Constructing a Historical Nordic Human Capital Database: An End-to-End Machine Learning Approach

Christian E. Westermann*, Christian M. Dahl†

Abstract

Automatic and robust transcription of scanned documents, especially those of lesser quality, is still a difficult area. Despite the recent and large advancements in the fields of machine learning and deep learning, the large heterogeneity in historical document layouts constrains the ability to generalize. We demonstrate that by approaching the problem from a lower level, a level where generalization is feasible, results in robust and effective transcription of historical documents. We showcase an end-to-end pipeline of novel machine learning techniques which is utilized to transcribe around 700 and 450 Norwegian and Danish historical documents respectively. The outcome is a Nordic human capital database with detailed individual-level data and with massive potential.

Keywords: Deep Learning, Character Recognition, Document Image Analysis, Tabular Data, Historical Documents

This is an early draft and all results are preliminary. Some sections and results are missing and unfinished.

*Department of Business and Economics, University of Southern Denmark, cew@sam.sdu.dk
†Department of Business and Economics, University of Southern Denmark, cmd@sam.sdu.dk

1 Introduction

Historical documents have long been a valuable source of information for historians, economists, and within the social sciences in general. They allow researchers to get a glimpse of the past and in doing so, they can shed light on past events and their consequences, raise our understanding of various demographics and even help us understand contemporary incidences (Beach et al. [2020]). While there is no end to the variety of historical documents, as well as the type of information contained within, we constrain ourselves to those of a tabular nature for several reasons. First, as econometricians, we often deal with tabular data and information that can be stratified into certain demographics is often represented in tabular form, such as census data or in this case, grade sheets. Second, approaches to automatic segmentation and transcription of historical tabular data remains, to the best of our knowledge, to be thoroughly addressed. Last, the outcome of this work results in a historic database on human capital with profound potential within the social sciences.

Lately, there have surfaced numerous methods and approaches to automatic segmentation of historical documents. The reason for multiple procedures can be attributed to the diversity of the documents and that each new type of document to be processed likely requires some notion of customization or fine-tuning of one's current approach.

In this paper, we wish to demonstrate our approach to automatic segmentation of tabular data in historical documents. Tabular data produce subtle but critical challenges different to those already solved by current methods. One such challenge is the correct assignment of table entries to their respective row and column. If one such assignment is inaccurate, it can in the worst of cases, create a domino effect of wrong assignments and the result is a full tables worth of incorrect transcriptions.

The contributions of this work is two-fold. The showcasing of the end-to-end machine learning pipeline serves as a step towards generalized approaches to tabular document image analysis. We present a pipeline of methods that operate on a lower level of the transcription process, where we believe that generalization is considerably more feasible. This includes

simpler tasks such as line identification, entry detection and transcription. The showcasing of more generic methods to historical document segmentation is analogous to the work of Ares Oliveira et al. [2018], but with the focus on and contribution to, tabular data extraction specifically. Furthermore, we present a unique Nordic historical database related to human capital. The database will contain detailed individual level information on school performance and career trajectories, just to name a few. As a result, the remainder of the paper will be structured as follows: Section 2 presents related work and frameworks, Section 3 gives an overview of the documents in question as well as a detailed explanation of the data within. Section 4 is a walk-through of the full pipeline with detailed information related to each step. Section 5 presents the results related to the pipeline as well as metrics for the individual steps and Section 6 discusses the prospects of the database by framing several research questions. Finally, Section 7 concludes.

2 Related Work

As already mentioned, many parallels can be drawn to dhSegment[1], an approach developed by Ares Oliveira et al. [2018] who demonstrate DL-based approaches to historical document segmentation where they minimize the effort required to transition from one segmentation task to the next. Ares Oliveira et al. [2018] focus on medieval documents and the applicability of a single neural network architecture on various segmentation tasks. They report their results, benchmarking against other approaches in competitions related to DIA, such as layout analysis and page detection. While such metrics also play an important role in our contribution, we demonstrate the end-to-end process from raw scan to CSV-file and thus the performance of intermediary steps is not as relevant as the overall segmentation accuracy. Additionally, also in similar spirit to their work, we wish to emphasize the generality of our methods and that they remain largely the same for all structured tabular data with only minor tweaks needed for more challenging cases.

[1] https://github.com/dhlab-epfl/dhSegment

Furthermore, in an important step towards the democratization of DIA, Shen et al. [2021] have developed a toolkit, LayoutParser[2], for the purpose of scanned document transcriptions. It is an extensive DL-based library for various tasks related to processing document images and while performance is high in general, it has yet to tackle the issues of densely populated tabular data.

DeepDeSRT is a DL-approach to detection and structure recognition of document tables (Schreiber et al. [2017]). Performance is evaluated on the ICDAR 2013 table competition, which is comprised of document image tables that are significantly more modern and consequently very different to the documents presented in this work.

Paliwal et al. [2020], while also providing an end-to-end example of a DL-based approach to table segmentation and extraction, it is similarly built and evaluated with respect to the ICDAR 2013 competition. Methods and frameworks developed for the ICDAR 2013 competition are not directly comparable to the work presented in this paper, as our documents can date several centuries back and consequently introduce issues not present in modern document tables.

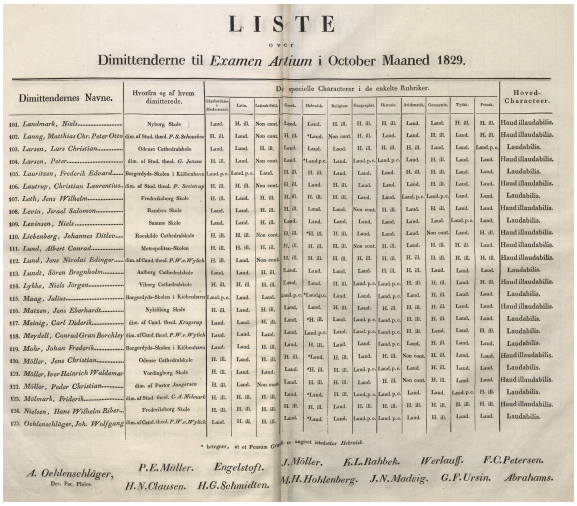
3 Data and Document Types

3.1 Data Description

The database will be comprised of information obtained from two types of Danish and Norwegian document sources: university annuals and high school graduate yearbooks. The annuals provide complete lists of grades for individual students in all subjects from high school as well as universities. The documents span several centuries, with documents from the University of Copenhagen being available as far back as the eighteenth century. Similarly, the documents from the University of Oslo has information on its students all the way back to its foundation in the beginning of the nineteenth century. While not presented in this paper,

[2]https://github.com/Layout-Parser/layout-parser

the high school graduate yearbooks include comprehensive biographies of entire cohorts of high school graduates from the 1820s to 1940. These biographies provide detailed information on the career trajectory of each individual, including where they worked and travelled, from graduation and until retirement. Figure 1 illustrates a few examples of the document images that are processed in this paper.



LISTE

over

Dimittenderne til *Examen Artium* i October Maaned 1829.

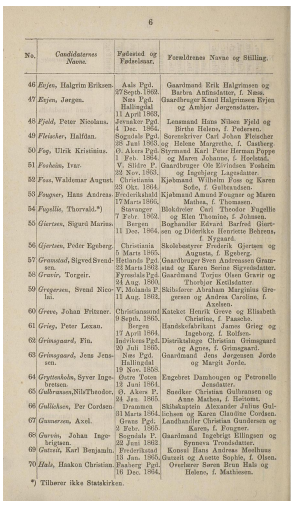
Dimittendernes Navne.	Hvorfra og af hvem dimitterede.	Hovedcharakter.
101. *Lundmark, Niels*	Nyborg Skole	Haud illaudabilis.
102. *Lang, Matthias Chr. Peter Otto*	Cin. af Stud. theol. *P. S. Schouboe*	Haud illaudabilis.
103. *Larsen, Lars Christian*	Odense Cathedralskole	Laudabilis.
104. *Larsen, Peter*	Cin. af Stud. theol. *G. Jessen*	Haud illaudabilis.
105. *Lauritzen, Frederik Edvard*	Borgerdyds-Skolen i Kiøbenhavn	Laudabilis.
106. *Lautrup, Christian Laurentius*	Cin. af Stud. theol. *P. Scavenius*	Haud illaudabilis.
107. *Leth, Jens Wilhelm*	Frederiksborg Skole	Laudabilis.
108. *Levin, Israel Salomon*	Randers Skole	Haud illaudabilis.
109. *Levinsen, Niels*	Samme Skole	Laudabilis.
110. *Lichenberg, Johannes Ditlev*	Roskilde Cathedralskole	Haud illaudabilis.
111. *Lund, Albert Conrad*	Metropolitan-Skolen	Haud illaudabilis.
112. *Lund, Jens Nicolai Eilingsø*	Cin. af Cand. theol. *P. W. Wyllich*	Haud illaudabilis.
113. *Lundt, Søren Bregnhøi*	Aalborg Cathedralskole	Laudabilis.
114. *Lykke, Niels Jørgen*	Viborg Cathedralskole	Haud illaudabilis.
115. *Maag, Julius*	Borgerdyds-Skolen i Kiøbenhavn	Laudabilis.
116. *Matzen, Jens Eberhardt*	Nykiøbing Skole	Laudabilis.
117. *Mønnig, Carl Didrik*	Cin. af Cand. theol. *Kragerup*	Laudabilis.
118. *Meydell, Conrad Gran Borchley*	Cin. af Cand. theol. *P. W. Wyllich*	Laudabilis.
119. *Møller, Johan Frederik*	Borgerdyds-Skolen i Kiøbenhavn	Laudabilis.
120. *Møller, Jens Christian*	Odense Cathedralskole	Haud illaudabilis.
121. *Møller, Iver Heinrich Wallentin*	Vordingborg Skole	Laudabilis.
122. *Møller, Peder Christian*	Cin. af Pastor *Jungersen*	Haud illaudabilis.
123. *Mølmark, Frederik*	Cin. af Stud. theol. *C. A. Nielsen*	Laudabilis.
124. *Nielsen, Hans Wilhelm Ribe*	Frederiksborg Skole	Haud illaudabilis.
125. *Oehlenschläger, Joh. Wolfgang*	Cin. af Cand. theol. *P. W. Wyllich*	Laudabilis.

A. Oehlenschläger, Dec. Fac. Philos.

P.E.Müller. *Engelstoft.* *J.Møller.* *K.L.Rahbek.* *Werlauff.* *F.C.Petersen.*

H.N.Clausen. *H.G.Schmidten.* *M.H.Hohlenberg.* *J.N.Madvig.* *G.F.Ursin.* *Abrahams.*

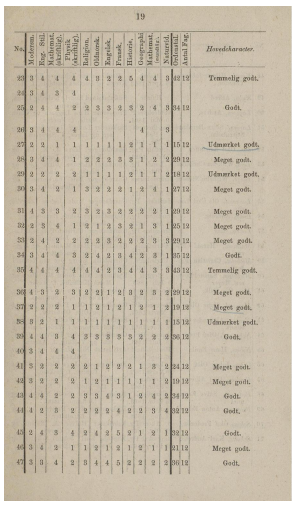
(a)



No.	Candidatarnes Navne.	Fødested og Fødselsaar.	Forældrenes Navne og Stilling.
46	*Eitjen*, Halgrim Eriksen.	Aals Pgd. 27 Septb. 1862.	Gaardmand Erik Halgrimsen og Barbra Anfinsdatter, f. Nass.
47	*Eitjen*, Jørgen.	Nes Pgd. Hallingdal 11 April 1863.	Gaardbruger Knud Halgrimsen Eitjen og Ambjør Jørgensdatter.
48	*Field*, Peter Nicolaus.	Jevanker Pgd. 4 Dec. 1864.	Lensmand Hans Nilsen Fjeld og Birthe Helene, f. Pedersen.
49	*Fleischer*, Halfdan.	Sogndals Pgd. 28 Juni 1863.	Sorenskriver Carl Johan Fleischer og Helene Margrethe, f. Qvalberg.
50	*Fog*, Ulrik Kristinius.	O. Akers Pgd. 1 Feb. 1864.	Styrmand Karl Peter Herman Poppe og Maren Johanne, f. Hofstad.
51	*Foshein*, Ivar.	V. Slidre P. 22 Nov. 1863.	Gaardbruger Ole Erlendsen Foshein og Ingebjørg Lagrelsdatter.
52	*Foss*, Waldemar August.	Christiania 23 Okt. 1864.	Kjøbmand Wilhelm Foss og Karen Sofie, f. Gulbrandsen.
53	*Fougner*, Hans Andreas.	Frederikshald 17 Marts 1866.	Kjøbmand Amund Fougner og Maren Mathea, f. Thomasen.
54	*Fugellie*, Thorvald.*)	Stavanger 7 Febr. 1863.	Blikdrivter Carl Theodor Fugellie og Elen Thomine, f. Johnsen.
55	*Gjertsen*, Sigurd Marius.	Bergen 11 Dec. 1864.	Bogholder Edvard Barfred Gjertsen og Dideriks Henriette Behrens, f. Bygaard.
56	*Gjertsen*, Fedor Egeberg.	Christiania 9 Marts 1865.	Skibsbygmester Frederik Gjertsen og Augusta, f. Egeberg.
57	*Grønsdal*, Sigvert Svendsen.	Hetlands Pgd. 23 Marts 1863.	Gaardbruger Sven Andreasen Grønsdal og Karen Sørine Sigvesdatter.
58	*Grønn*, Torgeir.	Fyresdals Pgd. 24 Aug. 1860.	Gaardmand Torjus Olsen Gravir og Thorbjør Reitlasdatter.
59	*Gregersen*, Svend Nicolai.	V. Moland P. 11 Aug. 1862.	Skibsbygger Abraham Margidius Gregersen og Andrea Caroline, f. Axelsen.
60	*Greve*, Johan Fritzner.	Christiansund 9 Septb. 1865.	Kateket Henrik Greve og Elisabeth Christine, f. Pasche.
61	*Grieg*, Peter Lexau.	Bergen 17 April 1864.	Handelsfuldmægtig James Grieg og Ingeborg, f. Rolfsen.
62	*Grimsgaard*, Fin.	Innkvarna Pgd. 20 Juli 1865.	Distriktslæge Christian Grimsgaard og Agnes, f. Grimsgaard.
63	*Grimsgaard*, Jens Jensen.	Nes Pgd. Hallingdal 19 Novbr. 1865.	Gaardmand Jens Jørgensen Jorde og Margit Jorde.
64	*Grytteboles*, Syver Ingebretsen.	Støre Toten 12 Juni 1861.	Engebret Damhougen og Petronelle Jensdatter.
65	*Gulbransen*, Nils Theodor.	Ø. Aker P. 24 Jan. 1865.	Snedker Christian Gulbransen og Anne Marhes, f. Hesloent.
66	*Gulliksen*, Per Cordsen	Drammen 31 Marts 1864.	Skibskaptein Alexander Julius Gulliksen og Karen Clandine Cordsen.
67	*Gunnerson*, Axel.	Grans Pgd. 2 Febr. 1865.	Landhandler Christian Gundersen og Karen, f. Fougner.
68	*Garvin*, Johan Ingebrigtsen.	Sogndals P. 22 Juni 1862.	Gaardmand Ingebrigt Ellingsen og Synneva Trondsdatter.
69	*Gutzeit*, Karl Benjamin.	Frederikstad 13 Jan. 1865.	Konsul Hans Andreas Meelhuus Gunzeit og Anette Sophie, f. Olsen.
70	*Hals*, Haakon Christian.	Faaberg Pgd. 16 Dec. 1864.	Overlærer Søren Hals og Helene, f. Mathiesen.

*) Tilhører ikke Statskirken.

(b)



No.	Navn.	Fødsels-Aar.	Maaned.	Dag.	Hovedcharakter.
33	3	4	4	4	12	Temmelig godt.
34	3	4	5	4	
35	2	4	4	2	12	Godt.
36	3	4	6	4	
37	2	2	1	1	15	Udmærket godt.
38	3	4	4	1	2	29	12	Meget godt.
39	2	2	2	2	1	18	12	Udmærket godt.
40	3	4	2	1	3	2	1	27	12	Meget godt.
41	3	3	3	2	3	2	3	2	2	1	20	12	Meget godt.
42	3	5	4	1	2	2	3	3	2	1	25	12	Meget godt.
43	2	4	2	2	2	3	3	2	2	3	3	29	12	Meget godt.
44	3	4	6	3	2	4	2	3	4	2	3	1	35	12	Godt.
45	4	4	4	4	4	4	2	3	4	4	3	3	43	12	Temmelig godt.
46	4	3	3	3	2	2	1	2	3	2	3	2	28	12	Meget godt.
47	2	2	2	1	1	2	1	2	1	2	1	2	19	12	Meget godt.
48	3	3	1	1	1	1	1	1	1	1	1	1	15	12	Udmærket godt.
49	4	4	3	4	3	3	3	3	2	2	3	2	36	12	Godt.
50	3	4	4	4											
51	3	2	2	2	2	3	3	3	1	2	2	2	29	12	
52	2	2	2	2	1	1	1	2	1	1	2	1	18	12	
53	3	4	2	1	3	2	2	2	1	2	4	1	27	12	
54	4	3	3	3	2	3	2	3	2	2	2	1	29	12	
55	2	3	4	1	2	1	2	3	2	1	3	1	25	12	
56	3	4	4	4	4	4	2	3	4	4	3	3	43	12	
57	3	4	4	4	4	4	2	3	4	4	3	3	43	12	
58	3	4	4	4	4	4	2	3	4	4	3	3	43	12	
59	3	4	4	4	4	4	2	3	4	4	3	3	43	12	
60	3	4	4	4	4	4	2	3	4	4	3	3	43	12	
61	3	4	4	4	4	4	2	3	4	4	3	3	43	12	
62	3	4	4	4	4	4	2	3	4	4	3	3	43	12	
63	3	4	4	4	4	4	2	3	4	4	3	3	43	12	
64	3	4	4	4	4	4	2	3	4	4	3	3	43	12	
65	3	4	4	4	4	4	2	3	4	4	3	3	43	12	
66	3	4	4	4	4	4	2	3	4	4	3	3	43	12	
67	3	4	4	4	4	4	2	3	4	4	3	3	43	12	
68	3	4	4	4	4	4	2	3	4	4	3	3	43	12	
69	3	4	4	4	4	4	2	3	4	4	3	3	43	12	
70	3	4	4	4	4	4	2	3	4	4	3	3	43	12	

(c)

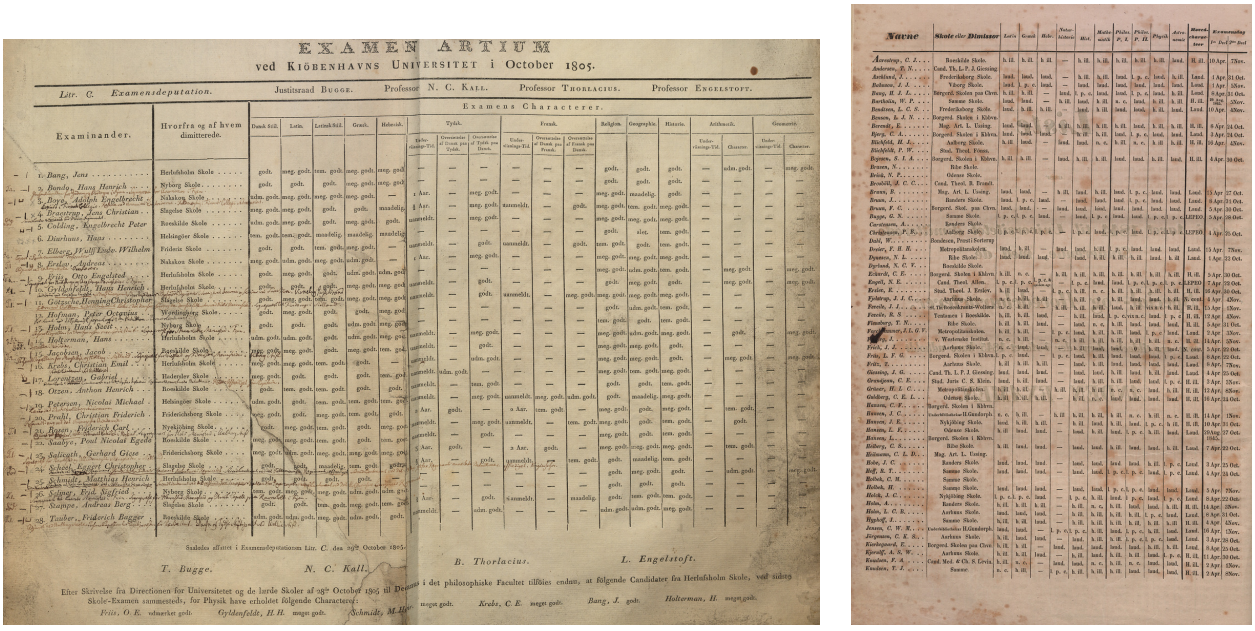
Figure 1: Images of a few of the document types that are transcribed. (a) An example of a Danish grade sheet which contains individual names, schools of graduation as well as grades in several subject, written Latin. (b) List of Norwegian student names, birthplace and birth year as well as their parents' names and occupations. (c) A Norwegian grade sheet that can be linked to the lists of names through the number identifier in the first column.

3.2 Document Types and Difficulties

The images displayed in Figure 1 represent some of the majority types of the documents available to us. While the layout does undergo changes over time, as school subjects are replaced or updated, the overall document layout stays largely the same. As long as there is a table outline available to us, our method can quickly be tweaked to accommodate new table structures.

There are several difficulties related to the document types illustrated in Figure 1 and the images shown are some of the neater examples. Comparatively, all three examples have

no clear separation of rows, which is especially prominent in document (b) where entries occasionally span several rows and the compactness makes separation non-trivial. (a) is scanned from a book, which results in a page break in the middle which in worse cases makes close columns almost unreadable or significantly offsets the pages from each other. Figure 2 displays a couple of the most difficult cases where the consequences of time are clear. There is an abundance of handwritten scribbles that intersects regions that we care about which makes both the entry detection as well as transcription processes extremely hard without tailored models. Furthermore, bends, different shading as well as various stains do not reduce the difficulty.



(a) (b)

Figure 2: Examples of difficult documents, both Danish grade sheets.

4 Pipeline

Analogous to Ares Oliveira et al. [2018], we give a brief overview of the outline followed by more detailed descriptions in the subsections. The pipeline can be decomposed into the following steps, accommodated by the flowchart of Figure 3:

1. The initial step entails a manual specification of a template (or reference) image as well as an overlay. The template specifies what point cloud we are interested in finding downstream and is usually chosen to be the table outline. The overlay comprises the specific areas of the documents that we are interested in. We define these areas to be the columns in the examples used in this paper but, given a static row and column count, could also be the individual table entries. After this step is completed, the pipeline is run on the remaining documents of the given layout type.

2. The second step involves locating the point cloud specified in the template. We use a neural network for semantic segmentation, which is the process of assigning each individual pixel to a specific class. The table outline is identified and the resulting classified pixels are used as the point cloud.

3. We use FilterReg, a point-set registration algorithm of Gao and Tedrake [2018], to align the template point cloud with the identified point cloud and consequently acquire the transformation parameters responsible for this alignment.

4. We can then apply the inverse of the transformation to the given document in question, transforming it to fit the specified overlay of step 1. Areas of interest can then easily be extracted.

5. Finally, the extracted areas can be transcribed.

   If the document layout has varying columns or rows, a few extra steps are needed prior to and post transcription:

   - Segment full columns or rows depending on layout as opposed to individual entries.
   - Apply Fast-RCNN of Girshick [2015] to extract entries.
   - Transcribe entries.
   - Assign entries to respective rows and columns using KMeans-clustering.

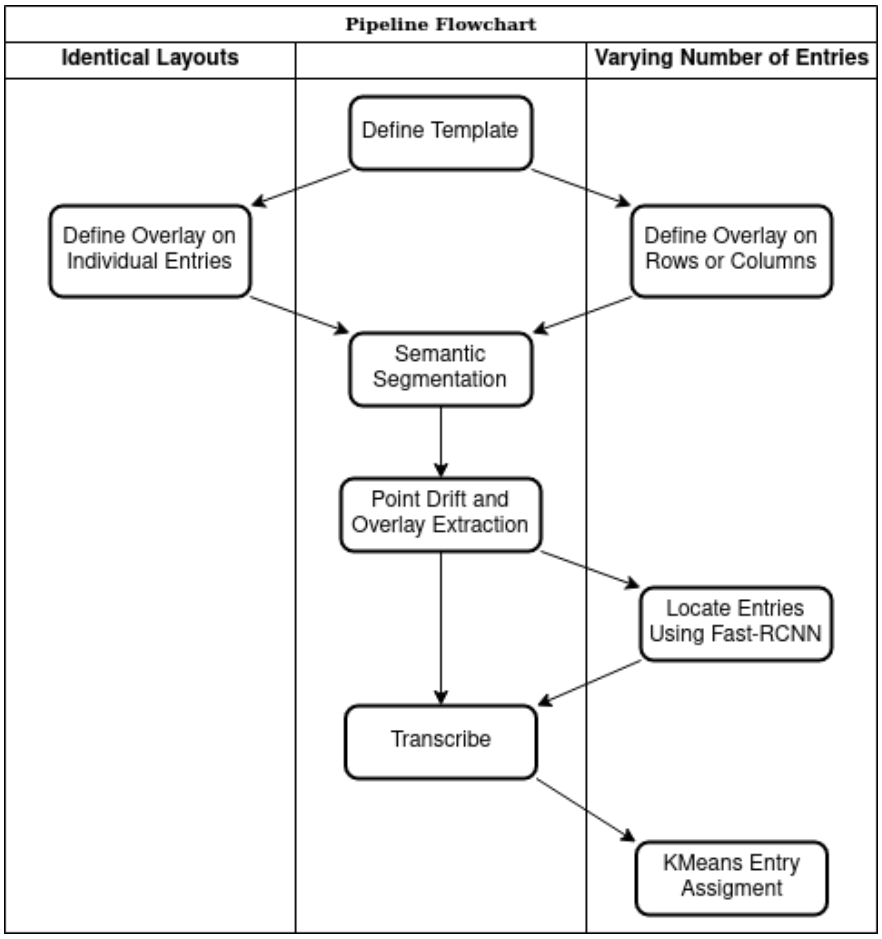
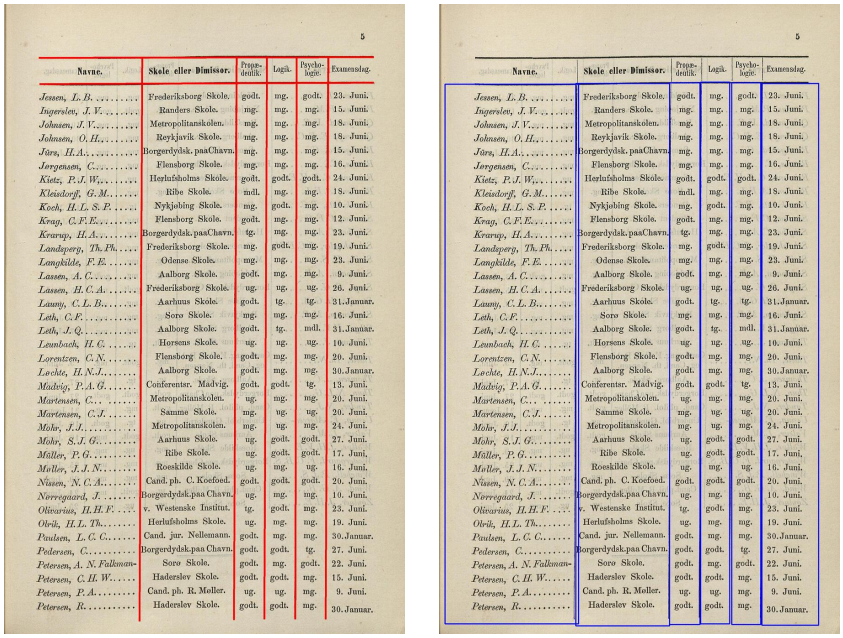


Figure 3: Overview of the described pipeline. Middle lane represents the operations common to both cases, whereas the left lane is for documents with identical layouts and the right lane is for documents with a varying number of rows or columns. We can easily see that having identical layouts significantly reduce the complexity of the problem.

4.1 Template and Overlay Definition

Defining the template and overlay is simple matter of highlighting the table outline as well as the areas of interest. We have developed a tool named TableParser, named after the work of Shen et al. [2021], which makes the manual part quite painless. As of this writing, TableParser has yet to become publicly available, but it abstracts away the majority of the pipeline process from the user, such that the only step to worry about is the definition of template and overlay. Figure 4 displays an example of a defined template and overlay for one type of Danish grade sheet.

In the absence of a tool like TableParser, you would manually define the corners of the



(a) Template (b) Overlay

Figure 4: An example of a template and overlay for a specific layout. The pixels making up the red lines on the template will act as the point cloud that we wish to locate on the target documents. In the overlay, we have highlighted the columns as the areas of interest.

rectangles in the overlay.

4.2 Semantic Segmentation and Point Cloud Detection

Having defined the template point cloud in the previous step, we now have to find a way to locate the corresponding point cloud on the image of interest. For this, we use semantic segmentation, assigning each pixel of the target image to one of three classes: background, text (scribbles) or outline. While we are solely interested in the table outline, extra classes can assist the network in establishing clearer decision boundaries and therefore make line detection more accurate.

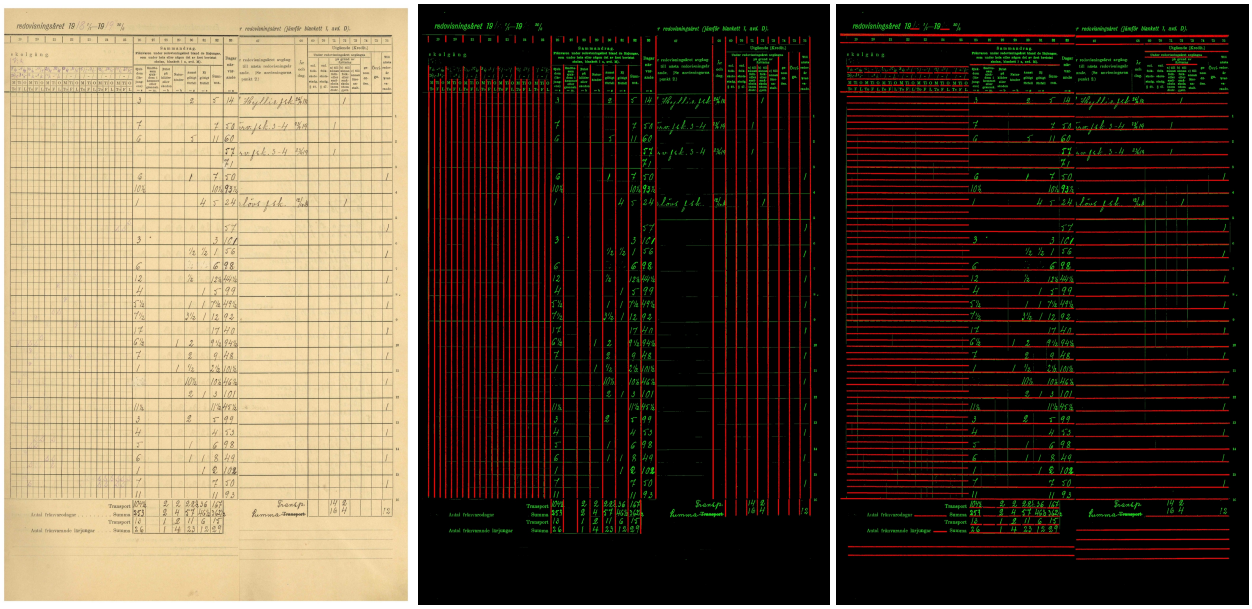
4.2.1 Network Architecture

Neural network architectures with a U-like structure have proven themselves to excel within the field of semantic segmentation. Since first introduced by Ronneberger et al. [2015] for the

purpose of biomedical image segmentation, many adaptations have been made to the network structure, one of which is DeeplabV3+ by Chen et al. [2018], which is the architecture used in this paper. We define two separate models of identical architecture, one responsible for vertical line identification and the other is responsible for horizontal line identification.

4.2.2 Training Data

The training data consists of a great variety of historical documents. Figure 5 illustrates one such document, in the form of a Swedish grade sheet.



(a) Original (b) Vertical Mask (c) Horizontal Mask

Figure 5: An illustration of an original document and its corresponding vertical and horizontal masks, with the classes for lines, text and background defined by red, green and black respectively.

Prior to being fed to the neural network, we split the images into 448×448 patches which are then further resized to 224×224 before entering the network . The rationales behind this decision are several. It is not uncommon for the document scans being several thousands in dimensions and simply resizing risks deteriorating important features in the process, such as thin lines. But most importantly, working with patches increases the generalization capabilities of the network. The amount of variance possible with respect to lines in a

patch is significantly smaller than that of a full image and we thus increase the chance of out-of-sample patches being within the distribution of the training data. Furthermore, we can randomly sample patches from a given image multiple times and coupled with random augmentations, we can quickly gather a substantial training set. Thus, when applying the pipeline to a target image, we similarly split it into patches, predict each individual patch, and then stitch them back together. Figure 6 illustrates examples of training images and their corresponding labels.

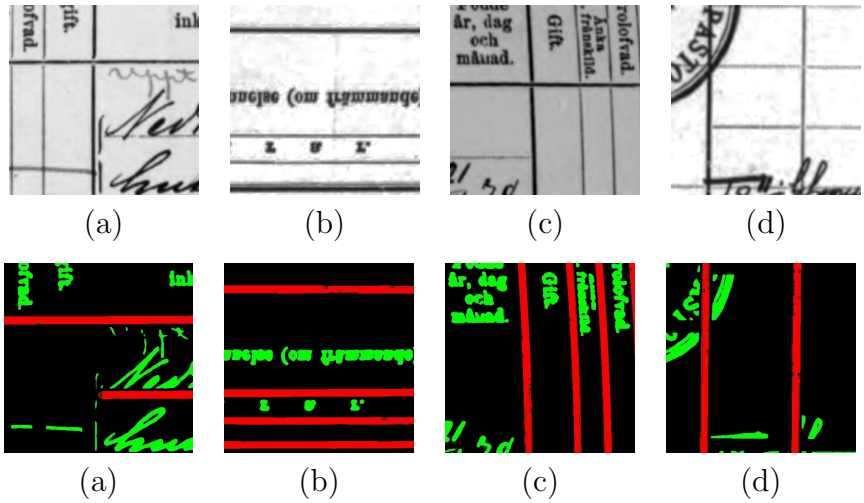


Figure 6: Input training images and their corresponding labels. Top row represents the raw input patches while the bottom row consists of the label counterparts. The masks are colored red, green and black for lines, text and background respectively. (a) and (b) are used as input to the horizontal model and (c) an (d) are used for the vertical model. (b) has additionally been augmented with a 180 degree rotation.

We apply several augmentations to the image patches and their labels. Since we are solely interested in lines, orientation and readability of text is of no large concern, and we can thus apply augmentations that would otherwise be nonsensical, such as upside-down text. As we strive for a generalizable approach to tabular document segmentation and transcription, we leave out all the document types of the Nordic database and consequently this paper, from the training set such that results presented are entirely out-of-sample. Patches included in the training set are thus collected from other historical documents that we have access to.

4.2.3 Results

Figure 7 showcases the result from applying the trained semantic segmentation network on a Danish grade sheet. While the line identification is the only thing that matters with respect to this approach, we still note high and precise classification accuracy of the other classes.

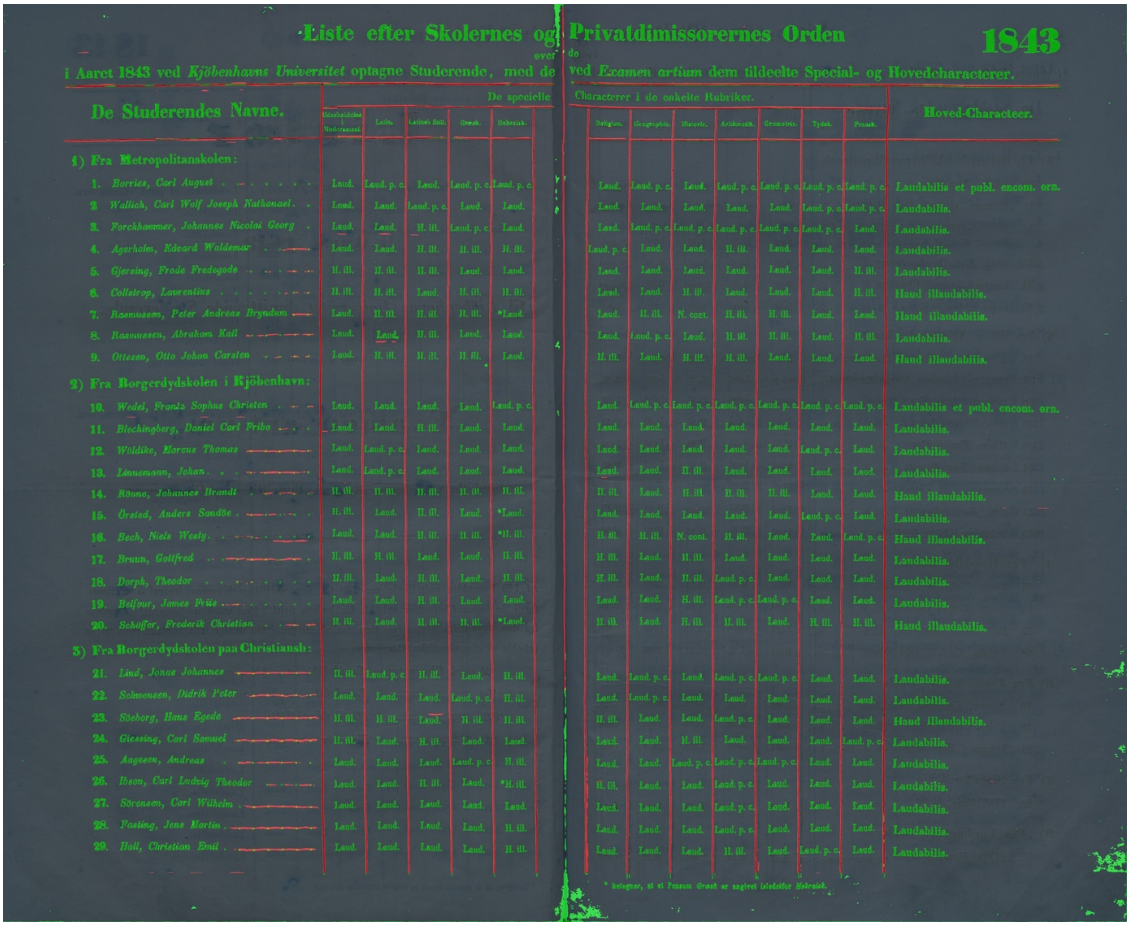


Figure 7

We can then easily extract the point cloud corresponding to classified line pixels.

4.3 Point Drift and Overlay Extraction

With the point cloud corresponding to line pixels in the target document identified, we can proceed and apply FilterReg. It is a point-set registration algorithm and as noted by the authors, Gao and Tedrake [2018], it can be considered an inverse of coherent point

drift by Myronenko and Song [2009]. Points are forced to move coherently, preserving their topological structure. This results in robust fits, even in the absence of identified points as well as in the case of over-identification, making it an ideal algorithm for our purposes. We do not know the number of points identified on a given document a priori, and it is therefore impossible to specify the exact number of points to look for. Furthermore, damaged or otherwise deteriorated documents might have table outlines unrecognizable to the semantic segmentation network which the coherent constraint alleviates.

With the template point cloud $\mathbf{X}$ and the identified point cloud $\mathbf{Y}$ on the given target document, we seek the mapping

$$\Delta\theta : \mathbf{X} \rightarrow \mathbf{Y}$$

where in accordance with the authors, we refer to $\Delta\theta$ as the motion parameters. They use the expectation maximization (EM) algorithm by Dempster et al. [1977], to iteratively align $\mathbf{X}$ with $\mathbf{Y}$. Having acquired $\Delta\theta$, we can apply its inverse, $\Delta\theta^{-1}$, to the target image which results in the columns aligning with the ones that we defined in the overlay. Since we defined the overlay on the same document as the template, we know the coordinates of the rectangles enclosing the areas of interest and we can easily extract the fields.

4.4 Locating Entries with Region Proposal Networks

After columns have been extracted using the column lines identified by the UNet, row entries are located using a Faster Region-based Convolutional Network (Faster-RCNN) by Ren et al. [2016]. By restricting entry localization to columns only, we reduce risks significantly, including the risk of accidentally targeting noise as well as the subsequent struggle of ordering and assigning located entries to the right columns and rows. The same rationale regarding using cutouts for the training of UNet applies here as well. We hypothesize that generalizing to columns is more feasible than generalizing to full sized document tables. We train the network to localize rows of text, everything from single characters to full names.

4.4.1 Data and Implementation

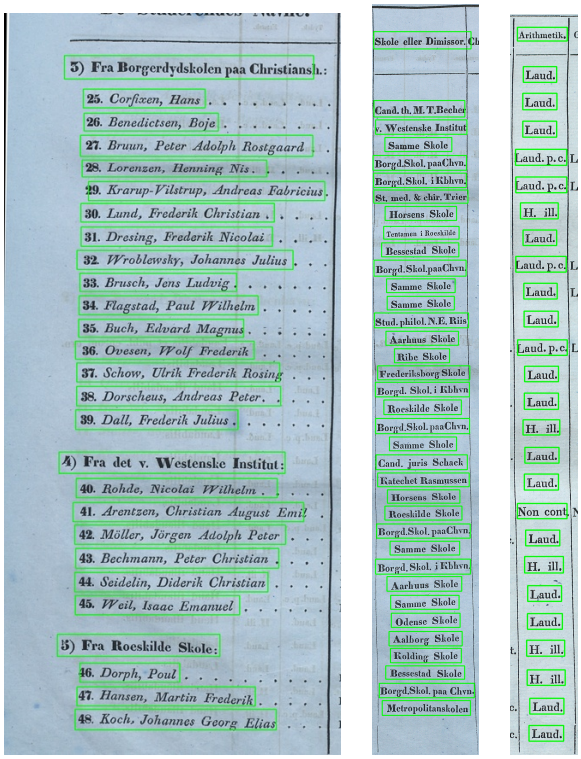
The network is pre-trained on the COCO dataset (Lin et al. [2015]) and then fine-tuned on our own data. We use labelImg[3], an open-source graphical image annotation tool made for labeling datasets for the task of object detection. Similarly to the UNet data, readability post-augmentation is not a concern as detection is the main objective. As such, we can apply horizontal and vertical flipping as well as rotations and shears. We use a non-maximum suppression threshold of 0.2, meaning that we allow boxes to overlap by 20% and otherwise, the box with the lowest prediction confidence is deleted. As opposed to standard object detection tasks, we expect few to no cases of any overlap but the threshold serves as a buffer as the network tends to predict multiple boxes on the same row, but with a much lower confidence attached. Figure 8 shows an example of three columns with the predicted bounding boxes drawn, after applying non-maximum suppression.

4.5 Transcription and Row Assignment

After acquiring the bounding boxes of every entry, we utilise Tesseract (Smith [2007]) for transcription. Tesseract often requires extensive preprocessing of input images before the transcription, including resizing, conversion from color image to grayscale and binarizing (turning the image black and white). It takes some experimentation to figure out the optimal preprocessing but the precise bounding boxes of the previous step boosts performance significantly, and simple resizing was sufficient in the majority of cases. In order to not rely on the region proposal network to not make any mistakes, we utilise K-Means clustering to ensure that entries are assigned the correct rows. Even given no mistakes, if a name of a person or school were to span several lines, counting the bounding boxes to determine the rows would not be feasible. The K-Means process can be described as follows:

1. Determine the correct number of rows by issuing a majority vote of the number of

[3]https://github.com/tzutalin/labelImg



(a) (b) (c)

Figure 8: Examples of row predictions. Columns are not necessarily from the same grade sheet. (a) Graduate names. (b) Names of schools or private teachers. (c) Grades written in Latin and achieved in arithmetic.

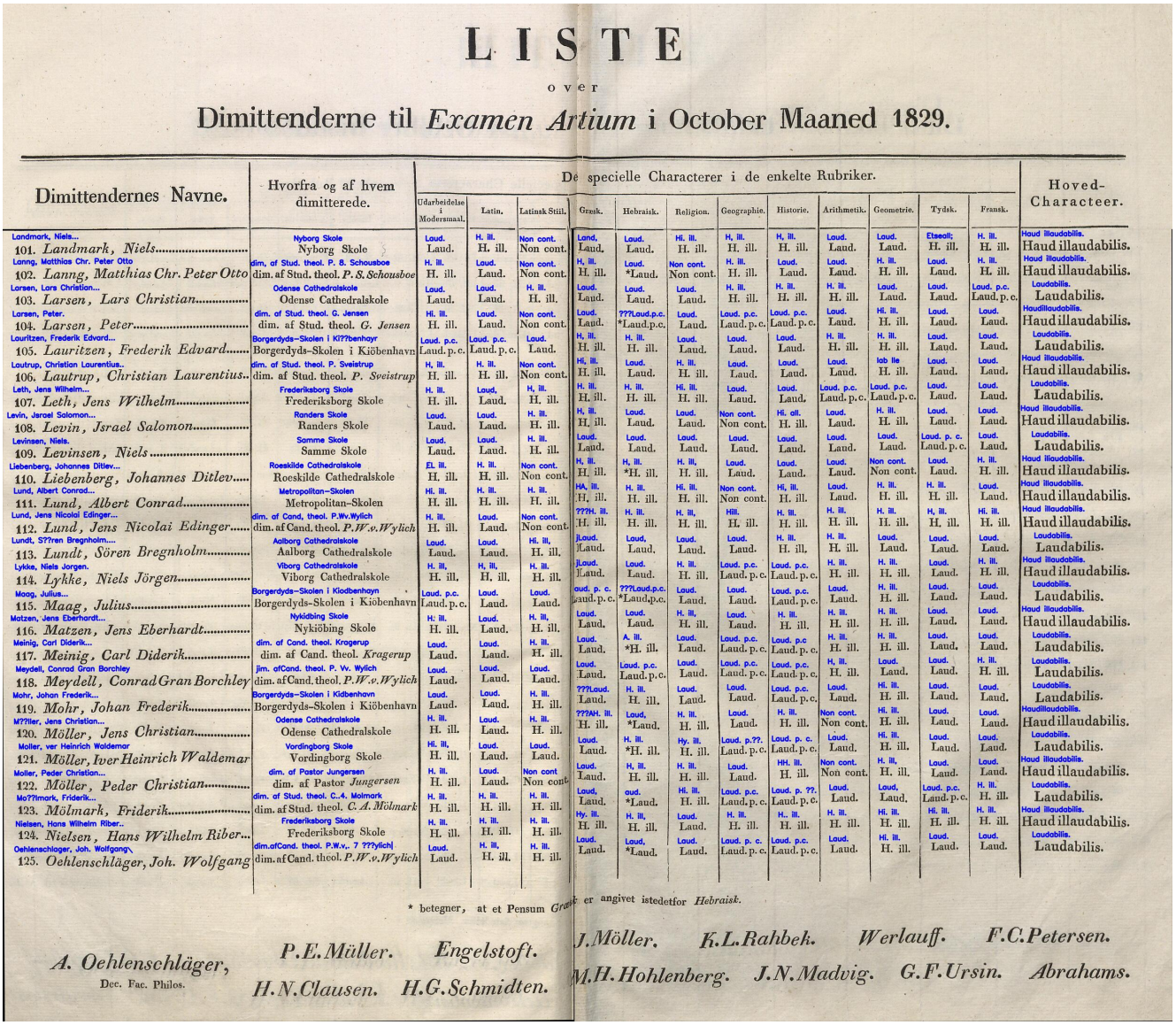
bounding boxes in each column

2. Using the columns that have the correct number of rows, create a cluster for each row based on the y-coordinate of the centers of the boxes

3. Iterate over all predicted bounding boxes and assign them to their respective clusters based on their center y-coordinate

Figure 9 illustrates a fully processed Danish grade sheet with the corresponding transcriptions positioned at their respective counterpart. While not shown on the image to minimize visual clutter, Tesseract also provides word-level confidences with respect to each prediction. This means that we actually get a measure of uncertainty related to every single transcription, that would otherwise be unobtainable using human operators. Such measures

can then be taken into account when dealing with the subsequent analysis of the data.



LISTE

over

Dimittenderne til *Examen Artium* i October Maaned 1829.

Dimittendernes Navne.	Hvorfra og af hvem dimitterede.	De specielle Characterer i de enkelte Rubriker.												Hoved-Characteer.
		Udarbeidelse i Modersmaal.	Latin.	Latinsk Stil.	Græsk.	Hebraisk.	Religion.	Geographie.	Historie.	Arithmetik.	Geometrie.	Tydsk.	Fransk.	
101. *Landmark, Niels*	Nyborg Skole	Laud. Laud.	H. ill. H. ill.	Non cont. Non cont.	Laud. Laud.	Laud. Laud.	Hi. ill. H. ill.	H. ill. H. ill.	H. ill. H. ill.	Laud. Laud.	Laud. Laud.	Etseal; H. ill.	H. ill. H. ill.	Haud illaudabilis. Haud illaudabilis.
102. *Lanng, Matthias Chr. Peter Otto*	dim. af Stud. theol. *P. S. Schousboe*	H. ill. H. ill.	Laud. Laud.	Non cont. Non cont.	H. ill. H. ill.	Laud. *Laud.	Non cont. Non cont.	H. ill. H. ill.	Laud. Laud.	H. ill. H. ill.	H. ill. H. ill.	Laud. Laud.	H. ill. H. ill.	Laudabilis. Laudabilis.
103. *Larsen, Lars Christian*	Odense Cathedralskole	Laud. Laud.	Laud. Laud.	H. ill. H. ill.	Laud. Laud.	Laud. Laud.	Laud. Laud.	H. ill. H. ill.	H. ill. H. ill.	H. ill. H. ill.	Laud. Laud.	Laud. Laud.	Laud. p.c. Laud. p. c.	Haudillaudabilis. Haudillaudabilis.
104. *Larsen, Peter*	dim. af Stud. theol. *G. Jensen*	Hi. ill. H. ill.	Laud. Laud.	Non cont. Non cont.	Laud. ???Laud.p.c.	Laud. *Laud.p.c.	Laud. p.c. Laud. p. c.	Laud. p.c. Laud. p. c.	Laud. p.c. Laud. p. c.	Laud. Laud.	Hi. ill. H. ill.	Laud. Laud.	Laud. Laud.	Laudabilis. Haud illaudabilis.
105. *Lauritzen, Frederik Edvard*	Borgerdyds-Skolen i Kiöbenhavn	Laud. p.c. Laud. p. c.	Laud. p.c. Laud. p. c.	Laud. Laud.	H. ill. H. ill.	H. ill. H. ill.	Laud. Laud.	Laud. Laud.	Laud. Laud.	H. ill. H. ill.	H. ill. H. ill.	Laud. Laud.	Laud. Laud.	Laudabilis. Haud illaudabilis.
106. *Lautrup, Christian Laurentius*	dim. af Stud. theol. *P. Sveistrup*	H. ill. H. ill.	H. ill. H. ill.	Non cont. Non cont.	Hi. ill. H. ill.	Laud. Laud.	H. ill. H. ill.	Laud. Laud.	Laud. Laud.	Laud. Laud.	Iab lle H. ill.	Laud. Laud.	Laud. Laud.	Laudabilis. Haud illaudabilis.
107. *Leth, Jens Wilhelm*	Frederiksborg Skole	H. ill. H. ill.	Laud, Laud.	H, ill. H. ill.	H. ill. H. ill.	H. ill. H. ill.	H. ill. H. ill.	Laud. Laud.	Laud. Laud.	Laud. p.c. Laud. p. c.	Laud. p.c. Laud. p. c.	Laud. Laud.	Laud. Laud.	Laudabilis. Haud illaudabilis.
108. *Levin, Israel Salomon*	Randers Skole	Laud. Laud.	Laud. Laud.	H. ill. H. ill.	H. ill. H. ill.	Laud. Laud.	Laud. Laud.	Non cont. Non cont.	Hi. ill. H. ill.	Laud. Laud.	H. ill. H. ill.	Laud. Laud.	Laud. Laud.	Haud illaudabilis. Haud illaudabilis.
109. *Levinsen, Niels*	Samme Skole	Laud. Laud.	Laud. Laud.	H. ill. H. ill.	Laud. Laud.	Laud. Laud.	Laud. Laud.	Laud. Laud.	Laud. Laud.	Laud. Laud.	Laud. p. c. Laud. p. c.	Laud. Laud.	Laud. Laud.	Laudabilis. Laudabilis.
110. *Liebenberg, Johannes Ditlev*	Roeskilde Cathedralskole	EL ill. H. ill.	H. ill. H. ill.	Non cont. Non cont.	H, ill. H, ill.	H, ill. *H. ill.	H. ill. H. ill.	Laud. Laud.	Laud. Laud.	Laud. Laud.	Non cont. Non cont.	Laud. Laud.	H. ill. H. ill.	Haud illaudabilis. Haud illaudabilis.
111. *Lund, Albert Conrad*	Metropolitan-Skolen	Hi. ill. H. ill.	H. ill. H. ill.	H. ill. H. ill.	HA. ill. H. ill.	H. ill. H. ill.	Hi. ill. H. ill.	Non cont. Non cont.	Hi. ill. H. ill.	Laud. Laud.	H. ill. H. ill.	H. ill. H. ill.	Laud. Laud.	Haud illaudabilis. Haud illaudabilis.
112. *Lund, Jens Nicolai Edinger*	dim. af Cand. theol. *P. W. v. Wylich*	H. ill. H. ill.	Laud. Laud.	Non cont. Non cont.	???H. ill. H. ill.	H. ill. H. ill.	H. ill. H. ill.	Hill. H. ill.	H. ill. H. ill.	H. ill. H. ill.	H. ill. H. ill.	H. ill. H. ill.	Hi. ill. H. ill.	Haud illaudabilis. Haud illaudabilis.
113. *Lundt, Sören Bregnholm*	Aalborg Cathedralskole	Laud. Laud.	Laud. Laud.	Hi. ill, H. ill,	jLaud. jLaud.	Laud, Laud.	Laud. Laud.	Laud. Laud.	H. ill, H. ill,	H. ill. H. ill.	Laud. Laud.	Laud. Laud.	Laud. Laud.	Laudabilis. Laudabilis.
114. *Lykke, Niels Jörgen*	Viborg Cathedralskole	H. ill, H. ill.	H, ill. H. ill.	H. ill. H. ill.	jLaud. jLaud.	Laud. Laud.	H. ill. H. ill.	Laud. p.c. Laud. p. c.	Laud. p.c. Laud. p. c.	H. ill. H. ill.	Laud. Laud.	Laud. Laud.	H. ill. H. ill.	Haud illaudabilis. Haud illaudabilis.
115. *Maag, Julius*	Borgerdyds-Skolen i Kiödbenhavn	Laud. p.c. Laud. p. c.	Laud. Laud.	Laud. Laud.	oud. p. c. Laud. p. c.	???Laud.p.c. *Laud.p.c.	Laud. Laud.	Laud. Laud.	Laud. p.c. Laud. p. c.	Laud. Laud.	H. ill. H. ill.	Laud. Laud.	Laud. Laud.	Laudabilis. Laudabilis.
116. *Matzen, Jens Eberhardt*	Nykiöbing Skole	H. ill. H. ill.	Laud. Laud.	H. ill, H. ill.	Laud. Laud.	Laud. Laud.	H. ill, H. ill.	Laud. Laud.	H. ill. H. ill.	H. ill. H. ill.	H. ill. H. ill.	Laud. Laud.	Laud. Laud.	Haud illaudabilis. Haud illaudabilis.
117. *Meinig, Carl Diderik*	dim. af Cand. theol. *Kragerup*	Laud. Laud.	Laud. Laud.	H. ill. H. ill.	Laud. Laud.	A. ill. *H. ill.	Laud. Laud.	Laud. p.c. Laud. p. c.	Laud. p.c Laud. p. c.	H. ill. H. ill.	H. ill. H. ill.	Laud. Laud.	H. ill. H. ill.	Laudabilis. Laudabilis.
118. *Meydell, Conrad Gran Borchley*	jim. afCand. theol. *P. W. v. Wylich*	Laud. Laud.	Laud. Laud.	Laud. Laud.	Laud. Laud.	Laud. p.c. Laud. p. c.	Laud. Laud.	Laud. p.c. Laud. p. c.	Laud. p.c. Laud. p. c.	H. ill. H. ill.	Laud. Laud.	Laud. Laud.	Laud. Laud.	Laudabilis. Laudabilis.
119. *Mohr, Johan Frederik*	Borgerdyds-Skolen i Kiöbenhavn	Laud. Laud.	Laud. Laud.	H. ill. H. ill.	???Laud. Laud.	H. ill. H. ill.	Laud. Laud.	Laud. Laud.	Laud. p.c. Laud. p. c.	Laud. Laud.	Hi. ill. H. ill.	Laud. Laud.	Laud. Laud.	Laudabilis. Haudillaudabilis.
120. *Möller, Jens Christian*	Odense Cathedralskole	H. ill. H. ill.	Laud. Laud.	H. ill. H. ill.	???AH. ill. H. ill.	Laud, *Laud.	H. ill. H. ill.	Laud. Laud.	H. ill. H. ill.	Non cont. Non cont.	Hi. ill. H. ill.	Laud. Laud.	Laud. Laud.	Haudillaudabilis. Laudabilis.
121. *Möller, Iver Heinrich Waldemar*	Vordingborg Skole	Hi. ill. H. ill.	Laud. Laud.	Laud. Laud.	Laud. Laud.	H. ill. *H. ill.	Hy. ill. H. ill.	Laud. p.?? Laud. p. c.	Laud. p. c Laud. p. c.	Laud. Laud.	H. ill. H. ill.	Laud. Laud.	Laud. Laud.	Laudabilis. Haud illaudabilis.
122. *Möller, Peder Christian*	dim. af Pastor *Jungersen*	H. ill. H. ill.	Laud. Laud.	Non cont. Non cont.	Laud. Laud.	H. ill. H. ill.	H. ill. H. ill.	Laud. Laud.	HH. ill. H. ill.	Non cont. Non cont.	H. ill. H. ill.	Laud. Laud.	Laud. Laud.	Haud illaudabilis. Haudillaudabilis.
123. *Mölmark, Friderik*	dim. af Stud. theol. *C. A. Mölmark*	H. ill. H. ill.	H. ill. H. ill.	H. ill. H. ill.	Laud, Laud.	aud. *Laud.	Hi. ill. H. ill.	Laud. p.c. Laud. p. c.	Laud. p. ?? Laud. p. c.	Laud, Laud.	Laud. p.c. Laud. p. c.	Laud. p.c. Laud. p. c.	H. ill. H. ill.	Laudabilis. Laudabilis.
124. *Nielsen, Hans Wilhelm Riber*	Frederiksborg Skole	H. ill. H. ill.	H. ill. H. ill.	H. ill. H. ill.	Hy. ill. H. ill.	H. ill. H. ill.	Laud. Laud.	H. ill. H. ill.	H.. ill. H. ill.	H. ill. H. ill.	Hi. ill. H. ill.	H. ill. H. ill.	H. ill. H. ill.	Haud illaudabilis. Haudillaudabilis.
125. *Oehlenschläger, Joh. Wolfgang*	dim. afCand. theol. *P. W. v. Wylich*	Laud. Laud.	H. ill. H. ill.	H. ill. H. ill.	Laud. Laud.	Laud, *Laud.	Laud. Laud.	Laud. p. c. Laud. p. c.	Laud. p.c. Laud. p. c.	Laud. Laud.	Hi. ill. H. ill.	Laud. Laud.	Laud. Laud.	Laudabilis. Laudabilis.

* betegner, at et Pensum *Græsk* er angivet istedetfor *Hebraisk*.

A. Oehlenschläger,
Dec. Fac. Philos.

P.E.Müller. *Engelstoft.* *H.N.Clausen.* *H.G.Schmidten.*

J.Möller. *K.L.Rahbek.* *Werlauff.* *F.C.Petersen.* *M.H.Hohlenberg.* *J.N.Madvig.* *G.F.Ursin.* *Abrahams.*

Figure 9: Example of a transcribed Danish grade sheet.

At the time of this writing, the pipeline has been used to transcribe approx. 700 Norwegian documents yielding around 7000 observations as well as approx. 450 Danish documents, adding close to 6300 observations to the database.

5 Results

Detailed results TBA

6 Prospects of Database

The documents subject to the presented pipeline, represent unique individual level information on grades achieved in individual subjects at various levels of education. They encompass several Nordic countries and range from the beginning of the nineteenth century until the Second World War. This period turned out to play an important role to the development of that region and the database can be used for exhaustive research of human capital and economic growth.

The database allow us to investigate several important research questions based on all new data, such as:

- **Construction of a new measure of human capital.** Our measure of human capital could be presented as both an average measure, as well as decomposed into for example knowledge of humanities, knowledge of science, diligence (grades), skills acquired while travelling, through practice, etc.

- **Was a quantity-quality trade-off behind the demographic transition?** The quantity-quality trade-off is considered to be a driving factor behind the fertility decline experienced during the demographic transition, and essentially states that families chose to have fewer but better educated children as the return to education increases. The empirical evidence is inconclusive, however. Existing studies on individual level data are based on rather specific subsamples of the population (Klemp and Weisdorf [2018] and Clark and Cummins [2015]), using census data but covering only a very limited time period (Fernihough [2017]), or using modern data and thus do not cover the demographic transition (Angrist et al. [2010]).

- **What was the role of human capital for development? And what was the role of the enlightenment, as measured by "the upper tail of human knowledge"?** Increased human capital and "good quality" basic, technical and higher education systems ("upper tail knowledge") have been found to have positive effects on "industrial development" and economic growth. However, constructed measures of average years of education, school enrolment rates, or (average) years of education are used as approximations. Squicciarini and Voigtländer [2015], for example, use subscriptions to the French Encyclopédie.

- **What is the relationship between health and education?** There is a strong positive correlation between education or social status and health, across countries and time periods. Studies on historical data often use height as a measure of health, but for restricted samples of e.g. military recruits (Bailey et al. [2016]) and relate this to socio-economic status indicated by occupation.

- **What was the "quality" of emigrants relative to those who chose to stay in the Nordic countries during the era of mass migration before the First World War?** Were these migrants positively or negatively selected from the general population? Put differently, was there a "brain drain" from the Nordic countries? Contrary to Abramitzky et al. [2012], for example, we would not have to rely on matching between US and Scandinavian censuses to identify migrants, which ultimately implies a relatively small sample size.

- **What is the role of human capital for social mobility, and how did the extent of mobility in terms of human capital outcomes change over time?** Was education restricted to only a few segments of society, or was it accessible to most people? Did education lead to work opportunities? Existing studies rely on small population samples or larger samples where the generational linkages are less accurate, for example for "families" rather than individuals (Clark and Cummins [2015]).

7 Conclusion

This paper demonstrated an end-to-end transcription pipeline of historical data exhibiting a tabular structure. We show a generic approach to tabular data with a generalizable method related to semantic segmentation. In the near future, we additionally seek to generalize the entry detection approach of Fast-RCNN. We demonstrated that due to the extreme heterogeneity in document layouts, it might be more feasible to approach the issue from a lower level, using patches and detecting entries column-wise (row-wise). Furthermore, we emphasized the importance of democratizing work within DIA as it can unlock data that could prove valuable not only to economists and historians. Lastly, we posed several important questions that can be thoroughly researched as a result of the database.

References

Brian Beach, Karen Clay, and Martin H Saavedra. The 1918 influenza pandemic and its lessons for covid-19. Working Paper 27673, National Bureau of Economic Research, August 2020. URL `http://www.nber.org/papers/w27673`.

Sofia Ares Oliveira, Benoit Seguin, and Frederic Kaplan. dhsegment: A generic deep-learning approach for document segmentation. *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Aug 2018. doi: 10.1109/icfhr-2018.2018.00011. URL `http://dx.doi.org/10.1109/ICFHR-2018.2018.00011`.

Zejiang Shen, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li. Layoutparser: A unified toolkit for deep learning based document image analysis. *arXiv preprint arXiv:2103.15348*, 2021.

Sebastian Schreiber, Stefan Agne, Ivo Wolf, Andreas Dengel, and Sheraz Ahmed. Deepdesrt: Deep learning for detection and structure recognition of tables in document images. In *2017*

14th IAPR International Conference on Document Analysis and Recognition (ICDAR), volume 01, pages 1162–1167, 2017. doi: 10.1109/ICDAR.2017.192.

Shubham Paliwal, Vishwanath D, Rohit Rahul, Monika Sharma, and Lovekesh Vig. Tablenet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images, 2020.

Wei Gao and Russ Tedrake. Filterreg: Robust and efficient probabilistic point-set registration using gaussian filter and twist parameterization. *CoRR*, abs/1811.10136, 2018. URL `http://arxiv.org/abs/1811.10136`.

Ross Girshick. Fast r-cnn, 2015.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation, 2018.

Andriy Myronenko and Xubo B. Song. Point-set registration: Coherent point drift. *CoRR*, abs/0905.2635, 2009. URL `http://arxiv.org/abs/0905.2635`.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. ISSN 00359246. URL `http://www.jstor.org/stable/2984875`.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

Ray Smith. An overview of the tesseract ocr engine. In *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition*, pages 629–633, Washington, DC, USA, 2007. IEEE Computer Society. ISBN 0-7695-2822-8. URL http://www.google.de/research/pubs/archive/33418.pdf.

Marc Klemp and Jacob Weisdorf. Fecundity, Fertility and The Formation of Human Capital. *The Economic Journal*, 129(618):925–960, 02 2018. ISSN 0013-0133. doi: 10.1111/ecoj.12589. URL https://doi.org/10.1111/ecoj.12589.

Gregory Clark and Neil Cummins. Intergenerational wealth mobility in england, 1858–2012: Surnames and social mobility. *The Economic Journal*, 125(582):61–85, 2015. doi: https://doi.org/10.1111/ecoj.12165. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ecoj.12165.

Alan Fernihough. Human capital and the quantity–quality trade-off during the demographic transition. *Journal of Economic Growth*, 22(1):35–65, March 2017. doi: 10.1007/s10887-016-9138-3. URL https://ideas.repec.org/a/kap/jecgro/v22y2017i1d10.1007_s10887-016-9138-3.html.

Joshua Angrist, Victor Lavy, and Analia Schlosser. Multiple experiments for the causal link between the quantity and quality of children. *Journal of Labor Economics*, 28(4):773–824, 2010. ISSN 0734306X, 15375307. URL http://www.jstor.org/stable/10.1086/653830.

Mara P. Squicciarini and Nico Voigtländer. Human Capital and Industrialization: Evidence from the Age of Enlightenment *. *The Quarterly Journal of Economics*, 130(4):1825–1883, 07 2015. ISSN 0033-5533. doi: 10.1093/qje/qjv025. URL https://doi.org/10.1093/qje/qjv025.

Roy E. Bailey, Timothy J. Hatton, and Kris Inwood. Health, height, and the household at the turn of the twentieth century. *The Economic History Review*, 69(1):35–53, 2016.

doi: https://doi.org/10.1111/ehr.12099. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ehr.12099.

Ran Abramitzky, Leah Platt Boustan, and Katherine Eriksson. Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration. *The American Economic Review*, 102(5):1832–1856, 2012. ISSN 00028282. URL http://www.jstor.org/stable/41724607.

Rangachar Kasturi, Lawrence O'Gorman, and Venu Govindaraju. Document image analysis: A primer. *Sadhana*, 27:3–22, 2002.

Xu Zhong, Jianbin Tang, and Antonio Jimeno-Yepes. Publaynet: largest dataset ever for document layout analysis. *CoRR*, abs/1908.07836, 2019. URL http://arxiv.org/abs/1908.07836.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

Gregory Clark and Neil Cummins. The Child Quality-Quantity Tradeoff, England, 1780-1880: A Fundamental Component of the Economic Theory of Growth is Missing. CEPR Discussion Papers 11232, C.E.P.R. Discussion Papers, April 2016. URL https://ideas.repec.org/p/cpr/ceprdp/11232.html.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015.

Ming-Wei Lin, Jules-Raymond Tapamo, and Baird Ndovie. A texture-based method for document segmentation and classification. *South African Computer Journal*, 36:49–56, 01 2006. doi: 10.46298/arima.1878.