Interessensbekundung

im Rahmen des Förderprogramms „e-Research-Technologien“

zur DFG-Ausschreibung:

„Implementierung der OCR-D-Software zur Volltextdigitalisierung historischer Drucke“

Anwendungsszenario:

Vollautomatische Verarbeitung von Scans historischer Drucke durch die OCR-D-Verfahren

eingereicht am 22.05.2020 vom

Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS

Ansprechpartner

Dr.-Ing. Joachim Köhler
Schloss Birlinghoven
53754 Sankt Augustin
Telefon 02241 14-1900
Fax 02241 144-1900
joachim.koehler@iais.fraunhofer.de

Fraunhofer IAIS

DFG

1 Absichtserklärung für die Implementierung der OCR-D-Software zur Volltextdigitalisierung historischer Drucke von Fraunhofer IAIS

Die OCR-D Projekt umfasst einen vollständigen Workflow zur Verarbeitung von historischen Drucken in digitaler Form. Das Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (Fraunhofer IAIS) forscht und entwickelt seit dem Jahr 2003 ebenso Methoden und Software zur Erkennung und Analyse für historische digitalisierte Drucke und hat diese in zahlreichen Projekten produktiv eingesetzt – von der analogen Druckvorlage bis hin zur maschinell lesbaren und volltextdurchsuchbaren Ressource.

2 IAIS sucht einen Partner aus dem Bereich der Kultureinrichtungen

Für die Antragstellung bei der DFG-Ausschreibung *„Implementierung der OCR-D-Software zur Volltextdigitalisierung historischer Drucke“* sucht das Fraunhofer IAIS einen oder mehrere ausgewählte Partner aus dem Bereich von Bibliotheken, Archiven sowie anderen bestandshaltenden bzw. bestandsverarbeitenden Einrichtungen, um einzelne Module aus OCR-D sowie den gesamten OCR-D-Workflow zu testen.

2.1 Suche nach einem oder mehreren Partnern für die Pilotierungsphase von bis zu sechs Monaten

Für Fraunhofer IAIS steht dabei im Mittelpunkt, dass gemeinsam mit einem oder mehreren Partnern während einer bis zu sechsmonatigen Pilotierungsphase praxisnahe Anwendungsszenarien erarbeitet werden, auf deren Grundlage die digitalisierten Drucke der jeweiligen Institutionen durch OCR-D-Verfahren verarbeitet werden können.

3 Vorarbeiten seitens IAIS

3.1 Produktiver Einsatz bereits vorhandener Verfahren

Auf Grundlage der von Fraunhofer IAIS entwickelten Technologien wurden zahlreiche Vorhaben zur Massenverarbeitung von historischen Zeitungen erfolgreich durchgeführt. Die mit Hilfe von IAIS-Technologien erzielten Projektziele erstrecken von der einfachen OCR-Erkennung von Frakturseiten bis hin zur komplexen Layoutanalyse inklusive Artikelseparierung mit einem anschließenden manuellen Qualitätssicherungsprozess.

3.2 Einsatz von OCR-D-Technologien durch IAIS

Die von Fraunhofer IAIS entwickelte Software zur Verarbeitung von digitalisierten Drucken ist effizient in C++ implementiert und trägt den Namen DocuLib.

DocuLib erlaubt die Einbindung von Komponenten des OCR-D-Frameworks. So lassen sich beispielsweise Technologien aus diversen Modulprojekten (siehe https://ocr-d.de/de/module-projects) effizient und einfach in den Workflow der DocuLib einbauen und verwenden.

Im Rahmen von ausführlichen Voruntersuchungen wurden bereits die in OCR-D entwickelten und trainierten Frakturmodelle mit bestehenden OCR-Methoden verglichen und der Gesamtworkflow für Digitalisate einer Universitätsbibliothek getestet.

3.3 Einsatz von Technologien zur Massenverarbeitung durch IAIS

Darüber hinaus werden zurzeit erste Projekte zur Massenverarbeitung von Digitalisaten durchgeführt. Hierbei soll die Software in eine übergeordnete Kubernetes-basierte Verarbeitungsumgebung integriert werden, um eine ausfallsichere und skalierbare Verarbeitung zu erreichen.

Das Fraunhofer IAIS verfügt bereits über umfangreiche Kenntnisse im Bereich Kubernetes und bei der Massenverarbeitung von Dokumenten.

4 Anwendungsszenarien mit Partner(n)

Das IAIS begreift sich als Technologiepartner von Bibliotheken, Archiven sowie anderen bestandshaltenden bzw. bestandsverarbeitenden Einrichtungen, mit denen gemeinsam praxisnahe Anwendungsszenarien erarbeitet werden sollen.

Im Fokus stehen für IAIS dabei folgende Modulprojekte (siehe auch https://ocr-d.de/de/module-projects):

1.) Skalierbare Verfahren der Text- und Strukturerkennung für die Volltextdigitalisierung historischer Drucke: Bildoptimierung
2.) Skalierbare Verfahren der Text- und Strukturerkennung für die Volltextdigitalisierung historischer Drucke: Layouterkennung
3.) Entwicklung eines Modellrepositoriums und einer Automatischen Schriftarterkennung für OCR-D

5 Projektziele

Das Ziel der Implementierung für die OCR-D-Software ist es, gemeinsam mit den Partnern die OCR-D-Technologien effizient und robust über die Fraunhofer IAIS DocuLib anzubinden und eine hohe Effizienz (Durchsatz) und Genauigkeit (Erkennungsqualität) zu erreichen. Dies soll dem Ziel dienen die OCRD-D-Technologien künftig ebenfalls in weiteren Kultureinrichtungen zum Einsatz zu bringen.

Um dieses Projektziel zu erreichen werden die einzelnen OCR-D-Module hinsichtlich ihrer Effizienz und ihrer Genauigkeit evaluiert und mit bestehenden Verfahren verglichen. Anschließend werden bei erfolgreicher Prüfung der OCR-D-Module diese in die bereits bestehende DocuLib-Softwarumgebung integriert, um eine optimale Gesamtperformance für die Massenverarbeitung zu erreichen.

5.1 Veröffentlichungen

Die von Fraunhofer IAIS entwickelten Technologien sowie die Erfahrungen bei der Implementierung von IAIS-Verfahren mit Partnern in bisherigen Projekten wurden in wissenschaftlichen Zeitschriften veröffentlicht.

Die Publikationen von Fraunhofer IAIS wurden hier gemäß des OCR-D-Workflows aufgelistet (siehe https://ocr-d.de/en/workflows).

5.2 Preprocessing

5.2.1 Binärisierung, Farbreduktion

- Constant-time locally optimal adaptive binarization
  I Konya, C Seibert, S Eickeler, S Glahn
  2009 International Conference on Document Analysis and Recognition, 738-742
- A novel preprocessing method for hectography prints based on independent component analysis
  T Kurbiel, I Konya, S Eickeler
  2011 International Conference on Document Analysis and Recognition, 1145-1149

5.2.2 Deskewing

- Fast seamless skew and orientation detection in document images
  I Konya, S Eickeler, C Seibert
  2010 20th International Conference on Pattern Recognition, 1924-1928
- Confidence measures for seamless skew and orientation detection in document images
  I Konya, S Eickeler, C Brandt
  2015 13th International Conference on Document Analysis and Recognition

5.2.3 Quality Assurance

- A new quality assessment and improvement system for print media
  M Liu, I Konya, J Nandzik, N Flores-Herr, S Eickeler, P Ndjiki-Nya
  EURASIP Journal on Advances in Signal Processing 2012 (1)

5.3 Optical Layout Recognition (OLR)

5.3.1 Layoutanalyse, Textdetektion und -segmentierung

- Machine learning for document structure recognition
  G Paaß, I Konya
  Modeling, Learning, and Processing of Text Technological Data Structures, Springer, 2011
- Histograms of Stroke Widths for Multi-script Text Detection and Verification in Road Scenes
  M Valdenegro-Toro, P Plöger, S Eickeler, I Konya
  IFAC-PapersOnLine 49 (15), 100-107

5.4 OCR, OCR-Optimierung für historische Drucke

- Character enhancement for historical newspapers printed using hot metal typesetting
  I Konya, S Eickeler, C Seibert
  2011 International Conference on Document Analysis and Recognition, 936-940
- Efficient, lexicon-free OCR using deep learning
  M Namysl, I Konya; 2019 International Conference on Document Analysis and Recognition

6 Relevante Projekte

Die o.g. generischen Module wurden als Teil von unterschiedlichen Workflows in Projekten zur Massendigitalisierung von (historischen) Zeitungen und Zeitschriften erfolgreich eingesetzt.

Name	Laufzeit	Inhalt	Ergebnis
Neue Zürcher Zeitung (NZZ)	2003 - 2005	Digitalisierung und Volltexterschließung von etwa 2 Millionen Zeitungsseiten (Jahrgänge 1780 - 2005)	Datenbank mit Bildern, Texten, Annotationen, PDF Artikel-Faksimiles integriert im NZZ Webportal
Donaukurier	2008-2009	Gesamtbestand 1,5 Mio Seiten, OCR-Zeitungsdigitaliserung,	Seitenarchiv beim Donaukurier
Liechtensteiner Volksblatt	2010-2011	Zeitungsseiten von 1900-2000, Artikelsegmentierung und -klassifikation	Auslieferung XML-Daten
Neues Deuschland	2011-2012	Zeitungsdigitalierung (200.000 Seiten), OCR, Layoutanalyse Artikelsegmentierung, manuelle Qualitätssicherung	Wurde im Stabi-Zeitungsportal veröffentlicht
THESEUS- Contentus (BMWi)	2007-2012	Entwicklung von Modulen zur semantischen Dokumentenanalyse und Verlinkung auf dem Datenbestand der Deutschen Nationalbibliothek (Bücher, Zeitungen, Ausschnitte, Fotos)	Dissertation I. Konya; Toolbox/Bibliothek zur vollautomatischen Layouterkennung
Succeed/Impact Centre of Competence (EU, FP-7)	2011-2014	Bereitstellung von Support, Training, Lernmaterial und Software-Tools für Dokumentenanalyse	Erweiterung und Optimierung von Software-Modulen für Layoutanalyse; Vorträge/Tutorials
Allgemeine Zeitung (Bayerische Staatsbibliothek BSB)	2014 - 2015	Volltexterschließung und (teilweise) Artikelsegmentierung der Allgemeinen Zeitung (~280.000 Seiten, Jahrgänge 1825 - 1929)	Seiten- und Artikel-XMLs (BSB-Schema); Schriftartenklassifizierung Antiqua/ Fraktur
Kicker-Sportmagazin (Olympia-Verlag)	2014 - 2016	Volltexterschließung und Artikelsegmentierung mit manueller Nachkorrektur des kompletten Datenbestands (~350.000 Seiten) des Kicker-magazins (Jahrgänge 1963 - 2015)	Datenbank (PDF, FhG- und Pressmatrix-XML) integriert im Webportal und Handy-App; Segmentierungs-Algorithmen für Farbscans und komplexe Layouts
KA3– eHumanities Projekt zur Analyse und Archivierung von AV-Daten (BMBF)	2015 - 2018	Segmentierungsalgorithmen für historische Schriften (lateinische, arabische und hebräische Handschriften, Initialen, Marginal- und Interlinearglossen). Etwa 5 Mill. Seitenscans vorhanden	Neue Algorithmen für Vorverarbeitung und Textzeilenerkennung aus Scans geringerer Qualität; Annotationstools
DeepER (BMBF)	2016–2018	Entwicklung einer OCR-Engine mittels Deep Learning (Neuronale Netzwerke)	OCR-Engine basierend auf Tensorflow und CNN/ RNN-Technologien

7 Andere relevante Technologien/Erfahrungen

- Importer/Exporter für verschiedene XML/XHTML/PDF Formate, wie METS/ALTO, PRImA PAGE XML, hOCR
- ML-basierte Dokumentenklassifikation