Automatic Font Group Recognition in Early Printed Books

Weichselbaumer, Nikolaus
weichsel@uni-mainz.de
JGU Mainz, Deutschland

Seuret, Mathias
mathias.seuret@fau.de
FAU Erlangen, Deutschland

Limbach, Saskia
limbach@uni-mainz.de
JGU Mainz, Deutschland

Christlein, Vincent
vincent.christlein@fau.de
FAU Erlangen, Deutschland

Maier, Andreas
andreas.maier@fau.de
FAU Erlangen, Deutschland

Introduction

Early modern books were printed with a large variety of different fonts. In the first decades after Gutenberg's invention, every printer had to start out by cutting his own punches and casting his own fonts. This diversity was somewhat standardised with the advent of an organised font trade in the 16th century. However, this was a very long process and one that was not completed before the 19th century. Only then did industrialised mass production make fonts as stable and - at least for text fonts - predictable as we know them today. The diversity of fonts is one of the cornerstones of analytical bibliography: with detailed descriptions of the individual fonts we can identify the printer of almost any given incunabula or at least narrow the possible candidates down to a very small group.

For OCR, however, this is a major drawback. Most OCR-models are trained to work with one of three different training sets, based on either just modern antiqua-fonts or on 19th-century standard Fraktur or on all fonts that ever existed. Specialised OCR-models, e. g. for Rotunda or Textura, almost don't exist as they would be very difficult to apply. One reason for this is that metadata for digitised books usually does not include the the font group or even the font of the main text face. Therefore, these models would - at the moment - only be applicable if the font is recognised manually. Given the vast amount of digital copies rendered by the large-scale digitisation projects like those for VD16, VD17 and VD18, this is out of the question.

Our project addresses this problem in two ways. Firstly, we will create a tool that can identify font groups automatically, i.e. fonts which are similar to each other and thus can be used jointly for training an OCR model (Christlein / Weichselbaumer 2016). Secondly, we will create OCR-models for various font groups. In this way, we hope to significantly improve the recognition rates of OCR for early printed books. In this paper, we will present and discuss first results on automatic font group recognition.

Basis

Fortunately, we have outstanding Ground Truth data: the Gesamtkatalog der Wiegendrucke (GW) (Staatsbibliothek zu Berlin 2019a) and its side project, the Typenrepertorium der Wiegendrucke (TW) (Staatsbibliothek zu Berlin 2019b). Both were initiated by Konrad Haebler at the turn of the last century and are still maintained today at the Berlin State Library. The GW provides us with bibliographical data for all known incunabula editions (books printed in the 15th century) as well as some 15,000 digital copies from all over the world. The corresponding records in the TW list over 6,000 fonts used for these books and later editions. This Ground Truth data was painstakingly collected in over a century and was recently - thankfully - converted to a database by the Berlin State Library (Eisermann / Duntze 2014).

Method

In a first step, we have also accumulated a large body of material from our collaboration partners: The University Library of Cologne, the Herzog-August Library in Wolfenbüttel, the University Library of Heidelberg, the University Library of Erlangen, the Berlin State Library, the Göttingen State Library, the Stuttgart State Library and the Bavarian State Library. We will also be able to work soon - for the first time - with digitised copies from the British Library, which is currently scanning its large incunabula collection. All in all, we have over a million images which provide a sound basis for our goals.

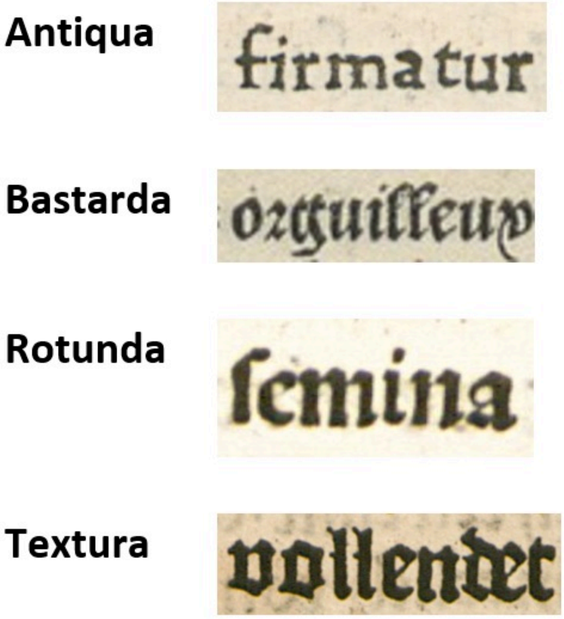
For evaluation purposes, random pages were taken out from the labelled data. We have two such subsets: validation and test. The validation data is used for tuning the classification method and evaluating it during its development, while the test data is kept until the end for an unbiased method evaluation on never seen, never used data.

A deep convolutional neural network (CNN) is used for the font recognition. To have both a robust and proven network architecture, we used one inspired by a residual network with 50 layers, also known as ResNet50 (He et al. 2016). A typical ResNet50 has layers with a large amount of neurons (from 64 to 512 in the convolutional layers), which can lead to overfitting the training data. To avoid this pitfall, we restrict the layers to having 96 neurons, with a penultimate fully-connected layer of 384 neurons. The training is done by stochastic gradient descent on batches of 64 samples, with a constant momentum of 0.9 and initial learning rate and weight decay of respectively 0.01 and 0.0005. The learning rate and weight decay are divided by 10 every 300,000 samples. The training is stopped after processing a million samples because the error stagnates.

The CNN has a receptive field of 224x224 pixels, which is insufficient for processing a whole page image at once. To identify the font of a page, we present 25 random crops from this page to the CNN, and average the results of the last linear layer (not the softmax output), then the class with maximum average value is taken. Crops full of text contain between 15 and 500 characters, depending on image resolution and text size. Typically, if a crop is misclassified (e.g., if it did not contain text), it will have little impact on the average result as the CNN is likely to produce results will low confidences.

Evaluation

We used, as training data, 280 pages with a median resolution of 2 megapixels from the 15th century containing text with fonts from four different groups: Antiqua, Bastarda, Rotunda, and Textura.



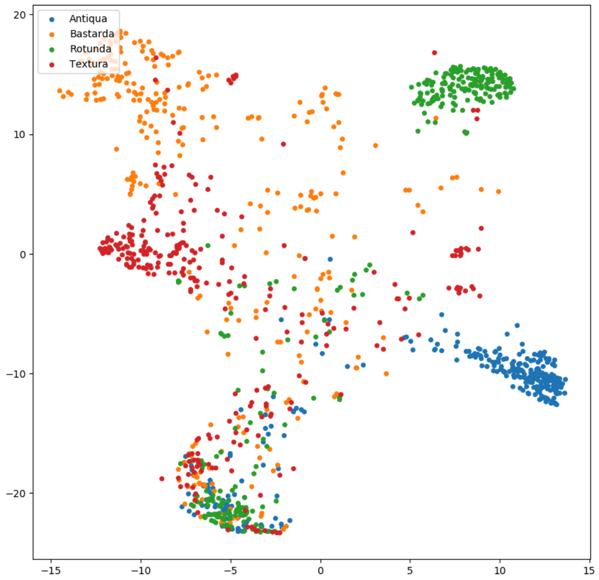
This means the CNN does not have the ability to answer something else. It is however useful to investigate what happens when pages with other fonts are given to the CNN. So as a test we provided 100 images of Fraktur - a font very closely connected to Bastarda. In addition to that, we also provided 30 images of each Greek, Hebrew and Italic - fonts that are rather different to the others. The results obtained on the test data for the four base fonts (15 pages each), as well as on the other pages, are given in the following confusion matrix:

	Antiqua	Bastarda	Rotunda	Textura
Antiqua	14			1
Bastarda		13	1	1
Rotunda			15	
Textura		1		14
Fraktur	0	92	4	4
Greek	5	20	4	1
Hebrew	3	17	6	4
Italic	1	26	2	1

Rows correspond to fonts, and columns to results given by the CNN. We can see that the test pages with the fonts known by the CNN are well classified, with an accuracy of 93%. The other fonts are more spread, but mostly classified as Bastarda.

This is a very significant result. Not only does the network recognize the connection between Bastarda and Fraktur but it also perceives the significant difference between Bastardas and the other three groups. After all, from the very beginning of typography the font group Bastarda differed considerably from other font groups. This is especially true for Texturas and Rotundas which have very uniform characteristics: Textura letters are upright, narrow and stand on crooked feet and Rotunda letters are small, curved and reveal a great contrast between thick and thin strokes. In contrast, Bastardas show much more variety - letters tend to slope forward and have flourished ascenders, yet many of them do not have these characteristics. Therefore it is very plausible that the CNN would categorize ‘unknown’ fonts as Bastardas.

This matches what can be seen from the data produced by the penultimate fully-connected layer of the CNN. As it has 384 dimensions, a t-Distributed Stochastic Neighbor Embedding (t-SNE) can be applied for visualization purpose. In the figure below, the dots correspond to individual random crops from test images. We can see five main areas. The one at the bottom corresponds to crops with little or no text, and therefore the CNN produces similar values regardless of the type group of the page. The points between this cluster and the center of the graphics might correspond to crops with text content, but not in a quantity large enough for identifying the script. Then, we have three well defined clusters for Antiqua, Rotunda, and Textura. Finally, the part corresponding to Bastarda is well spread and significantly less dense than for the other type groups. Thus, the CNN produces more variability in its penultimate layer for the Bastarda than for the other type groups, and a more important area of the feature space is considered, by the CNN, as corresponding to Bastarda. This could also explain why unseen type groups are frequently classified as belonging to the Bastarda.



Outlook

These results show that it is feasible to recognise font groups automatically. The authors are currently working on improving accuracy further and to expand the scope of the recognition tool from the 15th to the 16th-18th century. At the same time preliminary steps are taken to recognise not only font groups but exact fonts. This feature would not only make it much quicker to date and identify the printer of early modern books based on their fonts but also make this procedure much more accessible to scholars who are not highly specialised in analytical bibliography.

The source code used in this paper is available on github (https://github.com/seuretm/typegroups-classification-projection). Please note that the exact same results cannot be obtained due to the randomness of initial parameters, and that the data is currently not publicly available.

Bibliographie

Christlein, Vincent / Weichselbaumer, Nikolaus (2016): *Automatische Typenbestimmung in historischen Drucken.* Poster at DHd2016, Abstract online: http://dhd2016.de/boa-2.0.pdf [Access date 9 October 2018].

Eisermann, Falk / Duntze, Oliver (2014): *Auf der Spur der seltsamen Typen. Das digitale Typenrepertorium der Wiegendrucke*, in: Bibliotheksmagazin 3: 41–48 https://www.bsb-muenchen.de/fileadmin/imageswww/pdf-dateien/bibliotheksmagazin/BM2014-3.pdf [Access date 9. October 2018].

He et al. (2016): https://ieeexplore.ieee.org/abstract/document/7780459.

Staatsbibliothek zu Berlin (2019a): https://www.gesamtkatalogderwiegendrucke.de [Access date 11 January 2019].

Staatsbibliothek zu Berlin (2019b): https://tw.staatsbibliothek-berlin.de [Access date 11 January 2019].