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

Christian E. Westermann & Christian M. Dahl

University of Southern Denmark

December 31, 2021

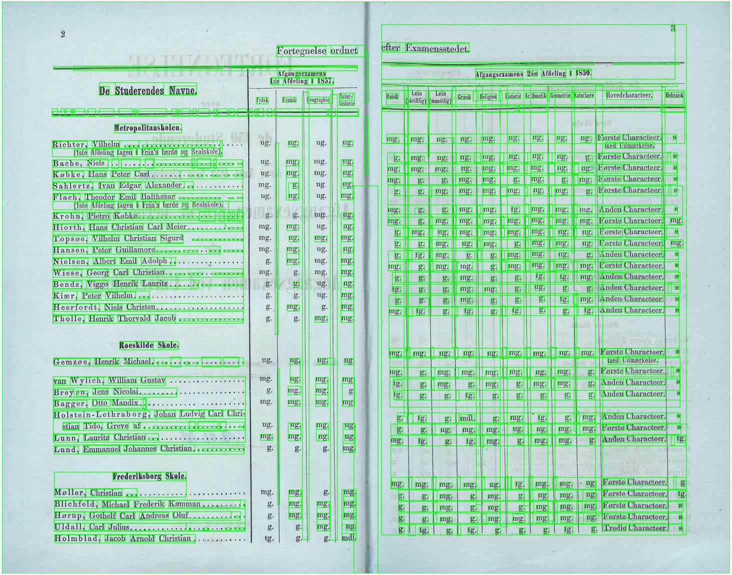
Motivation

- A unique database on human capital in the Nordic countries, spanning multiple centuries.
- With successful links to current registries, large potential for multi-generational analyses.
- To the best of our knowledge, no end-to-end demonstration of historical tabular data segmentation and transcription exists.
- Structured tabular data are extremely valuable, especially for economists (census data, etc.).
- Many subtle but extremely important complications.
- Contribute a generalizable approach with respect to structured document tables.

Related Work

- Many parallels to dhSegment, the work of Oliveira et al. (2019).
  - They focus on deep learning based generic segmentation of several tasks such as page-, image-, and text-detection.
  - We wish to introduce a generic approach to structured tabular data.
- LayoutParser by Shen et al. (2021).
  - Unified toolkit for Deep Learning based DIA.
  - Includes layout detection, OCR with Google Vision or Tesseract backend and more.
- Google's Tesseract
  - While sufficient in many tasks, it is too inaccurate for the tables we work with.
  - Too many or not enough boxes detected.
  - Very hard to sort relevant boxes.

Tesseract Example

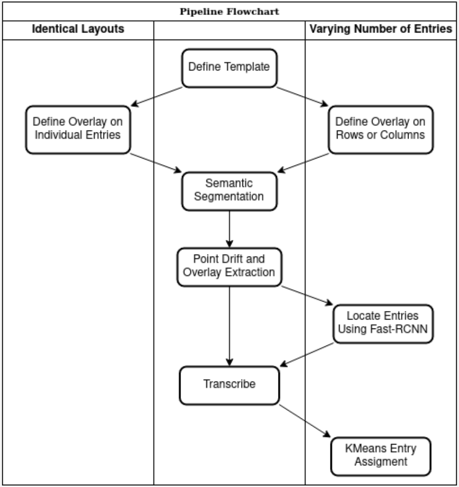


2 3

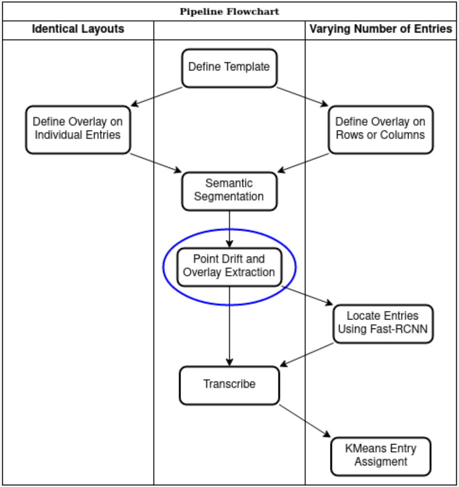
Fortegnelse ordnet efter Examensstedet.

De Studerendes Navne.	Afgangsexamens 1ste Afdeling i 1857. Tydsk	Fransk	Geographie	Natur-historie	Afgangsexamens 2den Afdeling i 1859. Dansk	Latin (skriftlig)	Latin (mundtlig)	Græsk	Religion	Historie	Arithmetik	Geometrie	Naturlære	Hovedcharacteer.	Hebraisk
Metropolitanskolen.															
Richter, Vilhelm (1ste Afdeling tagen i Friis's lærde og Realskole).	ug.	ug.	ug.	ug.	mg.	mg.	ug.	ug.	mg.	ug.	ug.	ug.	ug.	Første Characteer. med Udmærkelse.	»
Bache, Niels	ug.	mg.	mg.	ug.	g.	mg.	ug.	ug.	mg.	ug.	ug.	ug.	g.	Første Characteer.	»
Købke, Hans Peter Carl	ug.	mg.	mg.	ug.	mg.	mg.	mg.	ug.	ug.	mg.	mg.	ug.	ug.	Første Characteer.	»
Sahlertz, Ivan Edgar Alexander	mg.	g.	ug.	ug.	mg.	g.	g.	mg.	mg.	g.	mg.	g.	mg.	Første Characteer.	»
Flach, Theodor Emil Balthasar (1ste Afdeling tagen i Friis's lærde og Realskole).	ug.	ug.	ug.	mg.	g.	g.	mg.	mg.	mg.	mg.	ug.	mg.	g.	Første Characteer.	»
Krohn, Pietro Købke	g.	g.	mg.	ug.	mg.	g.	g.	mg.	mg.	tg.	mg.	mg.	mg.	Anden Characteer.	»
Hiorth, Hans Christian Carl Meier	mg.	mg.	ug.	ug.	mg.	g.	mg.	mg.	mg.	mg.	mg.	mg.	mg.	Første Characteer.	mg.
Topsøe, Vilhelm Christian Sigurd	mg.	ug.	mg.	mg.	g.	mg.	ug.	mg.	mg.	mg.	mg.	ug.	ug.	Første Characteer.	»
Hansen, Peter Guillamore	mg.	mg.	ug.	ug.	g.	g.	mg.	ug.	mg.	g.	mg.	mg.	ug.	Første Characteer.	mg.
Nielsen, Albert Emil Adolph	g.	mg.	mg.	mg.	g.	tg.	mg.	g.	g.	mg.	mg.	ug.	g.	Anden Characteer.	»
Wiese, Georg Carl Christian	mg.	g.	mg.	mg.	mg.	g.	mg.	mg.	g.	mg.	mg.	mg.	mg.	Første Characteer.	»
Bendz, Viggo Henrik Lauritz	g.	g.	ug.	ug.	g.	g.	g.	mg.	g.	g.	tg.	tg.	mg.	Anden Characteer.	»
Kiær, Peter Vilhelm	g.	g.	ug.	mg.	tg.	g.	g.	mg.	mg.	g.	ug.	g.	g.	Anden Characteer.	»
Heerfordt, Niels Christen	g.	mg.	g.	mg.	g.	g.	g.	mg.	g.	g.	g.	tg.	mg.	Anden Characteer.	»
Tholle, Henrik Thorvald Jacob	g.	g.	mg.	mg.	mg.	tg.	g.	tg.	g.	tg.	g.	g.	tg.	Anden Characteer.	»
Roeskilde Skole.															
Gemzøe, Henrik Michael	ug.	ug.	ug.	ug.	mg.	mg.	ug.	ug.	ug.	mg.	mg.	ug.	mg.	Første Characteer. med Udmærkelse.	»
van Wylich, William Gustav	mg.	ug.	mg.	mg.	mg.	g.	mg.	mg.	mg.	mg.	ug.	mg.	g.	Første Characteer.	»
Breyen, Jens Nicolai	g.	mg.	mg.	g.	tg.	g.	mg.	g.	mg.	g.	mg.	mg.	g.	Anden Characteer.	»
Bagger, Otto Mandix	mg.	mg.	mg.	mg.	tg.	g.	g.	tg.	g.	g.	g.	g.	g.	Anden Characteer.	»
Holstein-Lethraborg, Johan Ludvig Carl Christian Tido, Greve af	ug.	ug.	mg.	ug.	g.	tg.	g.	mdl.	g.	mg.	tg.	g.	mg.	Anden Characteer.	»
Lunn, Lauritz Christian	mg.	mg.	ug.	ug.	g.	g.	ug.	mg.	mg.	ug.	ug.	mg.	mg.	Første Characteer.	»
Lund, Emmanuel Johannes Christian	g.	g.	g.	mg.	mg.	tg.	g.	tg.	mg.	g.	mg.	mg.	g.	Anden Characteer.	tg.
Frederiksborg Skole.															
Møller, Christian	mg.	mg.	g.	mg.	mg.	mg.	mg.	mg.	ug.	tg.	mg.	mg.	ug.	Første Characteer.	g.
Blichfeld, Michael Frederik Kamman	g.	mg.	mg.	mg.	g.	g.	mg.	g.	mg.	g.	ug.	mg.	ug.	Første Characteer.	tg.
Hørup, Gothelf Carl Andreas Oluf	g.	mg.	mg.	mg.	g.	g.	mg.	g.	mg.	g.	mg	mg.	mg	Første Characteer.	»
Uldall, Carl Julius	g.	g.	mg.	ug.	g.	g.	mg.	g.	mg.	mg.	mg	mg.	ug.	Første Characteer.	»
Holmblad, Jacob Arnold Christian	tg.	g.	g.	mdl.	g.	tg.	g.	tg.	g.	g.	g.	tg.	g.	Tredie Characteer.	»

Approach Overview



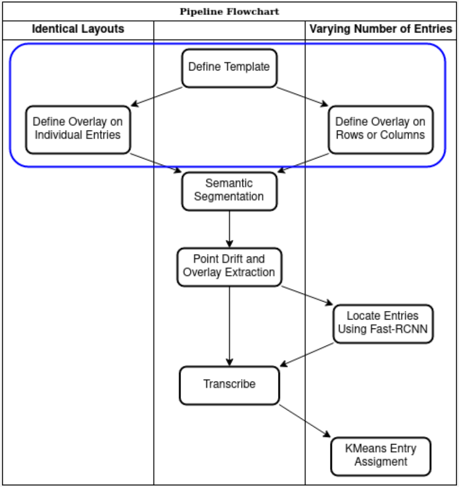
Approach Overview



Point Drift

- The engine behind the approach is FilterReg by Wei Gao and Russ Tedrake (2018), a Point-Set Registration algorithm.
- Learn the Motion (Transformation) Parameters, $\Delta\theta$, responsible for the alignment of two point clouds, X and Y.
- We can then apply $\Delta\theta^{-1}$ to Y, which aligns it with X.

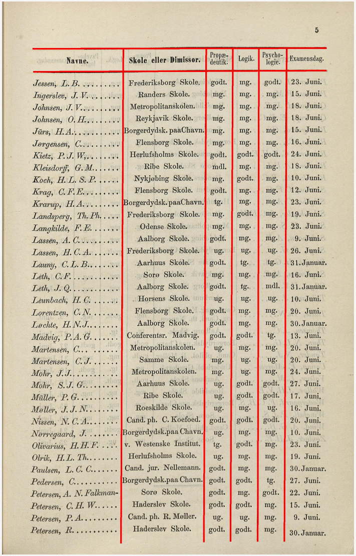
Approach Overview



Template and Overlay

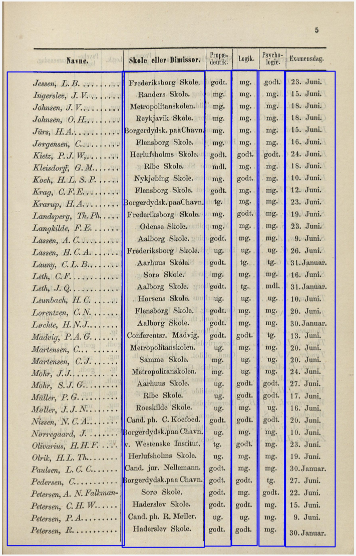
- Manually define two setup steps per document layout type: template and overlay.
- We define the template as the document table outline.
- How to define the overlay is dependent on whether number of entries vary.
  - Static: Entry-level overlay
  - Varying: Column- or Row-level overlay.

Template and Overlay



5

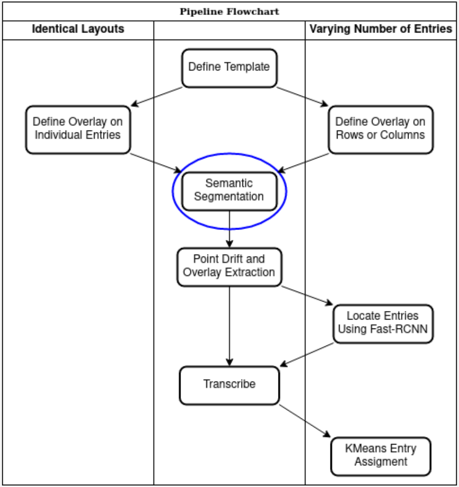
Navne.	Skole eller Dimissor.	Propædeutik.	Logik.	Psychologie.	Examensdag.
Jessen, L. B.	Frederiksborg Skole.	godt.	mg.	godt.	23. Juni.
Ingerslev, J. V.	Randers Skole.	mg.	mg.	mg.	15. Juni.
Johnsen, J. V.	Metropolitanskolen.	mg.	mg.	mg.	18. Juni.
Johnsen, O. H.	Reykjavik Skole.	mg.	mg.	mg.	18. Juni.
Jürs, H. A.	Borgerdydsk. paaChavn.	mg.	mg.	mg.	15. Juni.
Jørgensen, C.	Flensborg Skole.	mg.	mg.	mg.	16. Juni.
Kietz, P. J. W.	Herlufsholms Skole.	godt.	godt.	godt.	24. Juni.
Kleisdorff, G. M.	Ribe Skole.	mdl.	mg.	mg.	18. Juni.
Koch, H. L. S. P.	Nykjøbing Skole.	mg.	godt.	mg.	10. Juni.
Krag, C. F. E.	Flensborg Skole.	godt.	mg.	mg.	12. Juni.
Krarup, H. A.	Borgerdydsk. paaChavn.	tg.	mg.	mg.	23. Juni.
Landsperg, Th. Ph.	Frederiksborg Skole.	mg.	godt.	mg.	19. Juni.
Langkilde, F. E.	Odense Skole.	mg.	mg.	mg.	23. Juni.
Lassen, A. C.	Aalborg Skole.	godt.	mg.	mg.	9. Juni.
Lassen, H. C. A.	Frederiksborg Skole.	ug.	ug.	ug.	26. Juni.
Launy, C. L. B.	Aarhuus Skole.	godt.	tg.	tg.	31. Januar.
Leth, C. F.	Sorø Skole.	mg.	mg.	mg.	16. Juni.
Leth, J. Q.	Aalborg Skole.	godt.	tg.	mdl.	31. Januar.
Leunbach, H. C.	Horsens Skole.	ug.	ug.	ug.	10. Juni.
Lorentzen, C. N.	Flensborg Skole.	godt.	mg.	mg.	20. Juni.
Løchte, H. N. J.	Aalborg Skole.	godt.	mg.	mg.	30. Januar.
Madvig, P. A. G.	Conferentsr. Madvig.	godt.	godt.	tg.	13. Juni.
Martensen, C.	Metropolitanskolen.	ug.	mg.	mg.	20. Juni.
Martensen, C. J.	Samme Skole.	mg.	ug.	ug.	20. Juni.
Mohr, J. J.	Metropolitanskolen.	mg.	ug.	mg.	24. Juni.
Mohr, S. J. G.	Aarhuus Skole.	ug.	godt.	godt.	27. Juni.
Müller, P. G.	Ribe Skole.	ug.	godt.	godt.	17. Juni.
Møller, J. J. N.	Roeskilde Skole.	ug.	mg.	ug.	16. Juni.
Nissen, N. C. A.	Cand. ph. C. Koefoed.	godt.	godt.	godt.	20. Juni.
Nørregaard, J.	Borgerdydsk.paa Chavn.	ug.	mg.	mg.	10. Juni.
Olivarius, H. H. F.	v. Westenske Institut.	tg.	godt.	mg.	23. Juni.
Olrik, H. L. Th.	Herlufsholms Skole.	ug.	mg.	mg.	19. Juni.
Paulsen, L. C. C.	Cand. jur. Nellemann.	godt.	mg.	mg.	30. Januar.
Pedersen, C.	Borgerdydsk.paa Chavn.	godt.	godt.	tg.	27. Juni.
Petersen, A. N. Falkman-	Sorø Skole.	godt.	mg.	godt.	22. Juni.
Petersen, C. H. W.	Haderslev Skole.	godt.	godt.	mg.	15. Juni.
Petersen, P. A.	Cand. ph. R. Møller.	ug.	ug.	mg.	9. Juni.
Petersen, R.	Haderslev Skole.	godt.	godt.	mg.	30. Januar.



5

Navne.	Skole eller Dimissor.	Propædeutik.	Logik.	Psychologie.	Examensdag.
Jessen, L. B.	Frederiksborg Skole.	godt.	mg.	godt.	23. Juni.
Ingerslev, J. V.	Randers Skole.	mg.	mg.	mg.	15. Juni.
Johnsen, J. V.	Metropolitanskolen.	mg.	mg.	mg.	18. Juni.
Johnsen, O. H.	Reykjavik Skole.	mg.	mg.	mg.	18. Juni.
Jürs, H. A.	Borgerdydsk. paaChavn.	mg.	mg.	mg.	15. Juni.
Jørgensen, C.	Flensborg Skole.	mg.	mg.	mg.	16. Juni.
Kietz, P. J. W.	Herlufsholms Skole.	godt.	godt.	godt.	24. Juni.
Kleisdorff, G. M.	Ribe Skole.	mdl.	mg.	mg.	18. Juni.
Koch, H. L. S. P.	Nykjøbing Skole.	mg.	godt.	mg.	10. Juni.
Krag, C. F. E.	Flensborg Skole.	godt.	mg.	mg.	12. Juni.
Krarup, H. A.	Borgerdydsk. paaChavn.	tg.	mg.	mg.	23. Juni.
Landsperg, Th. Ph.	Frederiksborg Skole.	mg.	godt.	mg.	19. Juni.
Langkilde, F. E.	Odense Skole.	mg.	mg.	mg.	23. Juni.
Lassen, A. C.	Aalborg Skole.	godt.	mg.	mg.	9. Juni.
Lassen, H. C. A.	Frederiksborg Skole.	ug.	ug.	ug.	26. Juni.
Launy, C. L. B.	Aarhuus Skole.	godt.	tg.	tg.	31. Januar.
Leth, C. F.	Sorø Skole.	mg.	mg.	mg.	16. Juni.
Leth, J. Q.	Aalborg Skole.	godt.	tg.	mdl.	31. Januar.
Leunbach, H. C.	Horsens Skole.	ug.	ug.	ug.	10. Juni.
Lorentzen, C. N.	Flensborg Skole.	godt.	mg.	mg.	20. Juni.
Løchte, H. N. J.	Aalborg Skole.	godt.	mg.	mg.	30. Januar.
Madvig, P. A. G.	Conferentsr. Madvig.	godt.	godt.	tg.	13. Juni.
Martensen, C.	Metropolitanskolen.	ug.	mg.	mg.	20. Juni.
Martensen, C. J.	Samme Skole.	mg.	ug.	ug.	20. Juni.
Mohr, J. J.	Metropolitanskolen.	mg.	ug.	mg.	24. Juni.
Mohr, S. J. G.	Aarhuus Skole.	ug.	godt.	godt.	27. Juni.
Müller, P. G.	Ribe Skole.	ug.	godt.	godt.	17. Juni.
Møller, J. J. N.	Roeskilde Skole.	ug.	mg.	ug.	16. Juni.
Nissen, N. C. A.	Cand. ph. C. Koefoed.	godt.	godt.	godt.	20. Juni.
Nørregaard, J.	Borgerdydsk.paa Chavn.	ug.	mg.	mg.	10. Juni.
Olivarius, H. H. F.	v. Westenske Institut.	tg.	godt.	mg.	23. Juni.
Olrik, H. L. Th.	Herlufsholms Skole.	ug.	mg.	mg.	19. Juni.
Paulsen, L. C. C.	Cand. jur. Nellemann.	godt.	mg.	mg.	30. Januar.
Pedersen, C.	Borgerdydsk.paa Chavn.	godt.	godt.	tg.	27. Juni.
Petersen, A. N. Falkman-	Sorø Skole.	godt.	mg.	godt.	22. Juni.
Petersen, C. H. W.	Haderslev Skole.	godt.	godt.	mg.	15. Juni.
Petersen, P. A.	Cand. ph. R. Møller.	ug.	ug.	mg.	9. Juni.
Petersen, R.	Haderslev Skole.	godt.	godt.	mg.	30. Januar.

Approach Overview



Semantic Segmentation

- Now that we have X, how do we find Y?
- Semantic segmentation (pixel-wise classification) using Deep Learning.
- Three classes: Line, text / scribbles and background.
- Pixels that make up detected lines become the Y point cloud.

Semantic Segmentation: Neural Network Architecture

- We use the architecture of DeeplabV3+, by Chen et al. (2018).
- U-shaped and like other U-shaped architectures since originally introduced by Ronneberger et al. in 2015, it has proven to excel at semantic segmentation.
- We train two separate models, one for horizontal line identification and one for vertical.

Semantic Segmentation: Training Data

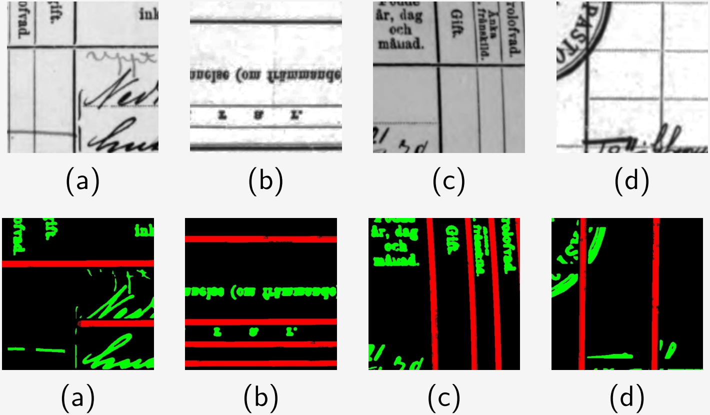
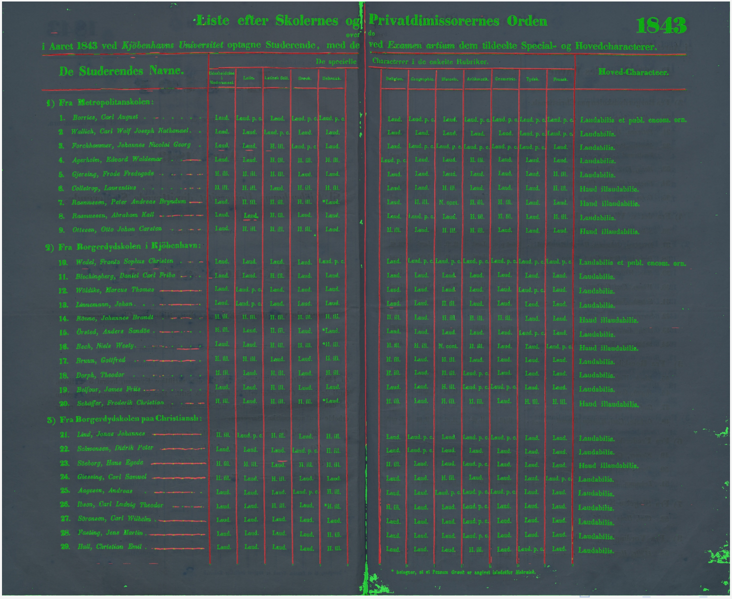


Figure: Top row: Input training images. Bottom row: Corresponding labels / masks. Red, green and black correspond to lines, text and background respectively.

Semantic Segmentation: Results



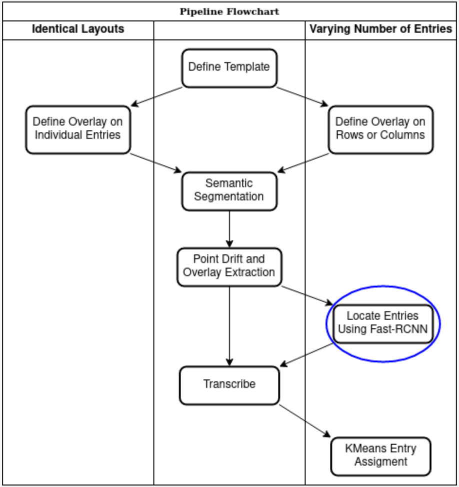
Liste efter Skolernes og Privatdimissorernes Orden 1843

over de i Aaret 1843 ved *Kjöbenhavns Universitet* optagne Studerende, med de ved *Examen artium* dem tildeelte Special- og Hovedcharacterer.

	De Studerendes Navne.	De specielle Characterer i de enkelte Rubriker.												Hoved-Characteer.
		Udarbeidelse i Modersmaal.	Latin.	Latinsk Stil.	Græsk.	Hebraisk.	Religion.	Geographie.	Historie.	Arithmetik.	Geometrie.	Tydsk.	Fransk.	
1)	Fra Metropolitanskolen:													
1.	*Borries, Carl August*	Laud.	Laud. p. c.	Laud.	Laud. p. c.	Laud. p. c.	Laud.	Laud. p. c.	Laud.	Laud. p. c.	Laud. p. c.	Laud. p. c.	Laud. p. c.	Laudabilis et publ. encom. orn.
2.	*Wallich, Carl Wolf Joseph Nathanael*	Laud.	Laud.	Laud. p. c.	Laud.	Laud.	Laud.	Laud.	Laud.	Laud.	Laud.	Laud. p. c.	Laud. p. c.	Laudabilis.
3.	*Forchhammer, Johannes Nicolai Georg*	Laud.	Laud.	H. ill.	Laud. p. c.	Laud.	Laud.	Laud. p. c.	Laud. p. c.	Laud. p. c.	Laud. p. c.	Laud. p. c.	Laud.	Laudabilis.
4.	*Agerholm, Edvard Waldemar*	Laud.	Laud.	H. ill.	H. ill.	H. ill.	Laud. p. c.	Laud.	Laud.	H. ill.	Laud.	Laud.	Laud.	Laudabilis.
5.	*Gjersing, Frode Fredegode*	H. ill.	H. ill.	H. ill.	Laud.	Laud.	Laud.	Laud.	Laud.	Laud.	Laud.	Laud.	H. ill.	Laudabilis.
6.	*Collstrop, Laurentius*	H. ill.	H. ill.	Laud.	H. ill.	H. ill.	Laud.	Laud.	H. ill.	Laud.	Laud.	Laud.	H. ill.	Haud illaudabilis.
7.	*Rasmussen, Peter Andreas Bryndum*	Laud.	H. ill.	H. ill.	H. ill.	*Laud.	Laud.	H. ill.	N. cont.	H. ill.	H. ill.	Laud.	Laud.	Haud illaudabilis.
8.	*Rasmussen, Abraham Kall*	Laud.	Laud.	H. ill.	Laud.	Laud.	Laud.	Laud. p. c.	Laud.	H. ill.	H. ill.	Laud.	H. ill.	Laudabilis.
9.	*Ottesen, Otto Johan Carsten*	Laud.	H. ill.	H. ill.	H. ill.	Laud.	H. ill.	Laud.	H. ill.	H. ill.	Laud.	Laud.	Laud.	Haud illaudabilis.
2)	Fra Borgerdydskolen i Kjöbenhavn:													
10.	*Wedel, Frantz Sophus Christen*	Laud.	Laud.	Laud.	Laud.	Laud. p. c.	Laud.	Laud. p. c.	Laud. p. c.	Laud. p. c.	Laud. p. c.	Laud. p. c.	Laud. p. c.	Laudabilis et publ. encom. orn.
11.	*Blechingberg, Daniel Carl Fribo*	Laud.	Laud.	H. ill.	Laud.	Laud.	Laud.	Laud.	Laud.	Laud.	Laud.	Laud.	Laud.	Laudabilis.
12.	*Wöldike, Marcus Thomas*	Laud.	Laud. p. c.	Laud.	Laud.	Laud.	Laud.	Laud.	Laud.	Laud.	Laud.	Laud. p. c.	Laud.	Laudabilis.
13.	*Linnemann, Johan*	Laud.	Laud. p. c.	Laud.	Laud.	Laud.	Laud.	Laud.	H. ill.	Laud.	Laud.	Laud.	Laud.	Laudabilis.
14.	*Rönne, Johannes Brandt*	H. ill.	H. ill.	H. ill.	H. ill.	H. ill.	H. ill.	Laud.	H. ill.	H. ill.	H. ill.	Laud.	Laud.	Haud illaudabilis.
15.	*Ørsted, Anders Sandöe*	H. ill.	Laud.	H. ill.	Laud.	*Laud.	Laud.	Laud.	Laud.	Laud.	Laud.	Laud. p. c.	Laud.	Laudabilis.
16.	*Bech, Niels Westy*	Laud.	Laud.	H. ill.	H. ill.	*H. ill.	H. ill.	H. ill.	N. cont.	H. ill.	Laud.	Laud.	Laud. p. c.	Haud illaudabilis.
17.	*Bruun, Gottfred*	H. ill.	H. ill.	Laud.	Laud.	H. ill.	H. ill.	Laud.	H. ill.	Laud.	Laud.	Laud.	Laud.	Laudabilis.
18.	*Dorph, Theodor*	H. ill.	Laud.	H. ill.	Laud.	H. ill.	H. ill.	Laud.	H. ill.	Laud. p. c.	Laud.	Laud.	Laud.	Laudabilis.
19.	*Belfour, James Friis*	Laud.	Laud.	H. ill.	Laud.	Laud.	Laud.	Laud.	H. ill.	Laud. p. c.	Laud. p. c.	Laud.	Laud.	Laudabilis.
20.	*Schäffer, Frederik Christian*	H. ill.	Laud.	H. ill.	H. ill.	*Laud.	H. ill.	Laud.	H. ill.	H. ill.	Laud.	H. ill.	H. ill.	Haud illaudabilis.
3)	Fra Borgerdydskolen paa Christiansh:													
21.	*Lind, Jonas Johannes*	H. ill.	Laud. p. c.	H. ill.	Laud.	H. ill.	Laud.	Laud. p. c.	Laud.	Laud. p. c.	Laud. p. c.	Laud.	Laud.	Laudabilis.
22.	*Schwensen, Didrik Peter*	Laud.	Laud.	Laud.	Laud. p. c.	H. ill.	Laud.	Laud. p. c.	Laud.	Laud.	Laud.	Laud.	Laud.	Laudabilis.
23.	*Söeborg, Hans Egede*	H. ill.	H. ill.	Laud.	H. ill.	H. ill.	H. ill.	Laud.	Laud.	Laud. p. c.	Laud.	Laud.	Laud.	Haud illaudabilis.
24.	*Giessing, Carl Samuel*	H. ill.	Laud.	H. ill.	Laud.	Laud.	Laud.	Laud.	H. ill.	Laud.	Laud.	Laud.	Laud. p. c.	Laudabilis.
25.	*Aagesen, Andreas*	Laud.	Laud.	Laud.	Laud. p. c.	H. ill.	Laud.	Laud.	Laud. p. c.	Laud. p. c.	Laud. p. c.	Laud.	Laud.	Laudabilis.
26.	*Ibsen, Carl Ludvig Theodor*	Laud.	Laud.	H. ill.	Laud.	*H. ill.	H. ill.	Laud.	Laud.	Laud. p. c.	Laud.	Laud.	Laud.	Laudabilis.
27.	*Sörensen, Carl Wilhelm*	Laud.	Laud.	Laud.	Laud.	Laud.	Laud.	Laud.	Laud.	Laud. p. c.	Laud.	Laud.	Laud.	Laudabilis.
28.	*Fasting, Jens Martin*	Laud.	Laud.	Laud.	Laud.	H. ill.	Laud.	Laud.	Laud.	Laud. p. c.	Laud.	Laud.	Laud.	Laudabilis.
29.	*Hall, Christian Emil*	Laud.	Laud.	Laud.	Laud.	H. ill.	Laud.	Laud.	Laud.	H. ill.	Laud.	Laud. p. c.	Laud.	Laudabilis.

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

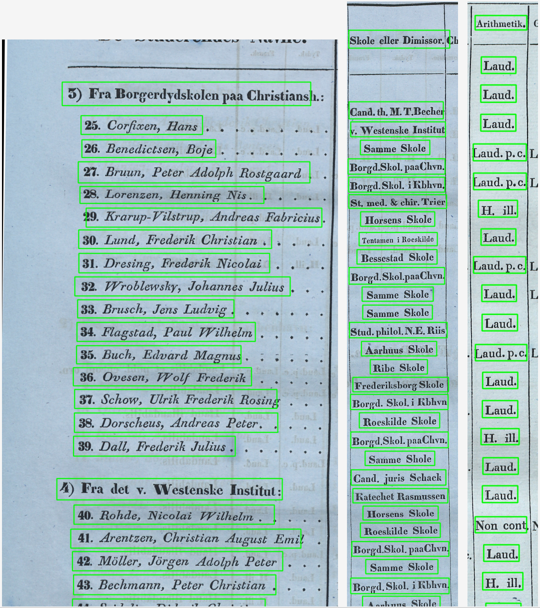
Approach Overview



Entry Detection Using Fast-RCNN

- Faster Region-based Convolutional Network (Faster-RCNN) by Ren et al. (2016).
- Performed column-wise on extracted areas using the overlay.
- We know what column we are dealing with and we can then proceed on locating entries top to bottom.
- Network was trained on data in this paper, but we seek generality here as well, in the future.

Entry Detection Using Fast-RCNN: Results



5) Fra Borgerdydskolen paa Christiansh.:

25. Corfixen, Hans
26. Benedictsen, Boje
27. Bruun, Peter Adolph Rostgaard
28. Lorenzen, Henning Nis
29. Krarup-Vilstrup, Andreas Fabricius
30. Lund, Frederik Christian
31. Dresing, Frederik Nicolai
32. Wroblewsky, Johannes Julius
33. Brusch, Jens Ludvig
34. Flagstad, Paul Wilhelm
35. Buch, Edvard Magnus
36. Ovesen, Wolf Frederik
37. Schow, Ulrik Frederik Rosing
38. Dorscheus, Andreas Peter
39. Dall, Frederik Julius

4) Fra det v. Westenske Institut:

40. Rohde, Nicolai Wilhelm
41. Arentzen, Christian August Emil
42. Möller, Jörgen Adolph Peter
43. Bechmann, Peter Christian

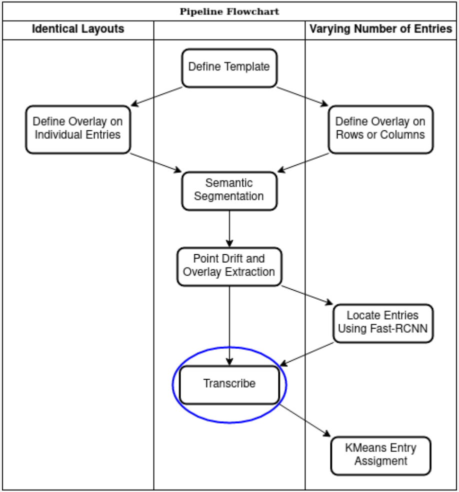
Skole eller Dimissor.

Cand. th. M. T. Becher
v. Westenske Institut
Samme Skole
Borgd. Skol. paa Chvn.
Borgd. Skol. i Kbhvn.
St. med. & chir. Trier
Horsens Skole
Tentamen i Roeskilde
Bessestad Skole
Borgd. Skol. paa Chvn.
Samme Skole
Samme Skole
Stud. philol. N. E. Riis
Aarhuus Skole
Ribe Skole
Frederiksborg Skole
Borgd. Skol. i Kbhvn
Roeskilde Skole
Borgd. Skol. paa Chvn.
Samme Skole
Cand. juris Schack
Katechet Rasmussen
Horsens Skole
Roeskilde Skole
Borgd. Skol. paa Chvn.
Samme Skole
Borgd. Skol. i Kbhvn.
Aarhuus Skole

Arithmetik.

Laud.
Laud.
Laud.
Laud. p. c.
Laud. p. c.
H. ill.
Laud.
Laud. p. c.
Laud.
Laud.
Laud. p. c.
Laud.
Laud.
H. ill.
Laud.
Laud.
Non cont.
Laud.
H. ill.

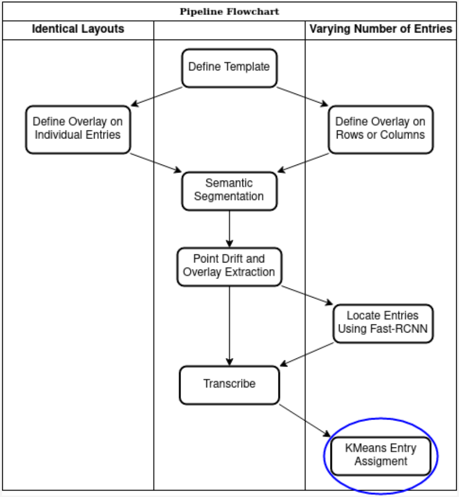
Approach Overview



Transcription

- Transcription can be done using any feasible model.
- Given the precise cutouts together with a little cleaning, Tesseract performs very well on these documents.
- Otherwise, you would have to do some fine-tuning / custom training.

Approach Overview



Assigning Entries Using KMeans

- Crucial step of the pipeline for documents with no constant number of entries.
- Missing an entry propagates the error which in the worst case leads to a full document of wrong assignments.
- We alleviate this by assigning entries using KMeans-clustering.
  - Issue a majority vote on the number of entries.
  - Create a cluster for each row based on the center y-coordinate of each bounding box.
  - Iterate through entries in each column and predict their respective row.