





Entwicklung eines Modellrepositoriums und einer Automatischen Schriftarterkennung für OCR-D

Vincent Christlein¹, Gregory Crane², Rui Dong³, Saskia Limbach⁴, Andreas Maier¹, Mathias Seuret¹, Nikolaus Weichselbaumer⁴

¹Pattern Recognition Lab, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen ²Lehrstuhl für Digital Humanities, Institut für Informatik, Universität Leipzig, Leipzig ³Khoury College of Computer Sciences, Northeastern University, Boston ⁴Gutenberg-Institut für Weltliteratur und schriftorientierte Medien Abteilung Buchwissenschaft, Mainz

Zielsetzung

Schriftarterkennung:

• Erkennung von Gruppen (nicht einzelnen Typen)

FACULTY OF ENGINEERING

- Vollautomatisch
- Ohne Information über den Textinhalt
- Vorstufe für schriftartspezifische OCR

Trainingsinfrastruktur:

- Training verschiedener OCR-Engines über eine Schnittstelle
- Repositorium f
 ür alle trainierten Modelle

Berücksichtigte Schriftarten

ut studium coniurationis vehementer simulent: ceteros adeant hene nolliceantur Denta oneram uti ena mina

τεροιδε γενόμοι, ταις Δυχαις ανόητι δοπόνως μου

סרייטמג החבן רי ווייבר דען אנהג מנאן זיח (חוֹה) זיח וחוֹן אעוֹ נעאן זח גרוש חוו פֿין חוֹש ררייח פירניג חייחר חול עטווחש אין חול ווען רער טייג קוענר ווער יחדערפטן זיח ניט די חוה דר פון נעאן

Hebräisch

Griechisch

olocaultū i laccificia ut dūs impaue

lapis in centro. Et per hoc arguit qu'finis viti quibi vel ponit zë fuit additio z loquit aliquatulu da

babendam credidit jus nostrum; Sed & plebejorum omnisg, sieure autres natios Et psien q par les fais diceulp

M So nym rettich der wol gedönet sey vnnd gutten frischen zitwan, vnd stoß das vnter einander vnnd mach darauß ein kleynes stup mit Schwabacher

Büchern, die man von der Zimmermannskunft hat, einige Anweisung finden mochte, wi den Reimern mehr als zu wohl bekannt fenn, und aus diefer Urfache glaube ich auch in ge-Fraktur

Methodik

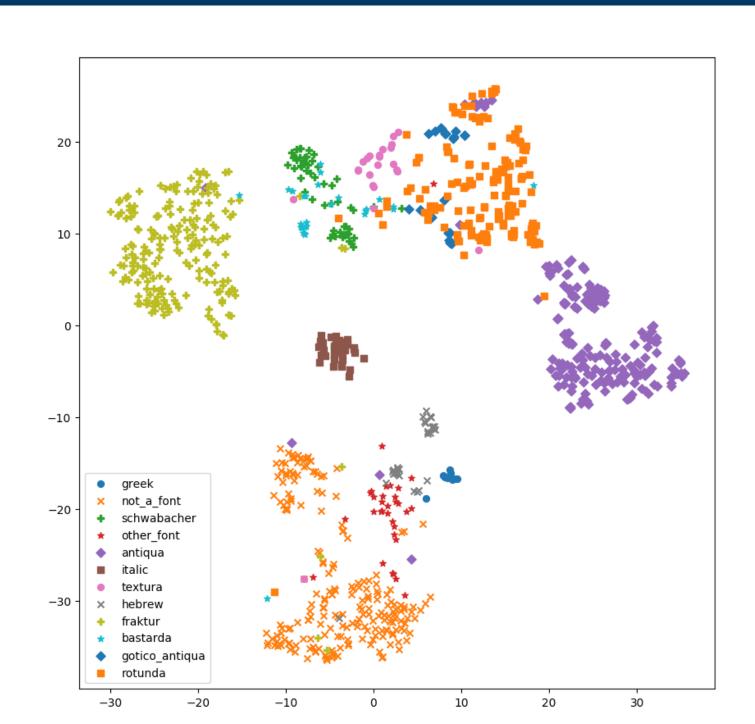
- Definition von Schriftartklassen auf Basis des buchhistorischen Forschungsstands (16.–18. Jh.)
- Überprüfung dieser Klassifikation durch exploratives Clustering
- Datengrundlage: 35 623 gelabelte Seiten, je 0–5 Schriftarten pro Seite
- Klassifikation durch ein tiefes neuronales Netzwerk (DenseNet-121)
- Webbasierte OCR-Trainingsinfrastruktur

Ergebnisse

Classification→				Gotico-									
↓Ground Truth	Antiqua	Bastarda	Fraktur	Antiqua	Griechisch	Hebräisch	Italic	Rotunda	Schwabacher	Textura	Not a font	Other font	Recall
Antiqua	1531		10				5	2				5	98,6%
Bastarda		286						6	10	1			94,4%
Fraktur			1933					1	5	1		2	99,5%
Gotico-Antiqua				269						1			99,6%
Griechisch					58	1					1		96,7%
Hebräisch					1	326							99,7%
Italic			1				187						99,5%
Rotunda				9				1495	5	11		1	98,3%
Schwabacher		16	4					2	452				95,4%
Textura				2						371		1	99,2%
Other font						1		1	1		288	15	94,1%
Not a font	4		2	2	1		5	1	7		4	2331	98,9%
Accuracy	99,7%	94,7%	99,1%	95,4%	96,7%	99,4%	94,9%	99,1%	94,2%	96,4%	98,3%	99,0%	

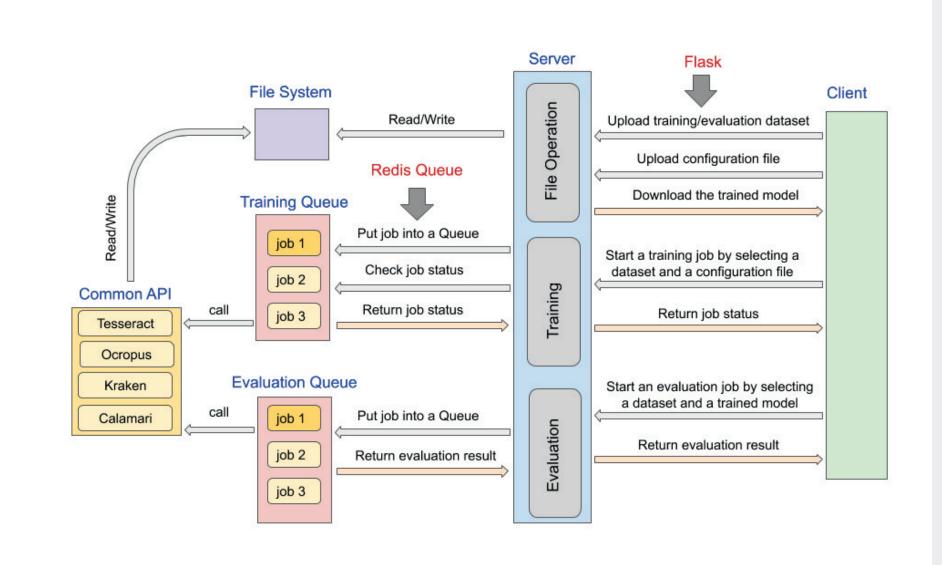
Testseiten: 9675 | korrekt klassifiziert: 9527 | Genauigkeit: 98,5 %

Clustering



- Jeder Punkt entspricht einer Seite
- Cluster der Schriftartgruppen klar erkennbar

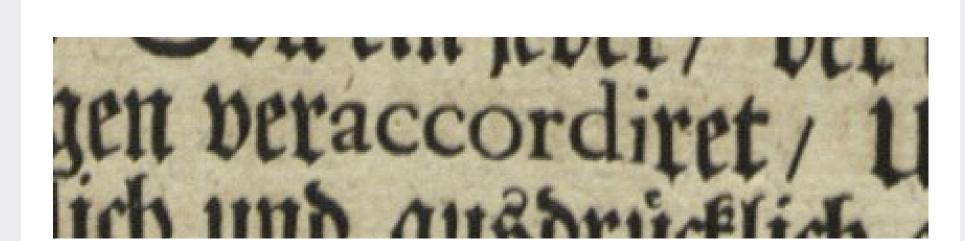
okralact



Webbasierte OCR-Trainingsinfrastruktur:

- Vereinheitlichte Schnittstelle von OCR engines
- Speicherung trainierter Modelle im Repositorium möglich

Herausforderungen



Feingranulare Erkennung:

- Kleinteiligere Schriftartgruppen
- Mehrere Schriftarten auf einer Seite
- Schriftartwechsel in der Zeile / im Wort

Synthetische Trainingsdaten:

- Mögliche Fontgenerierung aus Scans
- Potenzial zur Verbesserung sowohl der Schriftartklassifikation als auch OCR

Projektpublikationen

- Weichselbaumer, Nikolaus / Seuret, Mathias / Limbach, Saskia / Christlein, Vincent / Maier, Andreas (2019): "Automatic Font Group Recognition in Early Printed Books", in: Digital Humanities im deutschsprachigen Raum (DHd) 2019. 6th International Conference 25-29 March 2019, Universitäten zu Mainz und Frankfurt 84-87 (DOI: https://doi.org/10.5281/zenodo.2596095)
- Seuret, Mathias / Limbach, Saskia / Weichselbaumer, Nikolaus / Maier, Andreas / Christlein, Vincent (2019): "Dataset of Pages from Early Printed Books with Multiple Font Groups", in: Historical Document Imaging and Processing (HIP) 2019, 5th International Workshop 21-22 September 2019, Sydney, Australia 1-6. (DOI: https://doi.org/10.1145/3352631.3352640)
- Baierer, Konstantin / Dong, Rui / Neudecker, Clemens (2019): "Okralact a multi-engine Open Source OCR training system", in: Historical Document Imaging and Processing (HIP) 2019, 5th International Workshop 21-22 September 2019, Sydney, Australia 25–30. (DOI: https://doi.org/10.1145/3352631.3352638)
- Weichselbaumer, Nikolaus / Seuret, Mathias / Limbach, Saskia / Hinrichsen, Lena / Maier, Andreas / Christlein, Vincent (2020): "The rapid rise of Fraktur", in Digital Humanities im deutschsprachigen Raum (DHd) 2020. 7th International Conference 02-06 March 2020, Universität Paderborn

Links



classifier

Okralact

https://github.com/OCR-D/ocrd_typegroups_

Classifier https://doi.org/10.1145/3352631.3352640 https://github.com/OCR-D/okralact

Contacts Dr. Vincent Christlein

Pattern Recognition Lab Friedrich-Alexander-Universität Erlangen-Nürnberg Erlangen, Germany

+49 9131 85 20 281 vincent.christlein@fau.de Dr. Nikolaus Weichselbaumer

Buchwissenschaft Johannes Gutenberg-Universität Mainz, Germany

+4961313923180 weichsel@uni-mainz.de

