Dokumente segmentieren und Handschriften erkennen: Arbeiten mit der Plattform Transkribus

Hodel, Tobias
tobias.hodel@hist.uzh.ch
Staatsarchiv des Kantons Zürich

Lang, Eva-Maria
Eva.Lang@bistum-passau.de
Archiv des Bistums Passau

Fiel, Stefan
fiel@caa.tuwien.ac.at
Technische Universität Wien, Faculty of Informatics, Institute of Computer Aided Automation, Computer Vision Lab

Die Aufbereitung und Erkennung von handschriftlichen Dokumenten ist sowohl für Menschen als auch für Computeralgorithmen eine technische Herausforderung. Die Bearbeitung von handschriftlichem Material wird bislang von spezialisierten Experten durchgeführt, um technisch und qualitativ hochstehende Resultate aus historischen Dokumenten zu erhalten. Zur Erstellung hochwertiger Editionen ist dafür hilfswissenschaftliches Wissen (Paläographie, Editorik), historisches Hintergrundwissen und technisches Know-how gefragt.

Im Rahmen des Projekts READ (Recognition and Enrichment of Archival Data) werden unterschiedliche Aufgaben der Automatisierung (weiter-)entwickelt, um qualitativ gute Ergebnisse mit optimalem Ressourceneinsatz zu erhalten. Ein speziell dafür entwickeltes Tool ist die Software Transkribus, die die Arbeit von Experten und maschineller Erkennleistung verkoppelt. Die Software ist frei verfügbar unter www.transkribus.eu. Im Workshop wird Transkribus vorgestellt und kann durch die Teilnehmenden mit eigenen oder zur Verfügung gestellten Dokumenten getestet werden.

Transkribus unterstützt alle Prozesse vom Import der Bilder über die Identifikation der Textblöcke und Zeilen, die zu einer detaillierten Verlinkung zwischen Text und Bild führt sowie die Transkription und Annotation der Handschrift bis zum Export der gewonnen Daten in standardisierten Formaten.

Workflow in Transkribus

Um Texte zu transkribieren oder zu edieren, müssen digitale Bilder hochgeladen und danach mit Layouterkennungswerkzeugen bearbeitet werden. Die Analyse des Layouts kann automatisiert geschehen, wobei die manuelle Kontrolle und falls nötig die Nachbearbeitung im Moment noch sinnvoll ist.

Texte aus in Transkribus aufbereiteten Dokumenten können entweder mit bereits bestehenden HTR-Modellen (Handwritten Text Recognition) erkannt oder händisch erstellt werden und danach zum Training neuer Modelle genutzt werden. Insbesondere für die Bearbeitung grosser Dokumentenkorpora, die in ähnlichen Handschriften verfasst wurden, lassen sich bereits heute Effizienzgewinne und Vereinfachungen erzielen.

Aufbauend auf den Transkriptionen ist es möglich eine Vielzahl von Auszeichnungen und Annotationen innerhalb des Textes, aber auch darüber hinaus für Einzeldokumente und ganze Dokumentenbestände anzulegen. Neben der Anreicherung der Dokumente mit der Identifikation von Personen, Orten und Sachwörtern ist somit auch die Möglichkeit der Herstellung von Bestandsbeschreibungen und der Hinterlegung von Transkriptions- und Editionsvorschriften gegeben.

Ausgabeformate

Für den Export stehen unterschiedliche Formate und Ausgabeformen zur Verfügung. So ist es möglich XML-Dateien zu exportieren, die den Vorgaben der TEI entsprechen. Ausgehend davon können komplexe digitale Editionen erstellt werden, die jedoch im Unterschied zu herkömmlichen Editionen eine enge Verzahnung mit den verwendeten Bilddateien aufweisen. Dadurch werden Editionen ermöglicht, die den transkribierten Text in der Zusammenschau mit der faksimilierten Vorlage sichtbar machen (analog zu state-of-the-art Editionen, wie beispielsweise die Edition der Briefe Alfred Eschers: https://www.briefedition.alfred-escher.ch/). Daneben sind auch Ausgaben als Druckdaten (PDF) oder zur Weiterbearbeitung für Textverarbeitungsprogramme (DOCX) implementiert. Schliesslich ist auch ein Export im PAGE-Format (zur Anzeige in Viewern für OCR gelesene Dokumente, Pletschacher, 2010) sowie als METS (Metadata Encoding and Transmission) möglich.

Die Speicherung der Dokumente erfolgt in der Cloud (gehostet auf Servern der Universität Innsbruck). Die importierten Daten bleiben auch während der Bearbeitung unverändert im Dateisystem liegen und werden ergänzt durch METS und PAGE XML, letztere in eigenem Unterordner. Alle bearbeiteten Dokumente und Daten bleiben somit in den unterschiedlichen Bearbeitungsstadien nicht nur lokal verfügbar, sondern können für Projektmitarbeitende geteilt werden. Dank elaboriertem *user-management* ist die Zuteilung von Rollen möglich. Die Erkennprozesse werden serverseitig durchgeführt, sodass die Ressourcen auf den lokalen Rechnern nicht strapaziert werden. Transkribus ist mit JAVA und SWT programmiert und kann daher plattformunabhängig (Windows, Mac, Linux) genutzt werden.

Zielpublikum

Die Plattform ist für unterschiedliche Gruppen konzipiert. Einerseits für GeisteswissenschaftlerInnen, die selbst Transkriptionen und Editionen historischer Dokumente erstellen möchten. Andererseits richtet sich die Plattform an Archive, Bibliotheken und andere Erinnerungsinstitutionen, die handschriftliche Dokumente in ihren Sammlungen aufbewahren und ein Interesse an der Aufbereitung des Materials haben. Angesprochen werden sollen auch Studierende der Geistes-, Archiv- und Bibliothekswissenschaften mit einem Interesse an der Transkription historischer Handschriften.

Das Ziel, eine robuste und technisch hochstehende Automatisierung von Layout und Handschrift, lässt sich nur durch die enge Zusammenarbeit zwischen Geisteswissenschaftlern und Computerspezialisten erreichen, die bezüglich Datenqualität und Herstellung von Transkriptionen von unterschiedlichen Voraussetzungen und Ansprüchen ausgehen. Die Algorithmen werden daher nicht nur bis zu einem Status als *proof-of-concept* erarbeitet, sondern bis zur Praxistauglichkeit verfeinert und in grösseren Forschungs- und Aufbewahrungsumgebungen getestet und verbessert. Die Computerwissenschaftler sind entsprechend ebenfalls ein wichtiges Zielpublikum, wobei bei ihnen weniger die Nutzung der Plattform als das Beisteuern von Software(teilen) anvisiert wird. Die eingespeisten Dokumente und Daten bleiben privat und vor dem Zugriff Dritter geschützt. Von Projektseite können vorgenommene Arbeitsschritte zwecks besserem Verständnis der ausgeführten Arbeiten und letztlich der Verbesserung der Produkte ausgewertet werden.

Layout- und Texterkennung

Die zwei zentralen Automatisierungsprozesse basieren auf Algorithmen, die in laufenden Forschungsprojekten entwickelt und verbessert werden. Die *document image analysis* (DIA) versucht Textblöcke zu identifizieren und von Dreck, Scanfehlern und anderen Störsignalen zu unterscheiden, wobei zwischen handschriftlichen und gedruckten Textblöcken differenziert wird (Zagoris 2012; Stamatopoulos 2015).

In Transkribus werden auf der Layouterkennung aufbauend zwei *handwritten text recognition*-Engines (HTR) angeboten, die auf unterschiedlichen technischen Grundlagen basieren: Erstens kann eine nach dem Hidden Markov Model (HMM) operierende HTR der Technischen Universität Valencia angewählt werden (Toselli 2015, Puigcerver 2015). Zweitens kann ein Model basierend auf rekurrierenden neuronalen Netzwerken der Universität Rostock genutzt werden (Leifert 2016).

Transkribus und das gesamte Forschungsnetzwerk will die verfügbaren technischen Möglichkeiten den Endnutzern nach möglichst gängigen Workflows aufbereiten, so dass dem schnellen Praxiseinsatz keine Hindernisse im Weg stehen. Im Gegenzug wird die Nutzung im grossen Umfang erhofft, die den Subprojekten wichtige Trainingsdaten und Aufschlüsse bezüglich der Nutzung und den Problemen mit den Algorithmen sowie dem Graphical User Interface geben. Tests zum Einsatz der Technik in Archiven und Bibliotheken und unter unterschiedlichen Bedingungen werden momentan getestet und evaluiert.

Als Businessmodel ist eine Überführung des Forschungsprojekts in eine Kooperative geplant, die den Stakeholdern möglichst niederschwellige und kostengünstige Angebote unterbreiten soll (Mühlberger, Preprint). Somit vereint das Projekt READ die unterschiedlichsten Ansprüche an Automatisierungs- und Erkennungsroutinen und orientiert sich dabei an gängigen Arbeitsformen im Kontext mit handschriftlichen Dokumenten (siehe auch die Projekthomepage: http://read.transkribus.eu).

Aus- und Seitenblicke im Workshop

Zwei unterschiedliche Forschungsaspekte aus READ werden im Rahmen des Workshops als Inputs demonstriert:

Einerseits der Umgang mit einer speziellen Dokumentenform, Kirchenbüchern, in denen stark strukturierte Daten aus Pfarreien gesammelt wurden (Wurster, 2014 / 2015). Aufgrund der Strukturerkennung und der HTR wird es möglich, spezialisierte Suchroutinen zu produzieren.

Andererseits können aufgrund der erhobenen Daten und durch c *omputer vision* Profile der Schreibenden erstellt werden, die die Identifikation der Personen als Schreibende weiterer Dokumente naheliegend macht (Fiel, 2012). Beide Anwendungen versprechen für die Geisteswissenschaften neue Zugänge zu grossen Datenschätzen, die in den handschriftlichen Beständen gehoben werden können.

Programm/Ablauf des Workshops

Einführung in Transkribus (Tobias Hodel, Zürich): 30‘
Aufbau und Funktionieren des Programms, Demonstration des Gebrauchs anhand von Beispielen. Aufzeigen der Möglichkeiten zum Einsatz der Automatisierungen.

Strukturierte Daten in Kirchenbüchern (Eva-Maria Lang, Passau): 30‘
Demonstration vom Umgang mit Kirchenbüchern, einer spezifischen und stark standardisierten Dokumentform, die mit Transkribus aufbereitet werden. Eine Suche in den Dokumenten wird über eigene Routinen und Abfragemöglichkeiten gewährleistet.

Selbstständiges Arbeiten der Teilnehmenden mit Transkribus: 90‘
Die Möglichkeiten und Grenzen von Transkribus sollen von den Teilnehmenden (wenn möglich mit eigenen Dokumenten) selbst ausgetestet werden.

Schreiberidentifizierung (Stefan Fiel, Wien): 30‘
Ein über Transkribus hinausgehender Teil des Projekts beschäftigt sich mit *computer vision*. Ziel ist die Identifizierung unterschiedlicher Personen als Schreibende. Stefan Fiel berichtet über den Stand der Forschung und wie Teilnehmende die Hände wichtiger Schreibender zur Verfügung stellen können.

Diskussion über Vor- und Nachteile der Software: 45‘
Inklusive Evaluation des Tools und der Veranstaltung. Feedbacks werden eingeholt, zur Verbesserung der Software (usability, Umfang und Leistung der Automatisierungen etc.).

Das Projekt READ und somit die Weiterentwicklung von Transkribus werden finanziert durch einen Grant der Europäischen Union im Rahmen des Horizon 2020 Forschungs- und Innovationsprogramms (grant agreement No 674943).

Kontaktdaten aller Beitragenden (inkl. Forschungsinteressen)

Tobias Hodel, Staatsarchiv des Kantons Zürich, Winterthurerstrasse 170, CH-8057 Zürich, Schweiz; (Digital Humanities; Handwritten Textrecognition; eArchiving; Information Retrieval).

Eva-Maria Lang, Archiv des Bistums Passau, Luragogasse 4, DE-94032 Passau, (Automatic Text Recognition, Digital Archives, Image Recognition and Information Retrieval, Software Architecture).

Stefan Fiel, Technische Universität Wien, Faculty of Informatics
Institute of Computer Aided Automation, Computer Vision Lab, Favoritenstr. 9/183-2, A-1040 Vienna, Austria; (Bilderverarbeitung und Dokumentenanalyse).

Zahl der möglichen Teilnehmerinnen und Teilnehmer
30-40 Personen (auch abhängig von der Raumgrösse)

Benötigte technische Ausstattung:
Allgemein: Beamer, evtl. Whiteboard.
Teilnehmende: Eigener Rechner (wenn möglich Installation von Transkribus; Hilfe zur Installation von Transkribus wird 30 Minuten vor der Veranstaltung angeboten)

Anmeldungen und Rückfragen bitte an tobias.hodel@ji.zh.ch

Bibliographie

Fiel, Stefan / Sablatnig, Robert (2012): „Writer Retrieval and Writer Identification using Local Features“, in: *10th IAPR International Workshop on Document Analysis Systems* http://www.ict.griffith.edu.au/das2012/attachments/FullPaperProceedings/4661a145.pdf .

Leifert, Gundram / Strauß, Tobias / Grüning, Tobias / Labahn, Roger (2016): *Cells in Multidimensional Recurrent Neural Networks* https://arXiv.org/abs/1412.2620v02 .

Mühlberger, Günter / Colutto, Sebastian / Kahle, Philip (Preprint): *Handwritten Text Recognition (HTR) of Historical Documents as a Shared Task for Archivists, Computer Scientists and Humanities Scholars: The Model of a Transcription & Recognition Platform (TRP).*

Pletschacher, Stefan / Antonacopoulos, Apostolos (2010): „The PAGE (page analysis and ground-truth elements) format framework“, in: *Proc. ICPR* 257–260.

Puigcerver, Joan / Toselli, Alejandro Héctor / Vidal, Enrique (2015): „Probabilistic interpretation and improvements to the hmm-filler for handwritten keyword spotting“, in: *13th international conference on document analysis and recognition (ICDAR).*

Stamatopoulos, Nikolaos / Gatos, Basilis (2015): „Goal-oriented performance evaluation methodology for page segmentation techniques“, in: *13th international conference on document analysis and recognition (ICDAR)* 281–285.

Toselli, Alejandro Héctor / Vidal, Enrique (2015): „Handwritten text recognition results on the Bentham collection with improved classical n-gram-HMM methods“, in: *International workshop on historical document imaging and processing (HIP).*

Wurster, Herbert W. (2015): „Schritt für Schritt ins Internet – Europas Matriken online“, in: *insights: Archives and people in the digital age* 2: 16–17.

Wurster, Herbert W. (2014): „Matrikeln - Ein kulturhistorischer Blick auf die Kirchenbücher“, in: *Zeitschrift für bayerische Kirchengeschichte* 83: 87–93.

Zagoris, Konstantinos / Pratikakis, Ioannis / Antonacopoulos, Apostolos / Gatos, Basilis / Papamarkos, Nikos (2012): „Handwritten and Machine Printed Text Separation in Document Images Using the Bag of Visual Words Paradigm“, in: *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference* 103–108 10.1109/ICFHR.2012.207.