ChronSeg: Novel Dataset for Segmentation of Handwritten Historical Chronicles

Josef Baloun[1,2][a], Pavel Král[1,2][b] and Ladislav Lenc[1,2][c]

[1]*Department of Computer Science and Engineering, University of West Bohemia, Univerzitní, Pilsen, Czech Republic*
[2]*NTIS - New Technologies for the Information Society, University of West Bohemia, Univerzitní, Pilsen, Czech Republic*

Keywords: Page Segmentation, Dataset, Chronicle, Historical Document, Image, Text, Background, Fully Convolutional Neural Network, Pixel Labeling, Artificial Page.

Abstract: The segmentation of document images plays an important role in the process of making their content electronically accessible. This work focuses on the segmentation of historical handwritten documents, namely chronicles. We take image, text and background classes into account. For this goal, a new dataset is created mainly from chronicles provided by Porta fontium. In total, the dataset consists of 58 images of document pages and their precise annotations for text, image and graphic regions in PAGE format. The annotations are also provided at a pixel level. Further, we present a baseline evaluation using an approach based on a fully convolutional neural network. We also perform a series of experiments in order to identify the best method configuration. It includes a novel data augmentation method which creates artificial pages.

1 INTRODUCTION

Nowadays, considerable efforts are being made to digitize documents and make them accessible electronically in most areas like business or archives. The goal is usually to reduce storage costs or to make documents available to the general public. This effort is also covered by the project Porta fontium, which aims to combine extensive digitization and web presentation of the archival documents from the Czech-Bavarian border area. The goal of the project is the reconnection of related archival documents that were divided due to the events of the World War II. The archival documents should be published via the website[1] to the general public, the scientific world and regional researchers. There are also efforts to improve the search options by automatic annotation of the archival document pages.

This work focuses on pages of chronicles and their segmentation and classification into text, image and background classes. These segments are crucial for further processing. For example, an Optical Character Recognition (OCR) system requires a text input,

---

[a] https://orcid.org/0000-0003-1923-5355
[b] https://orcid.org/0000-0002-3096-675X
[c] https://orcid.org/0000-0002-1066-7269
[1] http://www.portafontium.cz/

but it could behave unpredictably when the input is an image of a house. In such a case, the usability of the result could be reduced due to the produced noise. Image segments can be used to search for related images. The result also provides information whether the page contains image or text.

To solve this task, there is a need for a precisely annotated dataset with similar characteristics as the chronicles. Ideally, the pages in the dataset should contain the annotation of correct regions and also should be realistic and representative. Because of inappropriate training data in existing datasets and the lack of training data for deep learning methods in general, it is beneficial to create a new annotated dataset.

The new dataset consists of scanned pages provided by Porta fontium portal and contains precise text, image and graphic region annotations. The main emphasis is placed on the chronicles that form the largest part of the dataset. In total, there are 58 annotated pages or double-sided pages of scanned documents.

Further, we performed a series of experiments on the dataset with a fully convolutional neural network (FCN) that is used as a baseline method. Generally, the pages do not have common or regular pattern, so the task is solved as a pixel-labeling problem. The network takes the whole image as input and produces masks for given classes. The experiments show that

the page segmentation task can be successfully solved even with a small amount of data. We also present an approach to automatically create artificial pages that can be used for data augmentation.

2 RELATED WORK

The issue of locating text in document images has a long history dating back to the late 1970s when OCR was addressed and it was necessary to extract these characters. "In order to let character recognition work, it is mandatory to apply layout analysis including page segmentation." (Kise, 2014) Today, there is also a need for extracting images from pages. The extracted images can be also further processed to allow image search for example.

This section first summarizes recent methods for page segmentation and then it provides a short overview of available datasets.

2.1 Methods

Today, there are lots of methods for page segmentation. The methods can be categorized into top-down and bottom-up categories. Historically, the segmentation problem was usually solved by conservative approaches based on simple image operations and on connected component analysis. Recent trend is to use neural networks for this task.

A method for segmenting pages using connected components and a bottom-up approach is presented in (Drivas and Amin, 1995). The method includes digitization, rotation correction, segmentation and classification into text or graphics classes. Another approach based on background thinning that is independent of page rotation is presented in (Kise et al., 1996). These conservative methods usually fail on handwritten document images, because it is hard to binarize pages due to degraded quality. It is also hard to extract characters since they are usually connected. These problems are successfully solved by approaches based on convolutional neural networks (CNN) that brought a significant improvement in many visual tasks. An example of a CNN for page segmentation of historical document images is presented in (Chen et al., 2017). Briefly, super-pixels (groups of pixels with similar characteristics) are found in the image and they are classified with the network that takes 28x28 pixels input. The result of the classification is then assigned to the whole superpixel.

Alternatively, every pixel could be classified separately in a sliding window manner. The problem is computational inefficiency because a large amount of computation is repeated as the window moves pixel by pixel. This problem is solved by FCNs like well-known U-Net (Ronneberger et al., 2015). U-Net was initially used for biomedical image segmentation but could be used on many other segmentation tasks including page segmentation. Another FCN architecture is presented by (Wick and Puppe, 2018), this network is proposed for page segmentation of historical document images. In contrast with U-Net, it does not use skip-connections and uses transposed convolutional layer instead of upsampling layer followed by convolutional layer. The speed improvement is achieved mainly thanks to the small input of 260x390 pixels. In order to achieve the best results in competitions, there are also networks like (Xu et al., 2017). This network uses the original resolution of images and provides many more details in the outputs.

2.2 Datasets

There are many architectures that solve the problem of segmentation very well. The main problem is the corresponding training data because appropriate data are the key point of approaches based on neural networks. There are several datasets for a wide range of tasks. Unfortunately, a significant number of datasets are inappropriate for our task or they are not publicly available.

Diva-hisdb (Simistira et al., 2016) is a publicly available dataset with detailed ground-truth for text, comments and decorations. It consists of three manuscripts and 50 high-resolution pages for each manuscript. These manuscripts have similar layout features. The first two manuscripts come from the 11th century. They are written in Latin language using the Carolingian minuscule script. The third manuscript is from the 14th century and shows a chancery script. The language is Italian and Latin. Unfortunately, there are no images on pages.

Handwritten historical manuscript images are available in the repository IAM-HistDB (Fischer et al., 2010) together with ground-truth for handwriting recognition systems. Currently, it includes three datasets: Saint Gall Database, Parzival Database and Washington Database. The Saint Gall Database (Fischer et al., 2011) contains 60 page images of a handwritten historical manuscript from 9th century. It is written in Latin language and Carolingian script. 47 page images of a handwritten historical manuscript from 13th century are available in the Parzival Database (Fischer et al., 2012). The manuscript is written in Medieval German language and Gothic script. The Washington database (Fischer et al., 2012) is created from the George Washington

Papers. There are word and text line images with transcriptions. The provided ground-truth is not intended for page segmentation, but Saint Gall Database contains line locations that can be used for text segmentation.

Another dataset is Layout Analysis Dataset (Antonacopoulos et al., 2009), that is precisely annotated for page layout analysis and contains suitable regions for our task. The dataset contains a huge amount of page images of different document types. There is a mixture of simple and complex page layouts with varying font sizes. The problem is that the documents are printed and consist mostly of modern magazines and technical journals so the page layout is totally different to our chronicles in most cases.

There are also competition datasets at PRImA website[2]. These datasets has to be requested first and consist mainly of newspapers, books and typewritten notes. The number of annotated pages is usually around ten per dataset. Like in the Layout Analysis Dataset, the text is printed and page layout is different to our chronicles.

3 DATASET DESCRIPTION

The dataset is a collection of documents for the task of historical handwritten chronicles segmentation. This collection is closely described in this section together with provided ground-truths and terms of use.

3.1 Content

The dataset is composed of scanned pages provided by Porta fontium portal. The main part of the dataset consists of 5 chronicles in a total of 38 double-sided pages from which 18 contain images (see Figure 1). We used chronicles from the rectory of Budětice, the rectory of Petrovice u Sušice, the town of Blovice, the town of Chudenice and the village of Hroznětín.

This set is chosen with emphasis on realism and representativeness in a wide range of chronicles that appear in Porta fontium. These documents had been created for many years. For example, the chronicle from the rectory of Budětice is dated from 1649 to 1981. However, the latest chronicle from the village of Hroznětín was written from 1949 to 1954. Usually, there are several writers and page layouts in one chronicle. There are also printed text cuttings glued into some pages. These facts make the dataset really challenging.
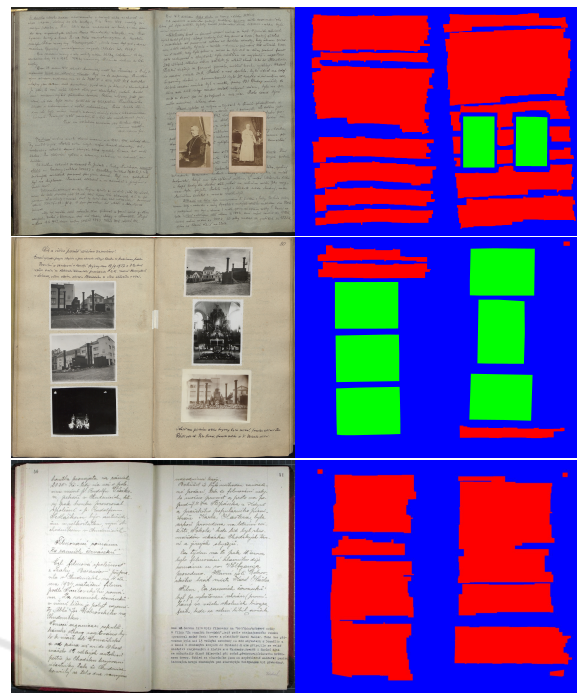
---

[2]https://www.primaresearch.org/datasets

Figure 1: Dataset examples: double-sided page (on the left) and its ground-truth (on the right).

We have selected a representative set of pages from the main part of the dataset and divided it into train, test and validation sets. The test and validation parts of the dataset consist of 8 and 4 pages with images respectively. In the training set, there are 26 pages: 6 pages with images and 20 no-image pages.

There is also an experimental part that contains 20 printed pages from documents of different types. This set contains rare pages (see Figure 2 for example) that can be used in different application areas. Additionally, this part contains 29 standalone images of old photographs. The experimental part is further used in experiments.

Totally, there are images of 58 good-quality pages of documents and their dimensions vary from 2000 to 5777 pixels. The detailed summary of the dataset content is provided in Table 1.

Table 1: Content of the created dataset.

	Main part	Exp. part	Total
Pages	38	20	58
Text regions	553	423	976
Image regions	33	31	64
Graphic regions	6	38	44
Images	0	29	29

Figure 2: Page example of the experimental part.

3.2 Ground-truth

Every page image in the dataset contains annotation in the widely used XML-based PAGE format[3] and also the generated pixel-labeled ground-truth.

The annotation is provided for text, graphic and image regions. The text regions are used for any block of text, e.g. handwritten text paragraph, glued printed text cutting or text over images. The image region represents a picture or drawing. Finally, the graphic regions are used for decoratives and stamps. The annotations are created with the document analysis system Aletheia (Clausner et al., 2011). The annotation process is visualized in Figure 3 and is as follows:

1. Page binarization using Otsu or adaptive binarization

2. Noise reduction in binary image

3. Annotation of regions
   - a tool *To Coarse Contour* is used for text
   - a tool *To Fine Contour* is used for image and graphic with complex shape
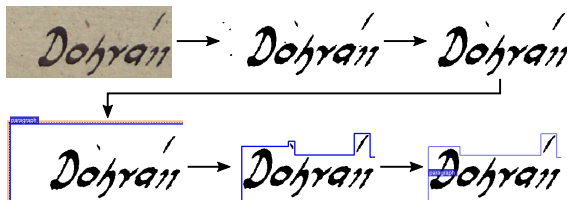
4. Inspection and manual correction

Figure 3: The process of annotation: After the image is binarized, the noise is filtered out. Then, the region is annotated and a tool To Coarse Contour is used for text. Finally, the annotation is manually corrected.

Additionally, the pixel-level ground-truth is generated and provided in *png* files, where R channel contains

---

[3]https://www.primaresearch.org/schema/PAGE/

ground-truth for text, G channel for image/graphic and B channel is used as a background. Since the experimental part contains the main part of the graphic regions (38 of 44), we treat graphic region as the same class as image. Since the task is solved as a pixel-labeling problem, the number of samples corresponds to the number of pixels rather than number of page images. Totally, there are more than 773 million pixels labeled, which represents enough samples for training, validation and testing.

3.3 Terms of Use

This dataset is licensed under the Attribution-NonCommercial-Share Alike 4.0 International License[4]. Therefore, it is freely available for research purposes at http://corpora.kiv.zcu.cz/segmentation.

4 EVALUATION

We also provide a baseline evaluation on this dataset to offer to researchers a possibility of straightforward comparison of their methods.

For the testing phase, the evaluation is as follows: The output of the system is compared with the ground-truth separately for each class. The final result is computed as the average of all images results for the given class. Finally, the average of classes is computed.

Since each pixel is binary labeled for a given class, it can be figured in True Positive (TP), True Negative (TN), False Positive (FP) or False Negative (FN) sets. Based on these sets the metrics *accuracy*, *precision*, *recall*, *F1 score* and *Intersection over Union* are applied (see Equations 1). (Wick and Puppe, 2018) proposed the *Foreground Pixel Accuracy* which is practically an accuracy calculated only over foreground pixels. In this work, foreground pixels are obtained with an adaptive document image binarization method (Sauvola and Pietikäinen, 2000).

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$
$$precision = \frac{TP}{TP+FP}$$
$$recall = \frac{TP}{TP+FN} \tag{1}$$
$$F1\ score = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$
$$IoU = \frac{TP}{TP+FP+FN}$$

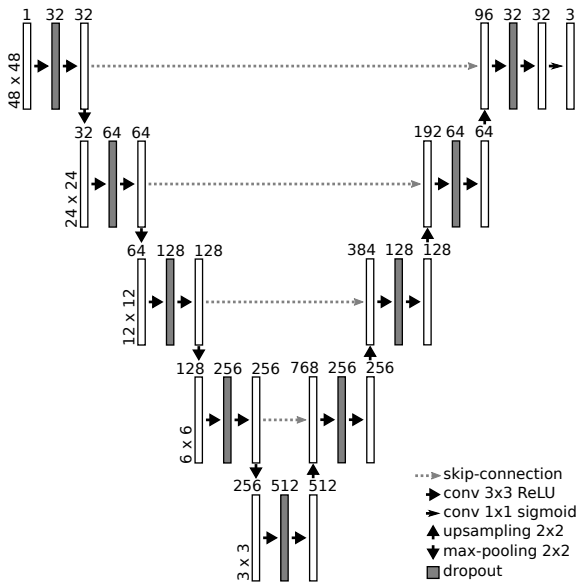[4]https://creativecommons.org/licenses/by-nc-sa/4.0/

Figure 4: Neural network architecture: Example for 48x48 input. Each box represents a feature map. The number of channels is marked above the box. Dimensions are the same for the whole level and are described on the left of the leftest box.

5 NETWORK MODEL

The U-Net like fully convolutional neural network architecture used as a baseline model is described in Figure 4. It is designed to process the whole page (double-sided page) all at once so it uses padding for convolutional layers to preserve dimensions. The padding also lets the network process differently borders of the input, where increased amount of noise occurs.

Convolutional layers use ReLU activation function except the last one that uses sigmoid and maps feature vectors to desired classes. Thanks to the shared parameters in the convolutional layers, the architecture can take practically any input dimension as long as the input dimension satisfies the Equation 2. This equation is given by four max-pooling layers. Max-pooling layers need to have even input. Otherwise, the dimension inconsistency will appear during concatenating feature maps for skip-connections after the upsampling layer.

$$x = i \cdot 16, \qquad i = 1, 2, \dots, \infty \qquad (2)$$

In image overlayed with text, we want the pixel to be classified as both text and image. For this goal, the Binary Cross-Entropy loss function is chosen, so every output channel is processed independently on the other output channels. The optimization during train-

ing is done with Adam optimizer (Kingma and Ba, 2014). For the training, dropout (dropout rate = 0.2) is applied and the technique of early stopping is used. If the IoU does not improve during the last 30 epochs (could be more, depends on the experiment), the training is stopped and the model with the best IoU score is selected for the evaluation.

The input image size is limited to 512x512 pixels to save time and satisfy the memory limitations. We found this setup reasonable according to the results of experiments in the Section 6.
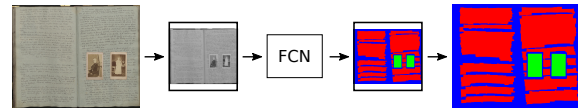

Figure 5: Prediction of an input image with FCN model. Boxes before and after FCN represent the limit of 512x512 input.

The processing of the page image starts by converting it to a gray level image. Secondly, the desired input image size is computed to satisfy the 512x512 input limit in a way that maintains the aspect ratio of the image. In this step, the smaller dimension may not fit Equation 2 so it is eventually fixed as the nearest correct dimension. Finally, the image is resized to the computed dimensions. The predicted output is then resized back to the input image dimensions as could be seen in Figure 5. We emphasize that the input size could be different during both training and prediction. For example, there could be 512x400 and also 512x416 input.

6 EXPERIMENTS

We made several experiments with extending the number of training samples, weighting loss function and input resolutions to find out their influence on the result and to choose the correct setup for the final model training. We also present a data augmentation approach for automatic creation of artificial pages from existing ones. This approach deals with the problem of class imbalances and significantly improves accuracy. The experiments are compared to the baseline model. For the baseline model, only the pages that contain images are used for training. The evaluation of the experiments is done on the validation part of the dataset. The test part of the dataset is used only for the evaluation of the final model.

Table 2: Average results (in %) of the experiments: Baseline is a referential model. Augmentation, artificial pages and printed pages are the experiments to extend the number of training samples presented in Subsection 6.1. Weighting loss function is an experiment described in Subsection 6.2. Final denotes the final model results.

	Accuracy	Precision	Recall	F1 score	IoU	FgPA
Baseline	95.3	91.8	92.6	92.0	85.5	98.5
Augmentation	95.5	93.2	92.7	92.8	86.7	98.4
Artificial pages	96.1	94.0	94.3	94.0	88.9	99.2
Printed pages	95.5	94.2	92.1	93.0	87.1	98.8
Weighting loss function	95.3	94.6	90.7	92.3	85.9	99.0
Final	96.4	94.5	94.3	94.2	89.2	99.2

6.1 Extending the Number of Training Samples

Although the task could be solved with a baseline model and 6 pages for training, more training data should provide better results.

A good approach for extending the number of training samples is an image augmentation (see Figure 6), where the same transformations are applied on the input image and its ground-truth. We were experimenting with the library Augmentor (Bloice et al., 2019) and methods *random_distortion*, *skew* and *rotate*. This combination led to an improvement (see *Augmentation* in Table 2). It is good to mention that the random_distortion is the most suitable for this architecture since it preserves the image borders and lets the network to learn that the borders are often labeled as a background and contain noise.
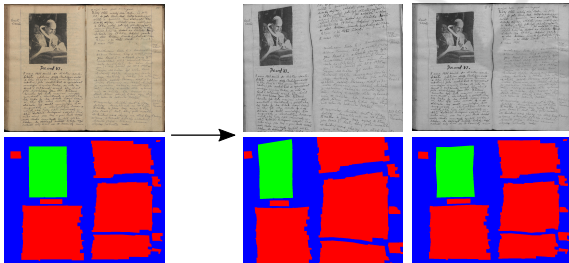

Figure 6: Image augmentation. From left: (1) input image and its ground-truth, (2) augmented image, (3) image augmented only with random_distortion.

There are pages in the dataset that do not contain any image. These no-image pages are causing class imbalances and problems during training. The solution is to add images into a no-image page in a way that is presented in Figure 7. The images are added randomly with reasonable position and size restrictions so that the image can not appear at the borders of the page image, the image size has to be at least 10x10 and image dimensions can not be bigger than 60 % of the page dimensions. This way, new artificial pages are created and the training set could be easily extended.
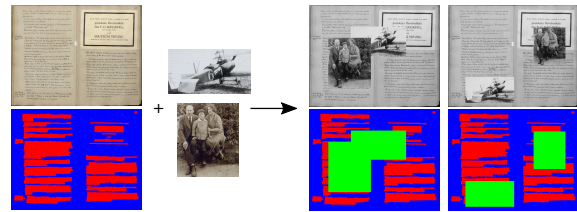

Figure 7: Artificial page creation: Document page is randomly extended with images.

Although the training process (see Figure 8) is not completely smooth, there is no evidence of problems caused by generated samples. Overall, the result of *artificial pages* presented in Table 2 shows a remarkable improvement.
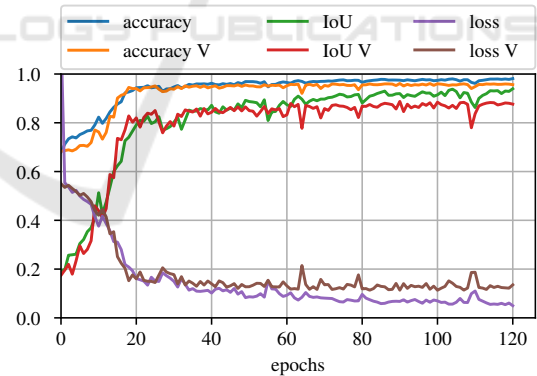

Figure 8: Training process during epochs (artificial pages): *accuracy*, *IoU* and *loss* on training and valid (denoted as *V*) set.

The training set could be also extended by the *printed pages* from the experimental part of the dataset and Layout Analysis Dataset (Antonacopoulos et al., 2009). This approach also led to an improvement but has not worked well in combination with previously described approaches.

Table 3: Average results (in %) with different input limits.

	Accuracy	Precision	Recall	F1 score	IoU	FgPA
512x512	96.1	94.6	93.8	94.1	88.9	99.1
1024x1024	96.6	94.7	94.0	94.2	89.2	99.2

6.2 Weighting Loss Function

This experiment faces the problem of separating components by weighting the loss function as proposed in the U-Net paper (Ronneberger et al., 2015). The weights are computed separately for text and image ground-truth channels. For these channels, the weight map is calculated according to Equation 3:

$$w(x) = w_0 \cdot \exp\left(-\frac{(d_1(x) + d_2(x))^2}{2\sigma^2}\right) \cdot (1 - \text{gt}(x)) + 1 \quad (3)$$

where x denotes the pixel position, $d_1(x)$ and $d_2(x)$ denotes the distance to the nearest and second nearest component. Value given by ground-truth at specified pixel is denoted as $\text{gt}(x)$ so more weight is added to the gaps between components as illustrated in Figure 9. During the experiments, we set $w_0 = 10$ according to the U-Net paper and increased $\sigma = 10$ because of wider gaps between components.
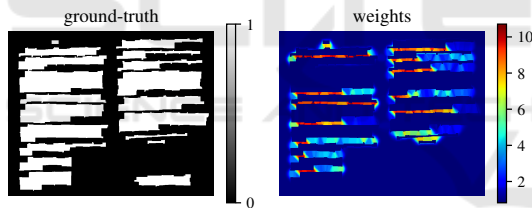
Figure 9: Example of calculated weights for weighting the loss function.

Although the results presented in Table 2 do not show noticeable improvement, the separation between components is much better as could be seen in Figure 10.

6.3 Input Resolution

We tried to increase the input resolution to provide more detail in the output and improve accuracy. The first approach was to train the network on random 512x512 crops of the image limited to 1024x1024. After the training, the prediction was done for the whole 1024x1024 page. This could be done thanks to the shared parameters in convolutional layers. This approach led to wrong predictions at image borders as presented in Figure 10. The explanation for this behaviour is the usage of padding in convolutional layers. Padding provides a lot of information for border
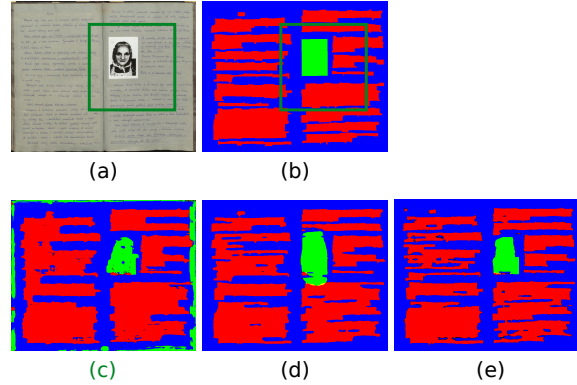
Figure 10: Prediction examples: (a) input image, (b) ground-truth, (c) prediction of the model trained on crops (example training sample in green box of (a) and (b)), (d) prediction of the base model, (e) prediction of a model with weighting loss function.

predictions. At the page borders, the input image contains noise, whereas training samples usually contain other classes than background. Then, it is much easier to mispredict the noise as the wrong class.

We also trained the network with 1024x1024 input but there was no huge improvement as could be seen in Table 3. During the training the augmentation and artificial pages were used for both 512 and 1024 inputs.
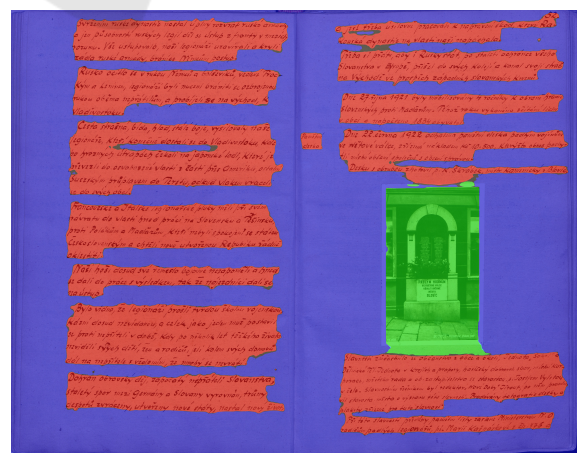
6.4 Final Model

Figure 11: Final model: An example prediction of the page from test set.

For the final model, an input was limited to

Table 4: Final evaluation on the test part of the dataset (in %).

	Accuracy	Precision	Recall	F1 score	IoU	FgPA
Text	96.3	95.8	92.1	93.8	88.4	98.7
Image	99.1	93.7	98.0	95.7	91.9	98.7
Background	96.1	96.5	96.4	96.4	93.1	99.0
Average	97.2	95.4	95.5	95.3	91.2	98.8

512x512 pixels and the weighting of the loss function with parameters set to $w_0 = 5$ and $\sigma = 10$ was used. The training part was extended by artificial pages and only random_distortion was applied for image augmentation since it preserves the image borders. With this training setup, the best results on the validation set were achieved. The results on the test part of the dataset are presented in Table 4 and an example prediction could be seen in Figure 11.

7 CONCLUSIONS AND FUTURE WORK

This paper introduces a novel dataset for page segmentation with a particular focus on handwritten chronicles. We consider text, image and background classes. The dataset contains 58 annotated pages in total from which 38 are double-sided pages of chronicles. Furthermore, we performed several experiments on the dataset to provide the baseline evaluation. Based on the experiments, we can say that the page segmentation task could be successfully solved even with a small amount of data. Also, the artificial pages can be created and used for training to improve accuracy with no problem. This step allows us to significantly increase the number of training samples.

Even though the presented baseline method achieved promising results of 97.2 % accuracy and 91.2 % IoU in average, there is still space for improvement.

We plan further processing of the network output with approaches based on the connected components to improve the results of segmentation. There is also a possibility of transfer learning. Finally, we plan to integrate the system into Porta fontium to improve search options.

ACKNOWLEDGEMENTS

REFERENCES

Antonacopoulos, A., Bridson, D., Papadopoulos, C., and Pletschacher, S. (2009). A realistic dataset for performance evaluation of document layout analysis. In *2009 10th International Conference on Document Analysis and Recognition*, pages 296–300.

Bloice, M. D., Roth, P. M., and Holzinger, A. (2019). Biomedical image augmentation using Augmentor. *Bioinformatics*, 35(21):4522–4524.

Chen, K., Seuret, M., Hennebert, J., and Ingold, R. (2017). Convolutional neural networks for page segmentation of historical document images. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 965–970. IEEE.

Clausner, C., Pletschacher, S., and Antonacopoulos, A. (2011). Aletheia - an advanced document layout and text ground-truthing system for production environments. In *2011 International Conference on Document Analysis and Recognition*, pages 48–52.

Drivas, D. and Amin, A. (1995). Page segmentation and classification utilising a bottom-up approach. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 2, pages 610–614 vol.2.

Fischer, A., Frinken, V., Fornés, A., and Bunke, H. (2011). Transcription alignment of latin manuscripts using hidden markov models. In *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, pages 29–36.

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

Fischer, A., Keller, A., Frinken, V., and Bunke, H. (2012). Lexicon-free handwritten word spotting using character hmms. *Pattern Recognition Letters*, 33(7):934–942.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kise, K. (2014). *Page Segmentation Techniques in Document Analysis*, pages 135–175. Springer London, London.

Kise, K., Yanagida, O., and Takamatsu, S. (1996). Page segmentation based on thinning of background. In *Proceedings of 13th International Conference on Pattern Recognition*, volume 3, pages 788–792 vol.3.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing.

Sauvola, J. and Pietikäinen, M. (2000). Adaptive document image binarization. *Pattern Recognition*, 33(2):225 – 236.

Simistira, F., Seuret, M., Eichenberger, N., Garz, A., Liwicki, M., and Ingold, R. (2016). Diva-hisdb: A precisely annotated large dataset of challenging medieval manuscripts. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 471–476. IEEE.

Wick, C. and Puppe, F. (2018). Fully convolutional neural networks for page segmentation of historical document images. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 287–292.

Xu, Y., He, W., Yin, F., and Liu, C.-L. (2017). Page segmentation for historical handwritten documents using fully convolutional networks. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 541–546. IEEE.