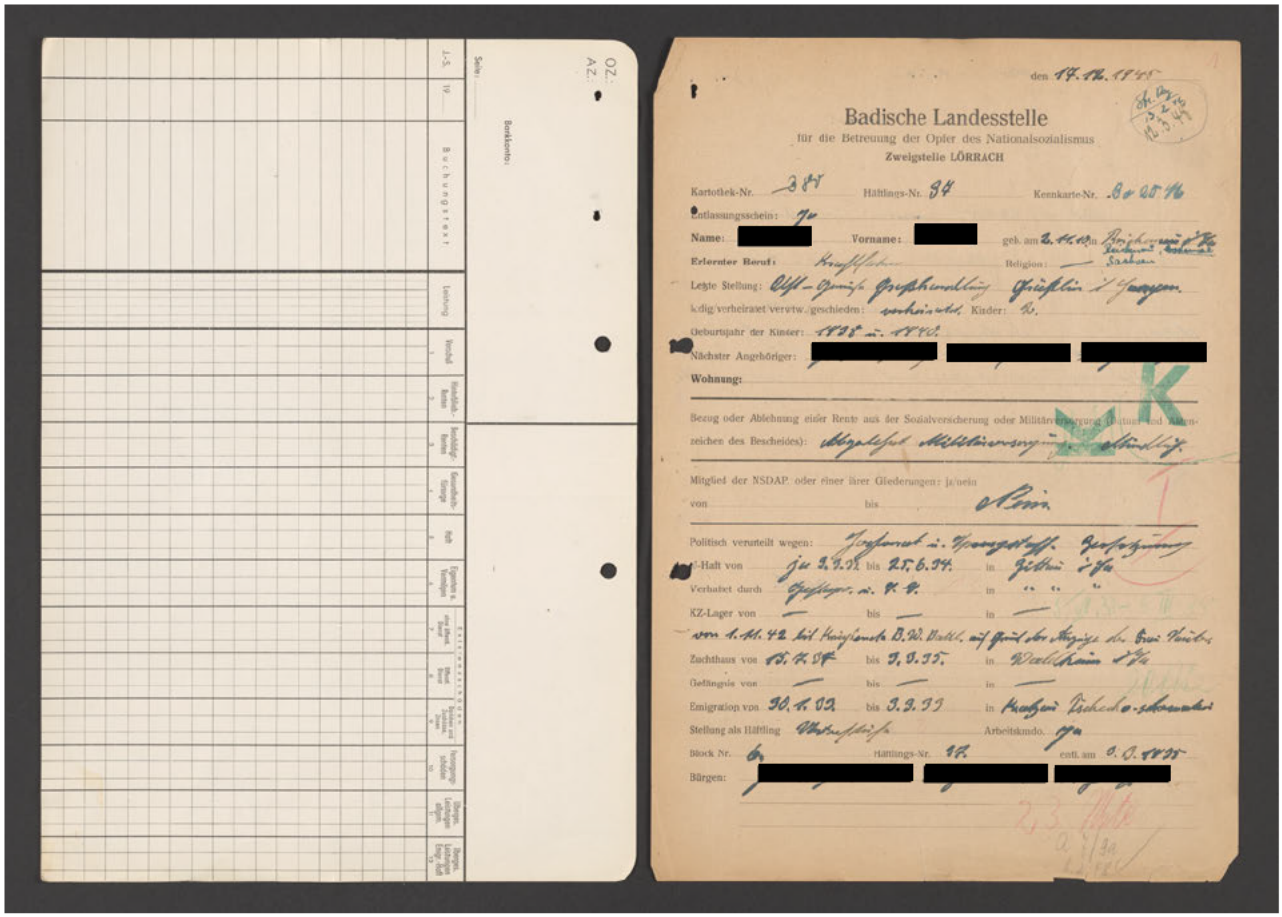


Figure 13: A sample of the color photos subset (*wgm*).

files[17], which can be easily read when creating synthetic samples.

As tiny regions and single words make the labeling process rather tedious and do not add much value to the input, for annotation a preference was expressed for homogeneous text regions like paragraphs. Indeed, great importance was placed on forms, tables, headlines, different font sizes, different stroke intensities, and rotation angles, as all that characteristics add to the diversity of the data, and therefore are considered particularly valuable. Printed, handwritten and mixed parts were labeled using a rectangular box. For annotation of stamps and signatures polygon was applied to avoid the inclusion of other text components in case they were too close. Otherwise, a regular rectangular box was used.

4.3.3 Crops Generation

Once regions in the document scans were assigned to respective label tags, there is a possibility to save the output to a JSON object as mentioned before. The JSON file[18]

[17]https://github.com/anaprikho/printed-hw-segmentation/tree/master/datasets/wgm/VoTT-output
[18]https://github.com/anaprikho/printed-hw-segmentation/tree/master/datasets/wgm/VoTT-output

provides all the information about tags there are in each image as well as coordinates of the image regions that were assigned to these labels. Having the coordinates and label tags' name of each annotated region, one can simply crop the region out of the original image and save it[19]. In total, out of 153 images of Microfilm subset 53 crops of handwritings and 44 crops of typewritings were generated[20]. As for the subset of colorful scans containing 150 images, there are 139 crops of handwritings and 296 crops of typewritings in total[21]. As can be seen, there is a significant difference in numbers of crops created from Microfilm and from color photos despite the almost equal number of the original images in both subsets. The reason for this is provided in the following paragraphs.

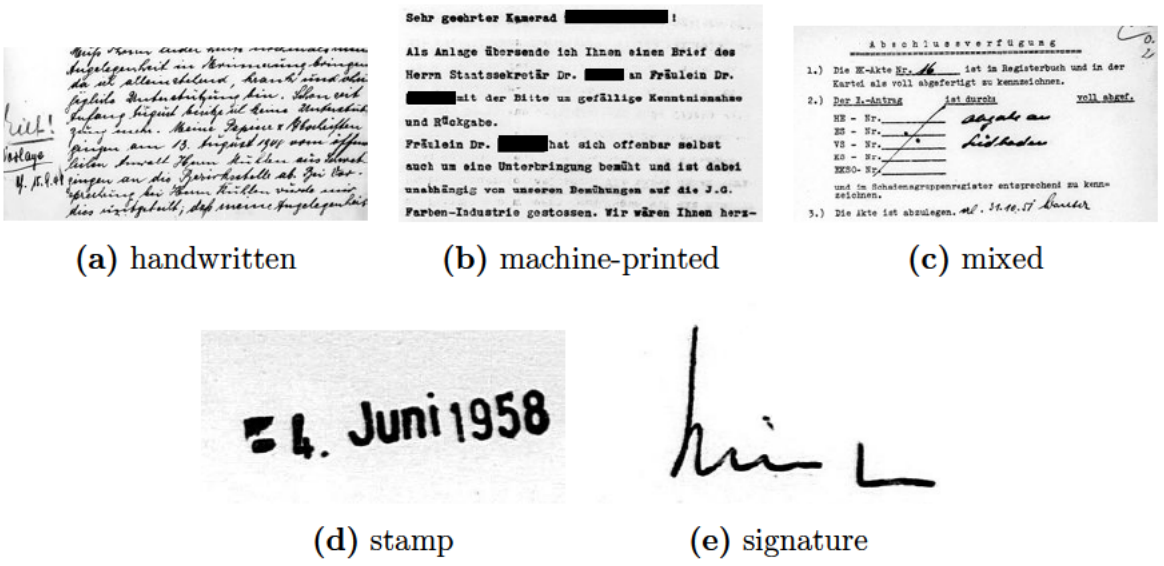
In the case of the Microfilm subset, there are many printed documents containing texts written in the same font. Obviously, annotating more regions there will be more crops, but crops of texts having the same characteristics (such as font and font size) do not add more value to data. Since the quality of data is more important than quantity, these similar regions of text in the document scans were not considered for annotation. Moreover, since the Microfilm documents were already converted into black and white by the scanning process, they have more noise. Many of them also contain foreign objects such as punch holes or an archivist's hand. Within the framework of the thesis, such noisy regions were skipped for better clarity in synthesized data, since noisy images can affect system performance [39].

In contrast, the number of crops derived from the subset of the color photos is several times as high as from the Microfilm data. The reason for this is that these two subsets include significantly different data. As a matter of fact, the color photos contain considerably less noise and therefore, no regions should be avoided by annotation. In addition, this data shows more diversity. Apart from the various types of official typewritten letters and certificates, there are plenty of filled-out forms and tables. Not only different kinds of documents but also a lot of different typewriting and handwriting family fonts, font sizes, ink colors, ink intensity, weights, and shapes. As a result, there are considerably more crops obtained with the second subset than with Microfilm. In Figures 14, 15 some samples of the crops generated from JSON data after annotating Microfilm and color photos subsets respectively can be seen.

[19] https://github.com/anaprikho/printed-hw-segmentation/blob/dev/dataset_generation/json2crops.py

[20] https://github.com/anaprikho/printed-hw-segmentation/tree/master/datasets/wgm/VoTT-output/Microfilm/.../crops

[21] https://github.com/anaprikho/printed-hw-segmentation/tree/master/datasets/wgm/VoTT-output/color_photos/.../crops



(a) handwritten (b) machine-printed (c) mixed

(d) stamp (e) signature

Figure 14: Samples of crops generated using annotations of the Microfilm subset (*wgm*).

4.3.4 Data Synthesis

As already mentioned before, most datasets containing overlappings do not provide pixel-wise annotations [5] required for model training and validation. As a solution, a new dataset using a data synthesis method introduced in [12] (see Figure 10) was generated by using handwritten and machine-printed crops of document images from both subsets (Microfilm and color photos)[22].

Since within the scope of the work green was used in pixel-level annotations for handwritten components and red for machine-printed components, some changes on the original script[23] were made, where Jo et al. used red for denoting the machine-printed and green for hand-generated parts. The overlapping areas are yellow as before. The patch size was set to 256x256 pixels as well as a rotation angle was set to 0 degrees. The rest of the parameters such as the number of sentences, scales, and stride of scales remained unchanged. The list of parameters used can be found in Table 4.

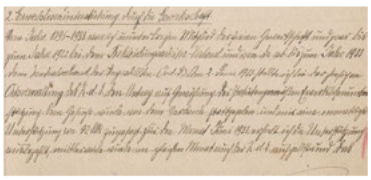
As for the Microfilm images, since they are already in black and white, the binarization step is optional. Some results of data synthesis using handwritten and printed crops of Microfilm are illustrated in Figure 16, where there are synthesized patches in the first row and the corresponding ground truth in the second row. Altogether, there are 1850 patches[24] synthesized from the Microfilm subset.

Regarding the color scans, applying the synthesis method, there are some patches

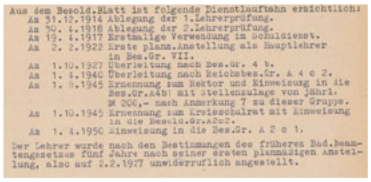
[22] https://github.com/anaprikho/HTSNet/blob/master/data_generation.py

[23] https://github.com/jottue/HTSNet

[24] https://github.com/anaprikho/printed-hw-segmentation/tree/master/datasets/wgm/wgm_Microfilm/synthesized_patches

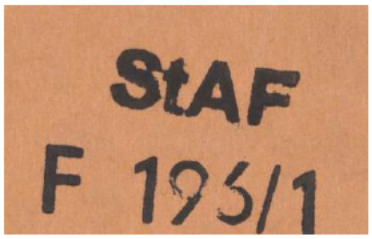


(a) handwritten



(b) machine-printed

(c) mixed



(d) stamp

(e) signature

Figure 15: Samples of crops generated using annotations of the color photos subset (*wgm*).

	wgm	
	Microfilm	color photos
NUM_SENTENCE	100	
SCALES	[0.7, 1., 1.5]	
PATCH_SIZE	256	
STRIDE_SCALE	1.5	
MAX_ROTATION	0	

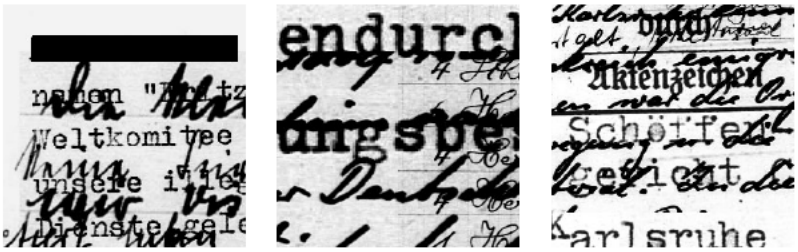
Table 4: List of parameters used in the data synthesizing method with the *wgm* crops.

where the background of hand-generated components was labeled incorrectly as handwritten (see Figure 17). The issue could be resolved by firstly binarizing the input crops with an *Otsu* binarization method with changed parameters (different than in [12])[25] and then using these crops as input data for the synthesis method. In this case, there is no need to binarize the data again while synthesizing. Some results of data synthesis using handwritten and printed crops of the color photos subset are illustrated in Figure 18. The first row shows synthesized patches and the second row indicates corresponding pixel-level annotations. In total, 734 patches[26] from the fragments cropped out of the color photos were generated.

Finally, using the data synthesis method, 2584 realistic samples and their pixel-level annotations were generated from the crops of the Microfilm and color document

[25] https://github.com/anaprikho/HTSNet/blob/master/binim.py

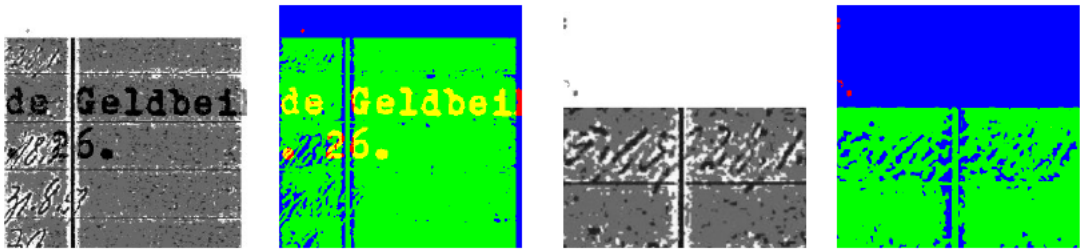
[26] https://github.com/anaprikho/printed-hw-segmentation/tree/master/datasets/wgm/wgm_color_photos/synthesized_patches



(a) patches

(b) labels

Figure 16: Some results of data synthesis using handwritten and machine-printed crops of Microfilm scans (*wgm*): **(a)** synthesized patches **(b)** pixel-level annotations of synthesized patches where the background was misclassified as handwriting.



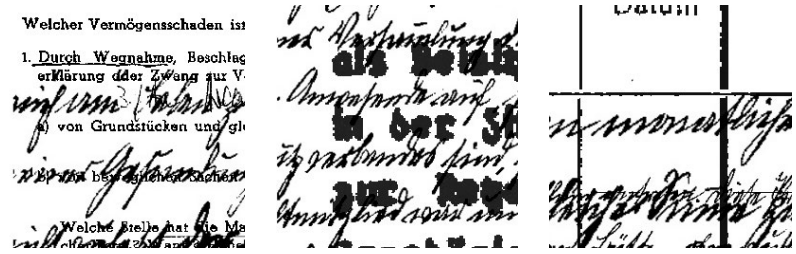
(a) **(b)** **(c)** **(d)**

Figure 17: Some failed results of data synthesis using handwritten and machine-printed crops of color scans (*wgm*): **(a),(c)**: synthesized patches **(b),(d)**: pixel-level annotations of synthesized patches.

scans[27]. A summary of numbers of synthesized images using the subsets of the *wgm* data is illustrated in Table 5.

As can be seen from Figures 16 and 18, the patches synthesized from the crops of the color scans seem to be less noisy than patches from Microfilm fragments. As a matter of fact, data quality is crucial for model training as it can affect system performance [39]. Since this noisy data is similar to realistic samples in the context of image segmentation of historical documents, it can be used for model training, validation, and testing. Also, less noisy patches generated of color photos crops can be of use bringing thereby diversity

[27]https://github.com/anaprikho/printed-hw-segmentation/tree/master/datasets/wgm/wgm



(a) patches

(b) labels

Figure 18: Some results of data synthesis using binarized handwritten and machine-printed crops of color photos subset (*wgm*): **(a)** synthesized patches **(b)** pixel-level annotations of synthesized patches.

	wgm	
	Microfilm	color photos
handwritten crops	53	44
machine-printed crops	139	296
synthesized patches	**1850**	**734**

Table 5: Overview of the number of crops in Microfilm and color photos subsets and of the synthesized patches using them (*wgm*).

to the data.

To summarize this chapter, 2584 and 16548 images were synthesized from the *wgm* and *cvl-jottueset* crops, respectively. A detailed overview of the collected data can be seen in Table 6.

	wgm		cvl-jottueset
	Microfilm	color photos	
handwritten crops	53	44	1604
machine-printed crops	139	296	141
synthesized patches	1850	734	**16548**
	2584		

Table 6: Overview of the collected data.

4.4 Model Architecture

4.4.1 FCN-light

In [5] Dutly et al. addressed the problem of printed and handwritten text recognition on a pixel-level. In addition, they aimed to deploy their model as a web service enabling in such way easy integration into existing workflows. This implies that the model should take up minimal disk space as well as its complexity should be preferably low, since complex models are computationally expensive. Because of the model's complexity and pixel-wise classification, a lightweight, pixel-based fully convolutional architecture (FCN-light) based on the FCN-8 architecture [22] was used, which due to the small disk footprint and low complexity is well suited for subsequent deployment as a web service [5]. A schematic overview of the lightweight architecture can be found in Figure 19 presented in [5].

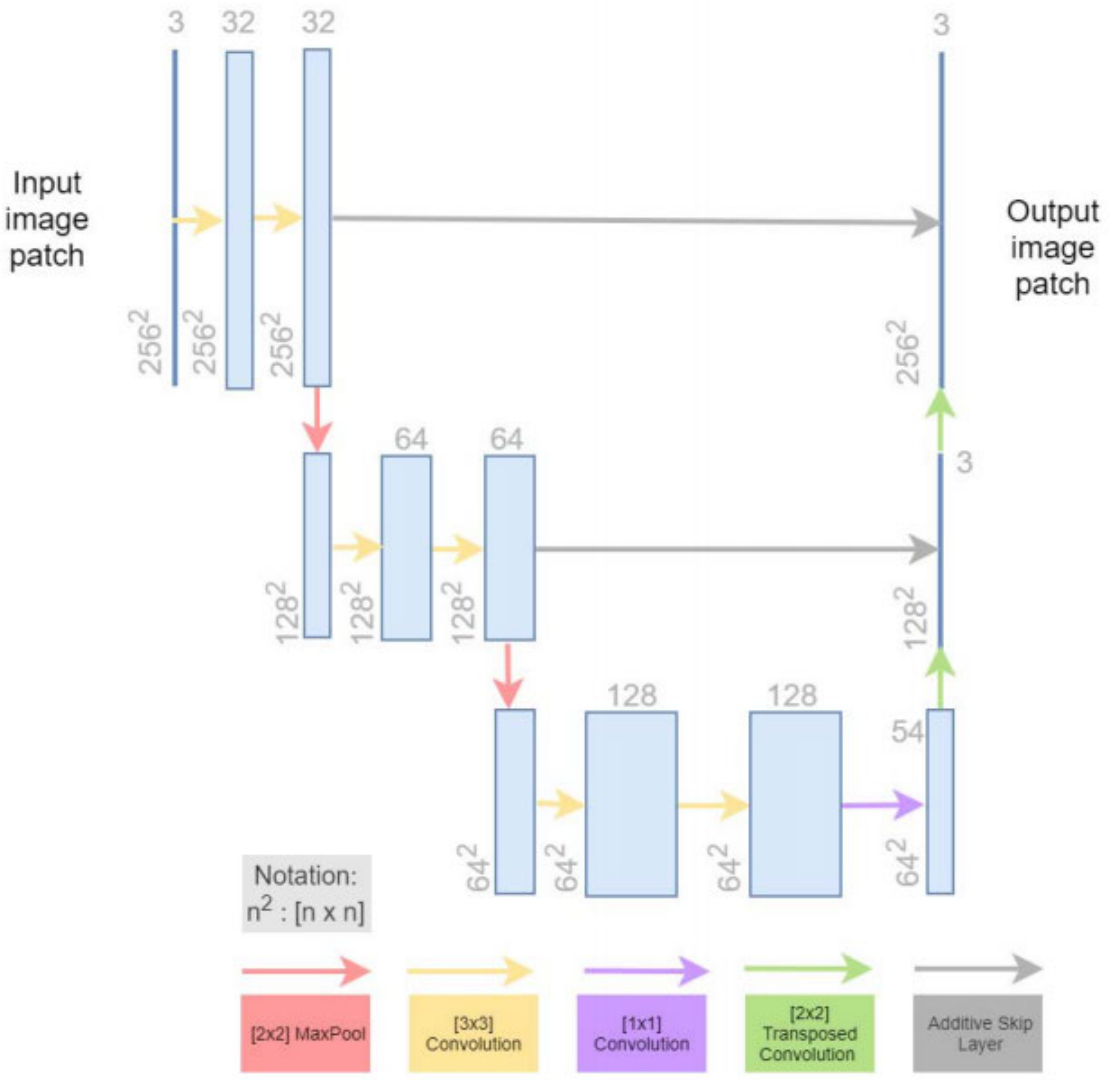


Figure 19: A lightweight variation of the FCN-8 model (the ReLu layers have been omitted from the diagram for clarity) [5].

4.4.2 CRF Post-processing

Since the FCN-light tends to misclassify small clusters of pixels [5], Dutly et al. [5] proposed to combine this lightweight, pixel-based model with a Conditional Random Field (CRF) for postprocessing, which increases the quality of predictions of the FCN-light. Indeed, CRFs can correct the labeling mistakes in some cases and consequently enhance the IoU scores [17]. Obviously, the end results of the post-processing step are highly dependent on the segmentation results of the FCN model in the first place.

5 Experiments

As mentioned before, the existing FCN-light model Dutly et al. proposed in [5] often fails in presence of mixed scripts and overlapping cases. To alleviate the problem, new feature vectors using previously synthesized data with *wgm* and *cvl_jottueset* datasets and their pixel-wise annotated labels were created (discussion in greater details below). The feature vectors contain the information about the color components of a pixel in an image. Subsequently, using the existing FCN lightweight architecture, new models was trained and evaluated using these newly created feature vectors. For model training, the same hyper-parameters as in [5] were used. The only exception was the number of epochs, which was limited to 15 instead of 50. To divide data into training, validation, and testing sets, a split ratio of 80/20/20 was chosen. To evaluate the models the IoU (Intersection over Union) metric introduced in the next section was used.

5.1 IoU

IoU is the most popular performance evaluation metric used in segmentation, object detection and tracking [23].

$$IoU = \frac{|A \cap B|}{|A \cup B|} = \frac{|I|}{|U|}$$

Where A and B are the prediction and ground truth bounding boxes. It is used to determine true positives (TP) and false positives (FP) in a set of predictions, where *positive* and *negative* are two arbitrary classes (e.g. "spam" or "not spam"). Therefore, TP is an outcome where our model was able to predict correctly the *positive* class while FP - is an outcome where our model incorrectly identified records as *positive*. Figure 20 Mahdi et al. used in [24] demonstrates more visibly what the IoU score measures.

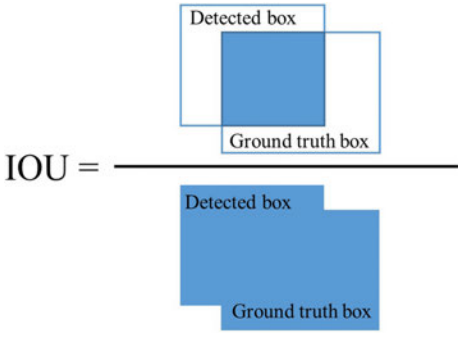


Figure 20: Illustration of Intersection over Union (IoU) [24].

	wgm	cvl_jottueset	wgm-cvl_ jottueset
training	2068	13238	15306
validation	258	1655	1913
testing	258	1655	1913

Table 7: Split ratios for *wgm*, *cvl_jottueset* and *wgm-cvl_ jottueset* datasets.

	fcnn_wgm	fcnn_cvl_jottueset	fcnn_wgm-cvl_jottueset
loss	0,029	0,028	0,031
validation loss	0,027	0,029	0,031
IoU	0,682	0,879	0,842
validation IoU	0,691	0,875	0,802

Table 8: Training history of the following models: *fcnn_ wgm*, *fcnn_ cvl_jottueset* and *fcnn_ wgm-cvl_jottueset*.

	cvl_jottueset_subset	wgm-cvl_jottueset_subset
training	2068	4136
validation	258	516
testing	258	516

Table 9: Split ratios for *cvl_jottueset_ subset* and *wgm-cvl_jottueset_ subset*.

	fcnn_cvl_jottueset_subset	fcnn_wgm-cvl _jottueset _subset
loss	0,033	0,035
validation loss	0,038	0,037
IoU	0,859	0,754
validation IoU	0,868	0,756

Table 10: Training history of the following models: *fcnn_ cvl_jottueset_ subset* and *fcnn_ wgm-cvl _jottueset _ subset*.

5.2 Experiment 1: cvl_jottueset

The first model was created using *cvl_jottueset* data. After creating new feature vectors with synthesized *cvl_jottueset* images and their ground truth, a model called *fcnn_ cvl_jottueset*[28] was trained with 13238 *cvl_jottueset* images and validated on 1655 images. Split ratios for the *cvl_jottueset* dataset can be seen in Table 7. The model training metrics such as loss, validation loss, IoU, and validation IoU values can be found in Table 8.

Due to the high number of images in the *cvl_jottueset* dataset (15306 images in the

[28]https://github.com/anaprikho/printed-hw-segmentation/blob/master/FCN/models/fcnn_cvl_jottueset.h5

training set), *fcnn_cvl_jottueset* model evaluation was interrupted due to the memory error (nevertheless, the trained model is still available online in open access). As a solution, a subset with the same number of crops as *wgm* datset contains (2584) was created. With this subset called *cvl_jottueset_subset*[29] a model *fcnn_cvl_jottueset_subset*[30] was trained (see the split ratios in Table 9). Its training metrics can be seen in Table 10. Evaluation results achieved by this model are represented in Table 11. While on average there is no deterioration in the performance after applying CRFs, this model shows better results on classification of printed components without post-processing. Like *fcnn_wgm*, this model tends to misclassify printed components as handwritten components especially in overlappings (see Figure 21). In order to test this model on real data, some Microfilm and color photos samples were used, and the segmentation results the model achieved can be seen in Figures 22 and 23. As is shown in these illustrations, the model indeed often fails to identify printed text especially in the presence of old fonts (in Fig. 22 **(c), (d), (f)**; in Fig. 23 **(a), (b)**). Actually, that is to be expected, since the *fcnn_cvl_jottueset_subset* model was trained only with modern printed documents of the *jottueset* questionnaires. In contrast, the model distinguishes between handwritten and printed components as shown in Fig. 22 **(a), (b), (e)**; in Fig. 23 **(c)**.

	FCN-light	FCN-light + CRF post-processing
Mean printed IoU	0,75	0,72
Mean handwritten IoU	0,67	0,71
Mean background IoU	0,96	0,98
Total mean IoU	0,79	0,80

Table 11: Evaluation results achieved by *fcnn_cvl_jottueset_subset* model.

5.3 Experiment 2: wgm

In the same way, another model called *fcnn_wgm*[31] was trained with and validated on *wgm* data (see Table 7). The model training history is shown in Table 8. Evaluation results achieved by the model *fcnn_wgm* are represented in Table 12. The reason for the lower mean IoU value after applying CRF post-processing can be a great number of overlaps in the synthesized crops of *wgm* data. As can be seen in Figure 24, which illustrates model segmentation results on crops synthesized from the *wgm* data, machine-printed text located close to and/or overlapped on handwritings tends to be identified

[29] https://github.com/anaprikho/printed-hw-segmentation/tree/master/datasets/cvl_jottueset_subset

[30] https://github.com/anaprikho/printed-hw-segmentation/blob/master/FCN/models/fcnn_cvl_jottueset_subset.h5

[31] https://github.com/anaprikho/printed-hw-segmentation/blob/master/FCN/models/fcnn_wgm.h5

(a) before CRF post-processing

(b) after CRF post-processing

Figure 21: Some evaluation results achieved by the *fcnn_cvl_jottueset_subset* model on crops synthesized from *cvl_jottueset* images.

as handwritten text after applying post-processing. However, the synthesized crops of *wgm* dataset are not quite realistic, since the number of overlapping is huge. To see how the model performs on real historical documents, some images from Microfilm and color photos subsets were used (see Figures 25 and 26).

	FCN-light	FCN-light + CRF post-processing
Mean printed IoU	0,55	0,35
Mean handwritten IoU	0,55	0,44
Mean background IoU	0,88	0,89
Total mean IoU	0,66	0,56

Table 12: Evaluation results achieved by *fcnn_wgm* model.

5.4 Experiment 3: wgm-cvl_jottueset

There are models which were separately trained with *wgm* and *cvl_jottueset* documents. To enhance diversity of the train data and thus to create an improved model, these two datasets were brought together into one called *wgm-cvl_jottueset*[32]. Using this com-

[32]https://github.com/anaprikho/printed-hw-segmentation/tree/master/datasets/wgm-cvl_jottueset

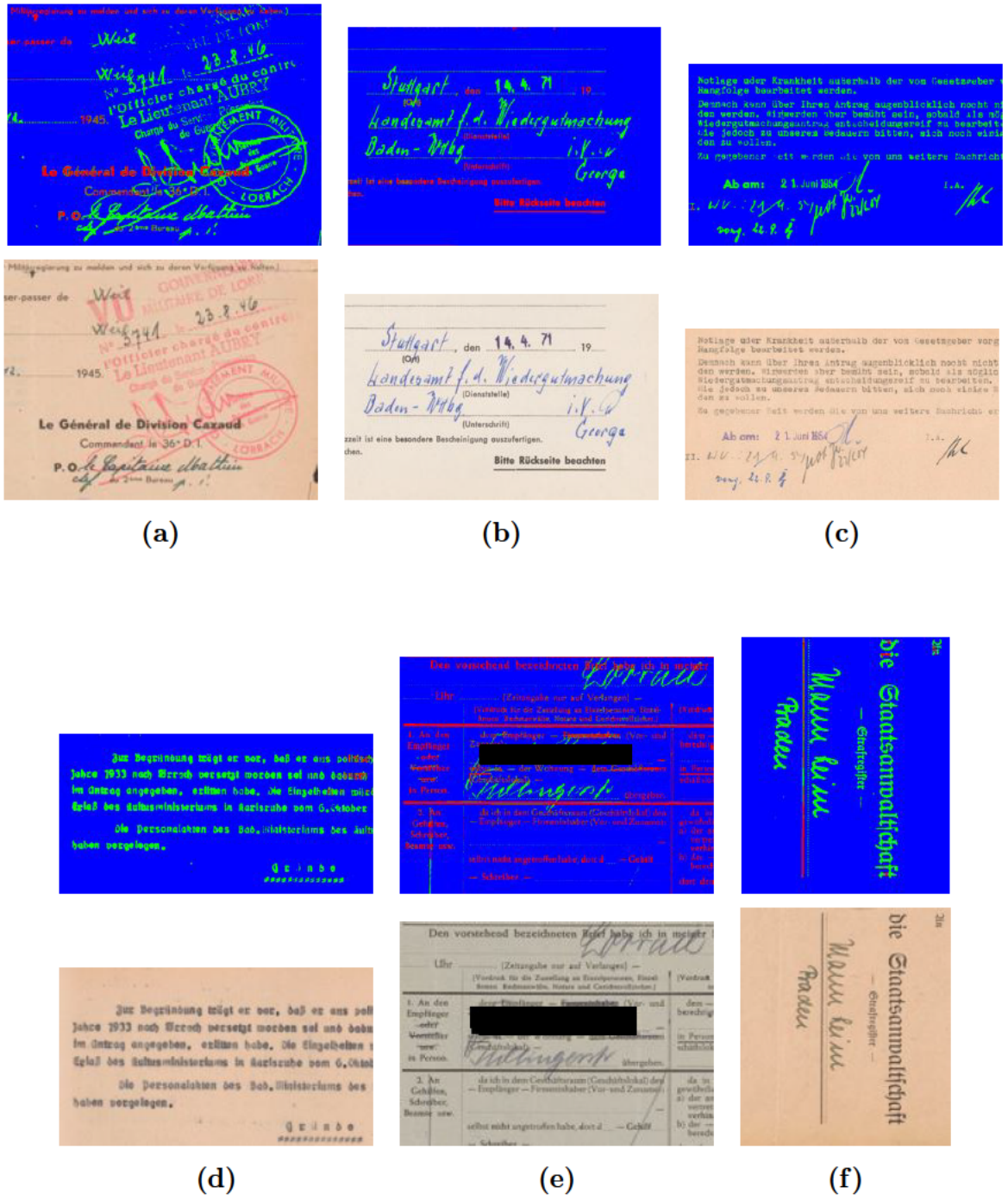


Figure 22: Some results achieved by the *fcnn_cvl_jottueset_subset* model on real historical documents of the color photos subset (*wgm*).

(a) (b) (c)

Figure 23: Some results achieved by the *fcnn_cvl_jottueset_subset* model on real historical documents of the Microfilm subset (*wgm*).

(a) before CRF post-processing

(b) after CRF post-processing

Figure 24: Some evaluation results achieved by the *fcnn_wgm* model on crops synthesized from *wgm* documents.

posit dataset, a model called *fcnn_wgm-cvl_jottueset*[33] was trained on 15306 images (see Tables 7 and 8). As in the case of the *fcnn_cvl_jottueset* model, evaluation of the

[33]https://github.com/anaprikho/printed-hw-segmentation/blob/master/FCN/models/fcnn_wgm-cvl_jottueset.h5

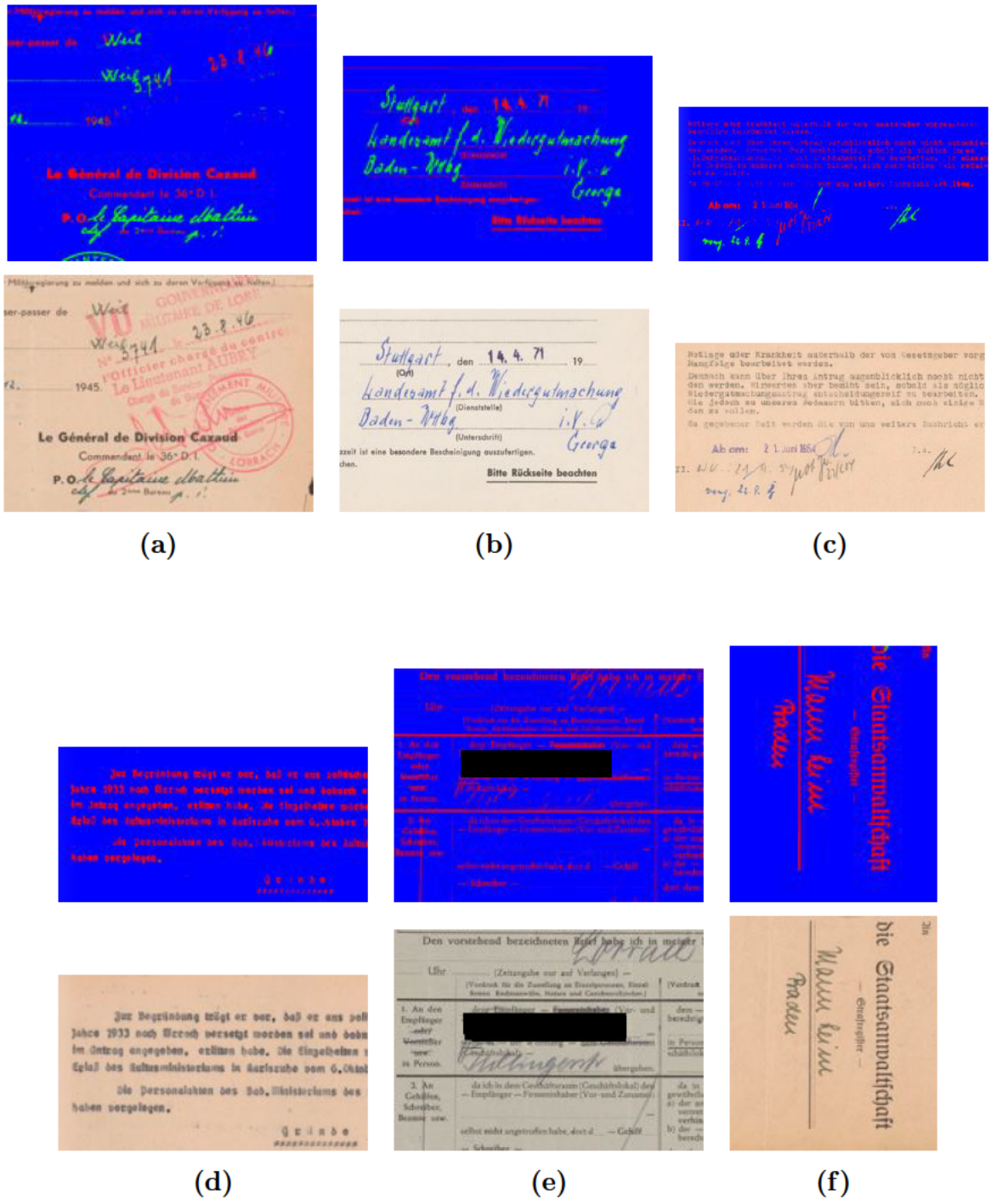
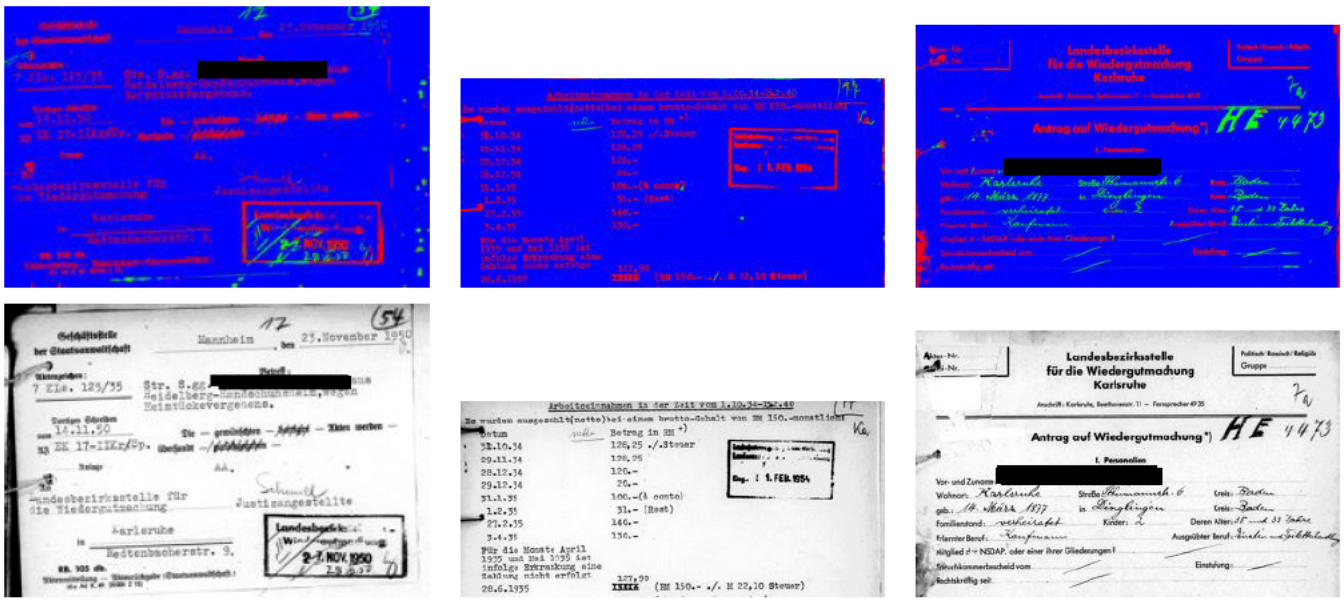


Figure 25: Some results achieved by the *fcnn_wgm* model on real historical documents of the color photos subset (*wgm*).



(a) **(b)** **(c)**

Figure 26: Some results achieved by the *fcnn_wgm* model on real historical documents of the Microfilm subset (*wgm*).

fcnn_wgm-cvl_jottueset model was also not possible due to the disk space demanded by such high number of crops in the test set. As a compromise, instead of *wgm-cvl_jottueset* data, a dataset[34] consisting from *wgm* and the newly created *cvl_jottueset_subset* was used to create a model called *fcnn_wgm-cvl_jottueset_subset*[35]. Its evaluation results can be seen in Table 13 showing that also here post-processing has no significant impact on the segmentation results. In Figure 27 some segmentation results on synthesized patches before and after applying CRF are illustrated. Using real historical documents of *wgm* data, the model yields results that can be found in Figures 28 and 29. This model shows improvement in identifying printed components as can be seen in Fig. 28 **(c)**, **(d)**; in Fig. 29 **(b)**, but at the same time misclassifies characters written in pencil (Fig. 28 **(e)**, **(f)**).

[34] https://github.com/anaprikho/printed-hw-segmentation/tree/master/datasets/wgm-cvl_jottueset_subset

[35] https://github.com/anaprikho/printed-hw-segmentation/blob/master/FCN/models/fcnn_wgm-cvl_jottueset_subset.h5

	FCN-light	FCN-light + CRF post-processing
Mean printed IoU	0,55	0,46
Mean handwritten IoU	0,50	0,53
Mean background IoU	0,89	0,93
Total mean IoU	0,65	0,64

Table 13: Evaluation results achieved by *fcnn_wgm-cvl_jottueset_subset* model.

(a) before CRF post-processing

(b) after CRF post-processing

Figure 27: Some evaluation results achieved by the *fcnn_wgm-cvl_jottueset_subset* model on crops synthesized from *wgm* and crops synthesized from *cvl_jottueset* images.

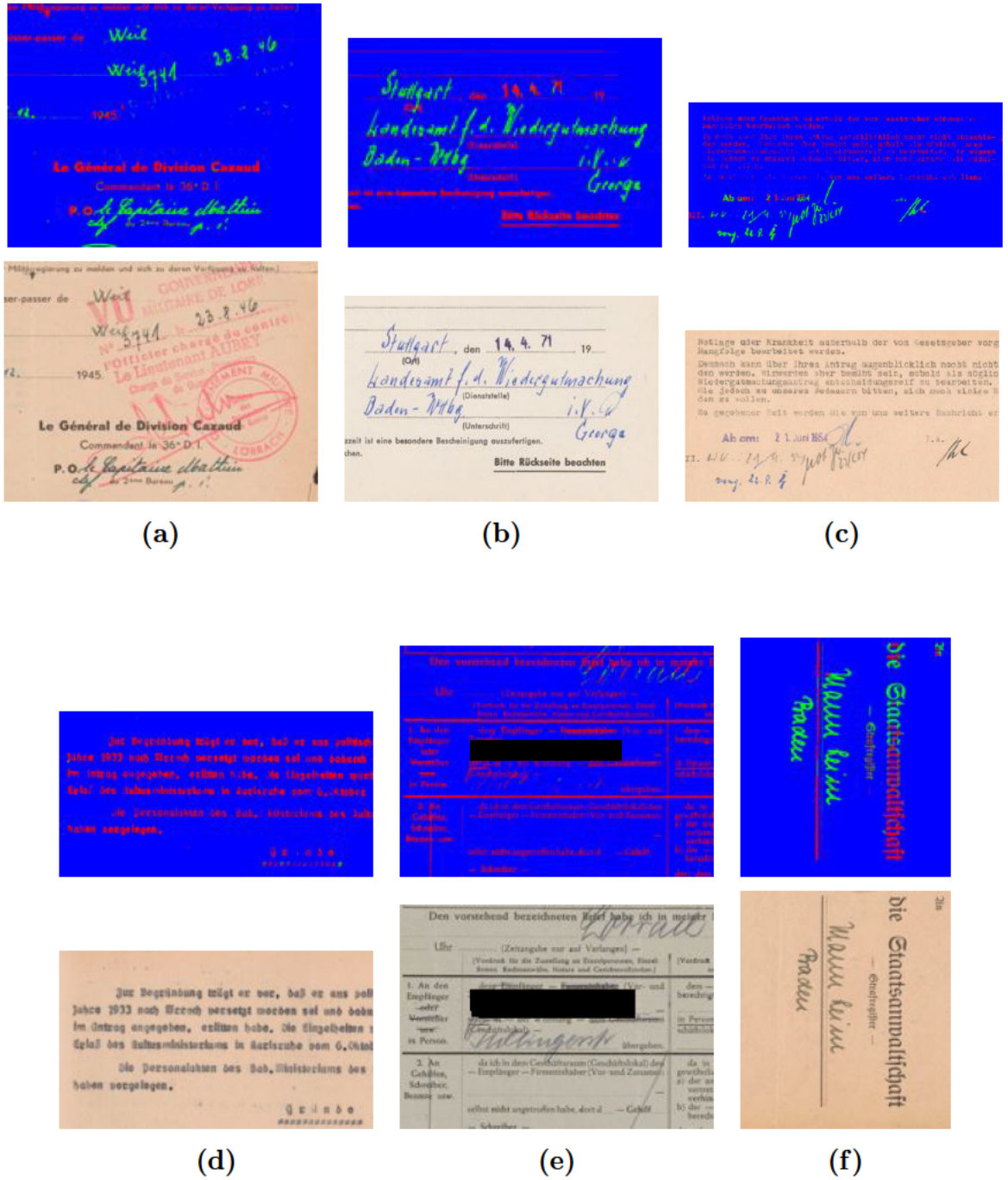


Figure 28: Results achieved by the *fcnn_wgm-cvl_jottueset_subset* model on real historical documents of the color photos subset (*wgm*).

(a) (b) (c)

Figure 29: Some results achieved by the *fcnn_wgm-cvl_jottueset_subset* model on real historical documents of the Microfilm subset (*wgm*).

6 Discussion and Conclusions

6.1 Discussion

As demonstrated in Tables 11, 12 and 13 showing models' performances before and after CRF post-processing, no significant increased benefit can be expected from using CRF post-processing in the newly created models. Especially in regions where printed and handwritten components are overlapping each other, printed components are often incorrectly identified. As previously noted, the synthesized patches used for model training, validation, and evaluation are excessively noisy to be realistic and thus might yield such unsatisfied results of the post-processing step. Models' segmentation results on real and therefore less noisy documents of the *wgm* dataset (Fig. 22, 23, 25, 26, 28, 29) indicate that the models provide very different and better outcomes compared to segmentation results of the original model *fcnn_ bin_ simple*[36] proposed in [5] (see Figures 30 and 31). Indeed, even though the *Sauvola* binarization method with unchanged parameters was used in the original model as well as in newly created models, considerably fewer text parts were recognized by the original model. Hence, the newly created models trained on synthesized data show the performance improvement in the text segmentation task in historical documents.

Considering the newly created and already evaluated models, the *fcnn_ cvl_ jottueset_ subset* model completes with the highest evaluation scores (see Table 11). Nevertheless, this model was trained and evaluated on synthesized patches created using only modern and not historical documents. In fact, these patches (Fig.11) are less noisy and contain fewer overlappings compared to other synthesized datasets (*wgm* and *wgm-cvl_ jottueset_ subset*). Therefore, the evaluation scores do not represent how well the model will perform on historical records which are typically extremely noisy. For this reason, each model, inclusive the *fcnn_ cvl_ jottueset_ subset*, was used for segmentation of some real documents taken from the *wgm* dataset (Fig. 25, 26, 22, 23, 28, 29). In such manner, the best segmentation results shows the *fcnn_ wgm-cvl_ jottueset_ subset* model, which was trained and evaluated with the compose *wgm-cvl_ jottueset_ subset* dataset consisting of *wgm* (historical documents) and of *cvl_ jottueset_ subset* (modern documents).

The model performance could be improved by expanding the dataset used for training. In fact, there is already the *fcnn_ wgm-cvl_ jottueset* model pre-trained with the *wgm-cvl_ jottueset* dataset, although this model was not evaluated yet due to the high computer capacity required. Nevertheless, the model was applied to real *wgm* documents (see Figures 32 and 33). Indeed, the *fcnn_ wgm-cvl_ jottueset* model shows the best seg-

[36] https://github.com/Jumpst3r/printed-hw-segmentation/blob/master/FCN/models/fcnn_bin_simple.h5

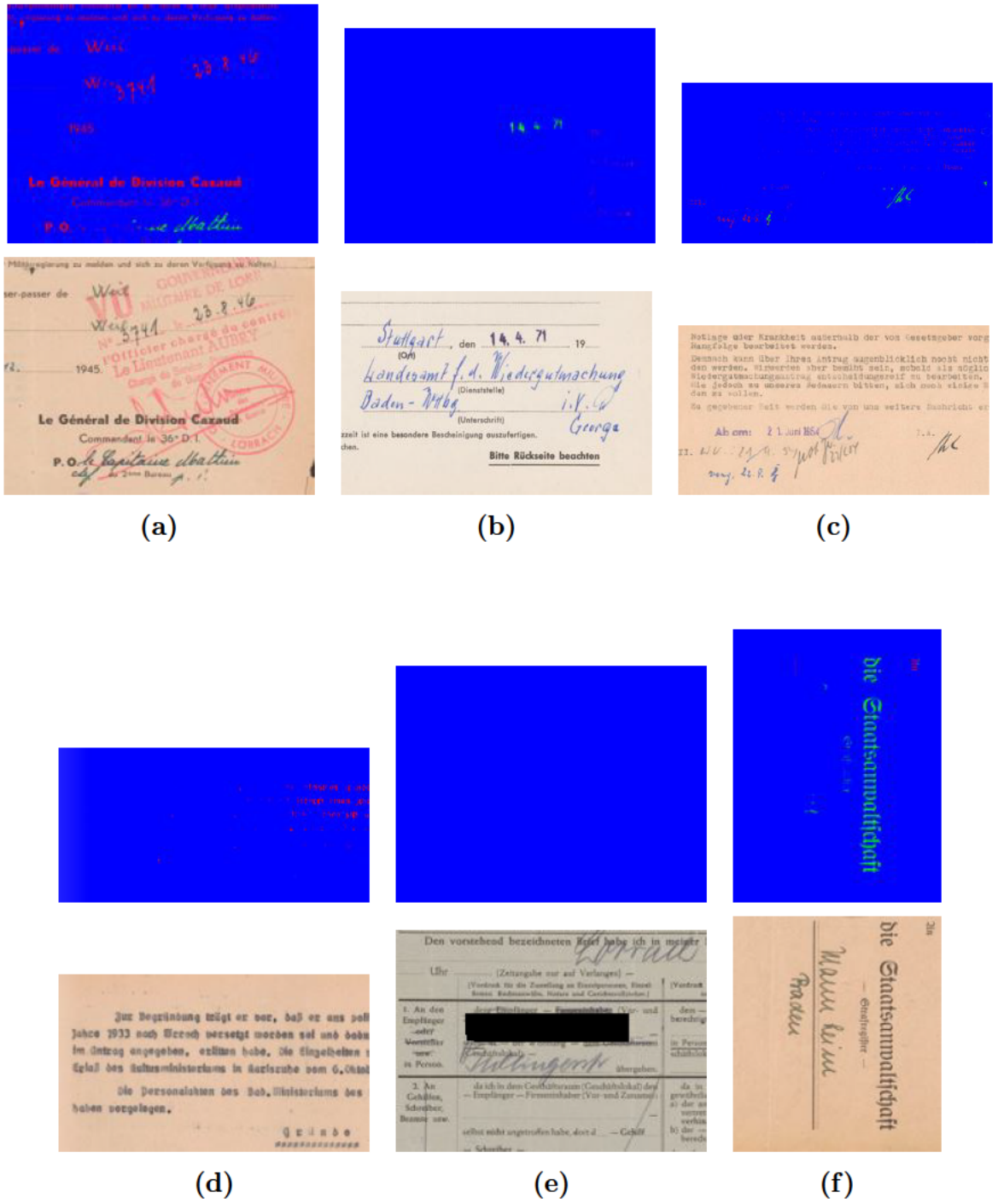
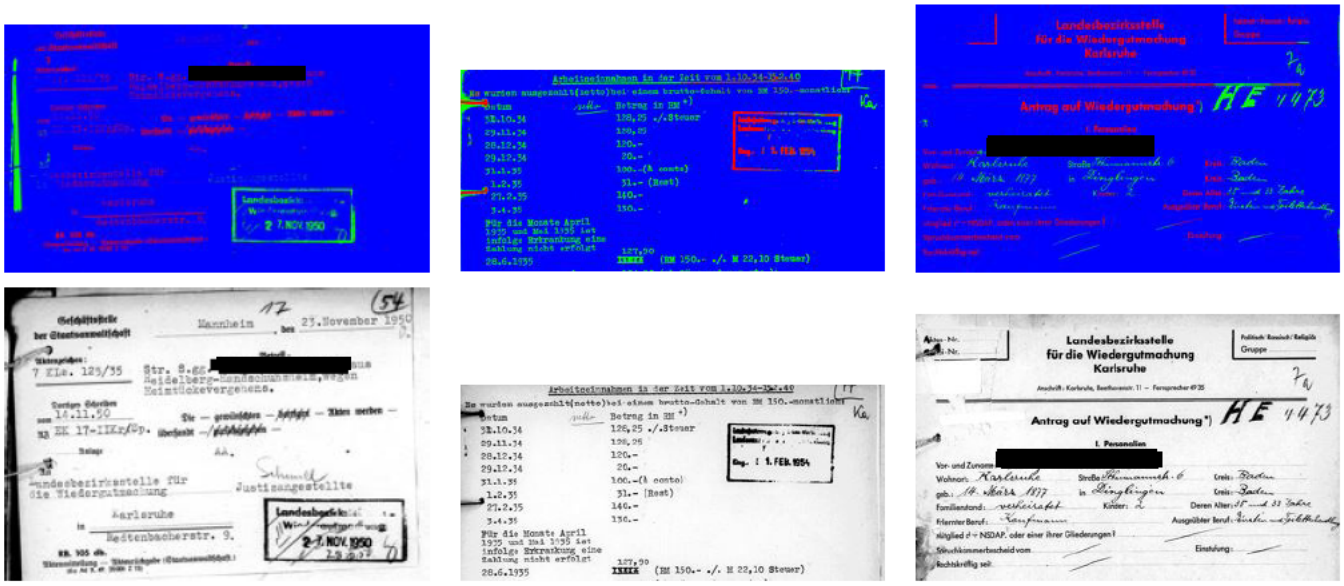


Figure 30: Results achieved by the original model [5] on real historical documents of the color photos subset (*wgm*).



(a) (b) (c)

Figure 31: Results achieved by the original model [5] on real historical documents of the Microfilm subset (*wgm*).

mentation results on the selected documents compared to other models created in this work. Moreover, this dataset could be extended by adding more diverse page scans of historical documents. Furthermore, to improve the model performance, use could be made of the other previously annotated labels such as noise, signature, and stamp areas of *wgm* documents for more clear segmentation results.

6.2 Conclusion and Future Work

In this work, the synthesized *wgm* dataset by overlaying printed and handwritten components of historical documents, as well as the *cvl_jottueset* dataset using modern documents, were introduced. Further, a cropped version of the *IAM* dataset containing only handwritten parts was created that could be used for the data synthesis purpose. Using the synthesized datasets, several models based on the model introduced in [5] for Latin-based texts were created. Compared to the original model in [5], these models show better results in the context of text segmentation in historical documents especially considering overlapping areas. The results achieved in this work show that the problem of text segmentation in historical records can be partially addressed by introducing more training data since historical documents are so challenging for the segmentation task due to the high number of different styles, old fonts, and spellings usually contained in them.

Therefore, the next steps in this area include a dataset extension by adding more diverse page scans of historical documents. Diversity in synthesized data could be repre-

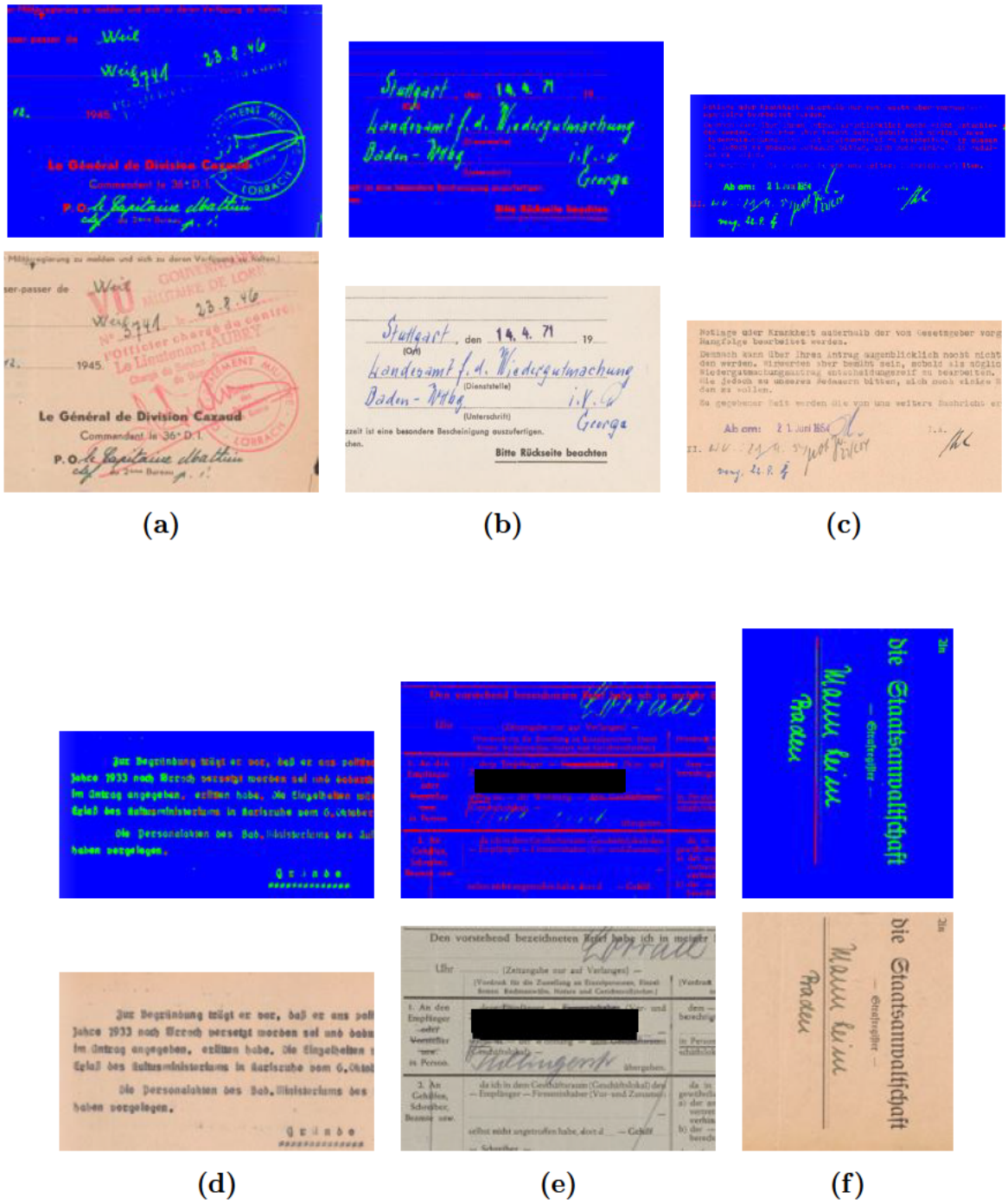
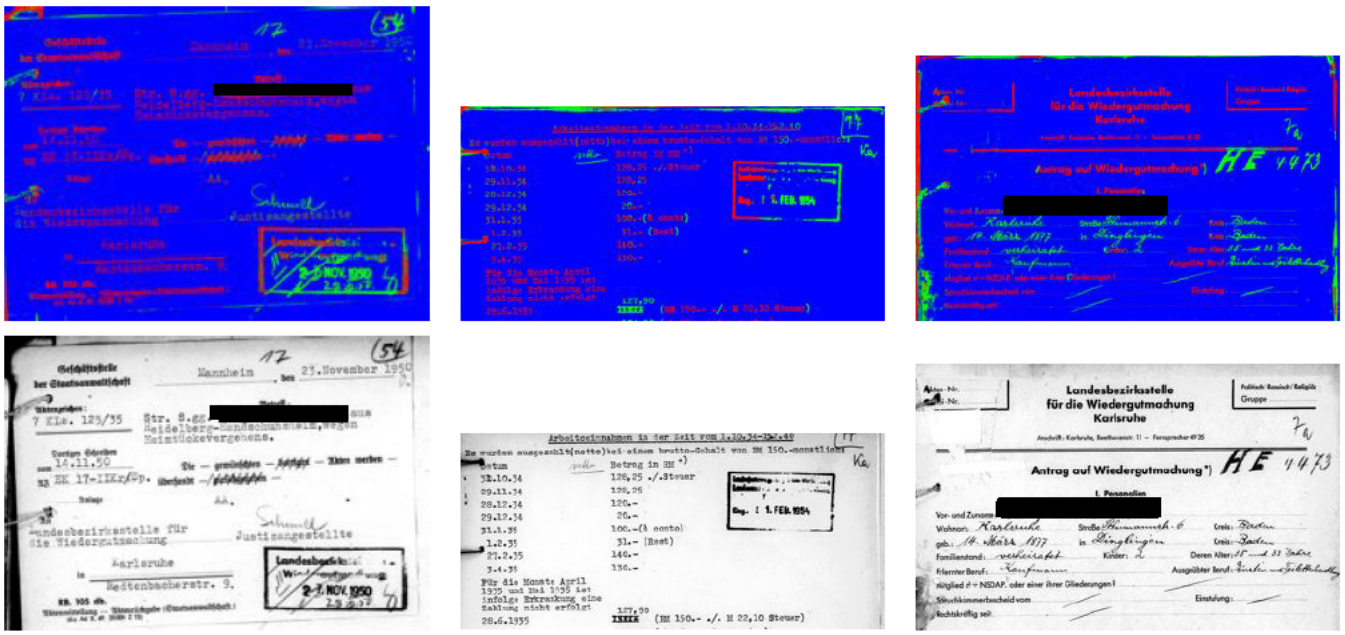


Figure 32: Results achieved by the *fcnn_wgm-cvl_jottueset* model (trained on the whole *wgm-cvl_jottueset* dataset) on real historical documents of the color photos subset (*wgm*).



(a) (b) (c)

Figure 33: Results achieved by the*fcnn_wgm-cvl_jottueset* model (trained on the whole *wgm-cvl_jottueset* dataset) on real historical documents of the Microfilm subset (*wgm*).

sented by creating not only noisy patches with many overlappings but also patches with fewer cases of overlaps might be considered. Thus, the parameters of the data synthesis method [12] could be changed. In addition to the synthesized data, real labeled modern documents such as annotated *IAM* data, and real labeled historical documents could be used for model training and evaluation as well. Further, other pre-processing techniques for noise removal, in particular the binarization method of Microfilm *wgm* images, could be taken into consideration, since the data contains excessive noise due to the degraded state of paper. For the same reason, annotation of noisy areas in *wgm* documents might be considered as well. Finally, to improve the model performance and achieve more clear segmentation results, use could be made of the other previously annotated labels of *wgm* data such as *mixed*, *signature* and *stamp* areas.

References

[1] Breuel, T. M., Ul-Hasan, A., Al-Azawi, M. A., and Shafait, F. High-performance ocr for printed english and fraktur using lstm networks. In *2013 12th international conference on document analysis and recognition* (2013), IEEE, pp. 683–687.

[2] Chammas, E., Mokbel, C., and Likforman-Sulem, L. Handwriting recognition of historical documents with few labeled data. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)* (2018), IEEE, pp. 43–48.

[3] Dhawan, A., Bodani, P., and Garg, V. Post processing of image segmentation using conditional random fields. In *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)* (2019), pp. 729–734.

[4] Drobac, S., Kauppinen, P., and Lindén, K. Ocr and post-correction of historical finnish texts. In *Proceedings of the 21st Nordic Conference on Computational Linguistics* (2017), pp. 70–76.

[5] Dutly, N., Slimane, F., and Ingold, R. Phti-ws: A printed and handwritten text identification web service based on fcn and crf post-processing. *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)* (2019), 20–25.

[6] Fischer, A., Indermühle, E., Bunke, H., Viehhauser, G., and Stolz, M. Ground truth creation for handwriting recognition in historical documents. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems* (2010), pp. 3–10.

[7] Garlapati, B. M., and Chalamala, S. R. A system for handwritten and printed text classification. *2017 UKSim-AMSS 19th International Conference on Modelling & Simulation* (2017), 50–54.

[8] Garz, A., Seuret, M., Simistira, F., Fischer, A., and Ingold, R. Creating ground truth for historical manuscripts with document graphs and scribbling interaction. *2016 12th IAPR Workshop on Document Analysis Systems (DAS)* (2016), 126–131.

[9] Guo, J. K., and Ma, M. Y. Separating handwritten material from machine printed text using hidden markov models. In *Proceedings of Sixth International Conference on Document Analysis and Recognition* (2001), pp. 439–443.

[10] Ho, Y., and Wookey, S. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE Access 8* (2020), 4806–4813.

[11] Islam, N., Islam, Z., and Noor, N. A survey on optical character recognition system. *arXiv preprint arXiv:1710.05703* (2017).

[12] Jo, J., Koo, H. I., Soh, J. W., and Cho, N. I. Handwritten text segmentation via end-to-end learning of convolutional neural networks. *Multimedia Tools and Applications 79*, 43-44 (2020), 32137–32150.

[13] Kandan, R., Reddy, N., Arvind, K., and Ramakrishnan, A. A robust two level classification algorithm for text localization in documents. In *International conference on Advances in visual computing* (2007), pp. 96–105.

[14] Kemker, R., Gewali, U. B., and Kanan, C. Earthmapper: A tool box for the semantic segmentation of remote sensing imagery, 2018.

[15] Kim, M.-S., Jang, M.-D., Choi, H.-I., Rhee, T.-H., Kim, J.-H., and Kwag, H.-K. Digitalizing scheme of handwritten hanja historical documents. *First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings.* (2004), 321–327.

[16] Kleber, F., Fiel, S., Diem, M., and Sablatnig, R. Cvl-database: An off-line database for writer retrieval, writer identification and word spotting. In *2013 12th International Conference on Document Analysis and Recognition* (2013), pp. 560–564.

[17] Krähenbühl, P., and Koltun, V. Efficient inference in fully connected crfs with gaussian edge potentials. In *Proceedings of the 24th International Conference on Neural Information Processing Systems* (2011), p. 109–117.

[18] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States* (2012), pp. 1106–1114.

[19] Lafferty, J., Mccallum, A., and Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning* (2001), pp. 282–289.

[20] Lins, R., Almeida, M., Bernardino, R., Jesus, D., and Oliveira, J. Assessing binarization techniques for document images. In *Proceedings of the 2017 ACM Symposium on Document Engineering* (2017), pp. 183–192.

[21] Liu, F., Lin, G., and Shen, C. CRF learning with CNN features for image segmentation. *CoRR abs/1503.08263* (2015), 2983–2992.

[22] Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 3431–3440.

[23] Lu, X., Li, B., Yue, Y., Li, Q., and Yan, J. Grid r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).

[24] Mahdi, F., Motoki, K., and Kobashi, S. Optimization technique combined with deep learning method for teeth recognition in dental panoramic radiographs. In *Scientific Reports* (2020), vol. 10.

[25] Memon, J., Sami, M., Khan, R. A., and Uddin, M. Handwritten optical character recognition (ocr): A comprehensive systematic literature review (slr). *IEEE Access 8* (2020), 142642–142668.

[26] Nafchi, H. Z., Ayatollahi, S. M., Moghaddam, R. F., and Cheriet, M. An efficient ground truthing tool for binarization of historical manuscripts. In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR* (2013), pp. 807–811.

[27] Nagy, G., and Lopresti, D. Interactive document processing and digital libraries. *Second International Conference on Document Image Analysis for Libraries (DIAL'06)* (2006), 2–11.

[28] Niblack, W. *An Introduction to Digital Image Processing.* Strandberg Publishing Company, DNK, 1985.

[29] Otsu, N. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics 9*, 1 (1979), 62–66.

[30] Peng, X., Setlur, S., Govindaraju, V., and Sitaram, R. Handwritten text separation from annotated machine printed documents using markov random fields. *International Journal on Document Analysis and Recognition (IJDAR) 16*, 1 (2013), 1–16.

[31] Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. *LNCS* (2015), 234–241.

[32] Rosasco, L., De Vito, E., Caponnetto, A., Piana, M., and Verri, A. Are loss functions all the same? *Neural computation 16* (2004), 1063–76.

[33] Sánchez, J. A., Bosch, V., Romero, V., Depuydt, K., and de Does, J. Handwritten text recognition for historical documents in the transcriptorium project.

In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage* (2014), pp. 111–117.

[34] Sauvola, J., and Pietikäinen, M. Adaptive document image binarization. pattern recognition. *Pattern Recognition 33* (2000), 225–236.

[35] Shetty, S., Srinivasan, H., and Beal, M. Segmentation and labeling of documents using conditional random fields. In *Proceedings of SPIE* (2007), vol. 6500, pp. 6500–1.

[36] Shivhare, P., and Gupta, V. Review of image segmentation techniques including pre & post processing operations. *International Journal of Engineering and Advanced Technology 4*, 3 (2015), 153–157.

[37] Springmann, U., Najock, D., Morgenroth, H., Schmid, H., Gotscharek, A., and Fink, F. Ocr of historical printings of latin texts: problems, prospects, progress. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage* (2014), pp. 71–75.

[38] Teichmann, M. T. T., and Cipolla, R. Convolutional crfs for semantic segmentation. *CoRR abs/1805.04777* (2018).

[39] Tensmeyer, C., and Martinez, T. Historical document image binarization: A review. In *SN Computer Science* (2020), vol. 1.

[40] Wolf, C., Jolion, J.-M., and Chassaing, F. Text localization, enhancement and binarization in multimedia documents. In *2002 International Conference on Pattern Recognition* (2002), vol. 2, pp. 1037–1040 vol.2.

Assertion

Ich versichere wahrheitsgemäß, die Arbeit selbstständig verfasst, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht zu haben, was aus Arbeiten anderer unverändert oder mit Abänderungen entnommen wurde sowie die Satzung des KIT zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet zu haben.

Karlsruhe, 14.10.2021 ANASTASIA PRIKHODINA