

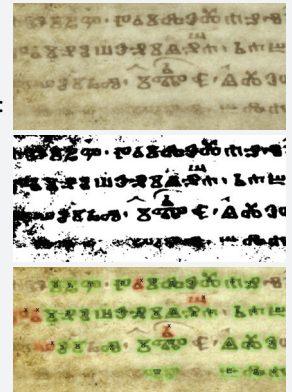
Einleitung

- Effiziente Analyse der digitalisierten Dokumente mithilfe von automatischen Dokumentenanalysetechniken
- Historische Dokumente weisen folgende Schwierigkeiten auf:
 - Initialen
 - Verblichene Zeichen
 - Hintergrundrauschen
- Aufgrund dieser Schwierigkeiten können Standardtechniken nicht angewandt werden
→ Entwicklung spezifischer Dokumentenanalysetechniken



OCR

- Unterstützung der philologischen Entzifferung von historischen Dokumenten
- Erkennung von Handschrift ist wesentlich schwieriger als von Maschinenschrift
- Gängige Software verwendet Binarisierung als Vorverarbeitungsschritt
 - Auf historischen Dokumenten nicht anwendbar
- Neue OCR-Methoden ([1] und [2]):
 - Speziell für ausgebliebene Buchstaben geeignet
 - Benötigen keine vorhergehende Binarisierung
 - Verwenden lokale Features auf Grauwertbildern



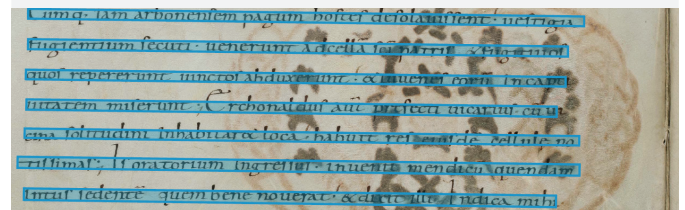
Schreiberidentifikation

- Zuordnung von historischen Dokumenten zu bestimmten Schreibern [3]
- Unterstützung von Paläographen
- Ermittelt wird die Ähnlichkeit des Schriftbildes von Eingabe- und Trainingsbildern
 - Kein Training benötigt
 - Annahme: Die n ähnlichsten Dokumente wurden von derselben Hand verfasst
- Test:
 - 361 Dokumente
 - 7 Schreiber
 - In 98.9% stammte das ähnlichste Bild von derselben Hand



Layout-Analyse

- Erkennung von Textlinien [4]
- Vorverarbeitungsschritt für weitere Algorithmen (z.B. OCR)
- Methodik:
 - Wort-Detektion
 - Textlinien-Detektion
 - Textblock-Analyse



Zusammenfassung

- Dokumentenanalyse historischer Dokumente oft komplexer als von modernen
- Techniken für spezifische Anforderungen wurden geschaffen
- Effiziente Analyse von digitalisierten Dokumenten
- Unterstützung für Philologen, Kodikologen und Paläographen

Literatur

- [1] Markus Diem, Robert Sablatnig: Are Characters Objects? ICPR 2010, pp. 565 - 570
- [2] Sajid Saleem, Fabian Hollaus and Robert Sablatnig: Recognition of degraded ancient characters based on dense SIFT. DATECH 2014, pp. 15-20
- [3] Stefan Fiel, Fabian Hollaus, Melanie Gau, Robert Sablatnig: Writer Identification on Historical Glagolitic Documents. DRR 2014, pp. 902102-1 -- 902102-9, 201
- [4] Markus Diem, Florian Kleber, Robert Sablatnig: Text Line Detection for Heterogeneous Documents, ICDAR 2013, pp. 560-564