Proceedings of the Joint Austrian Computer Vision and Robotics Workshop 2020 DOI: 10.3217/978-3-85125-752-6-20

Classification and Segmentation of Scanned Library Catalogue Cards using Convolutional Neural Networks

Matthias Wödlinger, Robert Sablatnig
Computer Vision Lab, TU Wien
{mwoedlinger,sab}@cvl.tuwien.ac.at

Abstract. *The library of the TU Wien has been documenting changes in its inventory in the form of physical library archive cards. To make these archive cards digitally accessible, the cards and the text regions therein need to be categorized and the text must be made machine-readable. In this paper we present a pipeline consisting of classification, page segmentation and automated handwriting recognition that, given a scan of a library card, returns the category this card belongs to and an xml file containing the extracted and classified text.*

1. Introduction

A library catalogue is a register where all bibliographic entries found in a library are listed. In this paper we present a pipeline that automatically processes scanned images of library catalogue documents such that they can be made available and also searchable in an online database. While earlier work in this direction uses hand crafted rules and regular expressions to classify text in extracted OCR data, in recent years Convolutional Neural Network (CNN) based methods that operate on pixel level have formed the state-of-the-art in this task [4].

The library catalogue at hand consists of 113073 mostly handwritten documents, mostly collected in the time period from 1815 to 1930. The scanned images contain exactly the card with no surrounding content (see Fig. 1). Documents are classified into two groups: library cards with a "*Signatur*" (a unique identifier) that we call *S cards* and cards without it (*V cards*). V cards are not relevant for the online database and must be sorted out.

For training 2000 S cards and 500 V cards where manually extracted. The S cards where further sorted into 5 classes based on their layout. The text regions were manually annotated and verified by experts.

Model	Accuracy
ResNet18	0.988
ResNet34	0.988
ResNet50	**0.994**

Table 1. The accuracy scores on the test set. The accuracy is computed with respect to all 6 classes.

In this paper we describe a pipeline that, given a scanned library card image, determines if it is type S or V and then returns an xml file with the extracted and classified text. We describe the components of our pipeline in Section 2 and give a conclusion in Section 3.

2. Methodology and Results

The pipeline developed in this project is summarized in Fig. 1.

Classification of S and V cards We use a ResNet [2] pretrained on ImageNet and finetuned on our documents to sort out V cards. We do not freeze any layers during finetuning but instead train the full model with a smaller initial learning of $4 \cdot 10^{-4}$. To prevent large class imbalances we train the network on all 6 classes. The 2500 annotated documents are randomly split into train, test and validation sets and rescaled to 512×512. Table 1 shows the accuracy scores on the test set for three ResNets with different depth parameters.

Page segmentation of S cards The text regions in S cards are categorized in 7 classes that each contain document specific information like title, author, publisher or unique identifiers. The text region classes are distinguished from one another by location, font size and content. We use a CNN for image segmentation to detect and classify the text regions

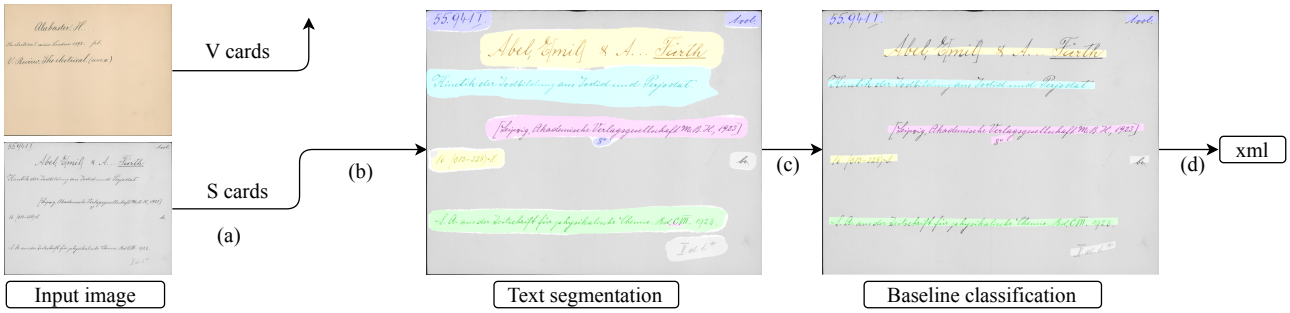

Figure 1. The proposed pipeline consisting of (a) an image classifier to sort out V cards (b) a segmentation network to detect and classify text regions and (c) baselines and finally (d) an HTR model whose output is combined with the baseline segmentation and saved as an xml file. Colors denote the different text categories.

Model	mIoU
Large Kernel Matters (ResNext101)	0.793
DeepLabV3+ (ResNet152)	**0.799**
dhSegment (ResNet50)	0.772

Table 2. The mIoU scores. The image classifiers in brackets denote the frontend used.

and later also the text baselines therein. We experiment with the models dhSegment [4], Global Convolutional Network (GCN) [5] and DeepLabV3+ [1]. The 2000 documents were first split in 50% train and 25% test and validation data each and then resized to 512×512. We found that adding a border around text regions (a line with constant width along the outline of text regions) as an additional class during training helps the network in learning to separate different text regions. Table 2 shows the mean intersection over union (mIoU) scores for the three best performing models. The segmentation is then used to classify the extracted text as described below.

Handwriting Recognition For the detection of text baselines and handwritten text recognition (HTR) model from Transkribus [3] are used. The Transkribus platform contains models for baseline detection and HTR pretrained on german Kurrent writing (with a character error rate of 7% on a seperate reference dataset [3]), which is the predominant writing style in our dataset. We apply the baseline detection of Transkribus, then classify the baselines according to the segmentation and add missing baselines for common errors. Afterwards the HTR model is applied and the result is saved as an xml file.

3. Conclusion

We have presented an approach for the automatic digitization of a library catalogue. We compared state-of-the-art models for semantic segmentation and found that DeepLabV3+ performs well in the task of page segmentation for historic handwritten documents. On the levels of baselines the classification of text using our segmentation appraoch performs reasonably well for the application however the character error rate of 7% needs improvement either through retraining on documents from our dataset or by manual corrections. For futher work, we believe that a better recognition of baselines has the largest potential for further improvements.

References

[1] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European conference on computer vision (ECCV)*, pages 801–818, 2018.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[3] P. Kahle, S. Colutto, G. Hackl, and G. Mühlberger. Transkribus – a service platform for transcription, recognition and retrieval of historical documents. In *International Conference on Document Analysis and Recognition (ICDAR)*, volume 4, pages 19–24. IEEE, 2017.

[4] S. A. Oliveira, B. Seguin, and F. Kaplan. dhsegment: A generic deep-learning approach for document segmentation. In *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 7–12. IEEE, 2018.

[5] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4353–4361, 2017.