Institut für Nachrichtentechnik, TU Braunschweig (IFN), Herzog August Bibliothek Wolfenbüttel (HAB)

Vom Zeichen zur Schrift.

Mit Mustererkennung zur automatisierten Schreiberhanderkennung in mittelalterlichen und frühneuzeitlichen Handschriften

"Deutschland befindet sich in einer Phase intensiv betriebener und mit einem hohen finanziellen Aufwand verbundener Digitalisierung seiner historischen Bestände. Für die Mediävistik und Frühneuzeitforschung stellt hierbei die Digitalisierung der dem Mittelalter und der Renaissance entstammenden Hss. ein zentrales Feld dar."[1] Die Nutzung der Digitalisate allein als digitale Lesekopie durch den betrachtenden Forscher würde das Erkenntnispotential, das dem Digitalisat selbst innewohnt, schlechterdings vergeuden.

Die Gewinnung und Nutzbarmachung der digital vorhandenen Informationen, welche die Analyse des dem Digitalisat zugrunde liegenden materiellen Objekts flankieren und sinnvoll ergänzen können, ist eine der ältesten Fragen digitaler Geisteswissenschaften. Eine zentrale Anwendung ist die Optical Character Recognition (OCR oder automatisierte Texterkennung), die der Herstellung eines maschinenlesbaren Textes aus bildhaft vorliegender Information dient. Sie stößt allerdings bei der Verarbeitung von Handschriftendigitalisaten an ihre Grenzen, was sich anhand des formalen Ablaufs einer Texterkennung veranschaulichen lässt:

1. Anfertigen eines Digitalisates in geeigneter Qualität gegebenenfalls Bildkorrekturen wie das Geraderücken schräg aufgenommener Seiten, Glättung von Rundungen aufgrund von Materialbiegung, etc.
2. Binarisierung der Farbwerte zur deutlichen Trennung von Schrift und Hintergrund
3. Segmentierung der Schrift, z.B. in Linien und Einzelworte
4. Mustererkennung, bei der zu erkennende Formen mit einem vorhandenen Zeichenvorrat verglichen werden
5. Im Falle der Übereinstimmung Zuweisung eines Zeichencodes nach üblicher Textkodierung (=UTF-8)

Die eigentliche Herausforderung an die OCR für Handschriften liegt in Arbeitsschritt 5, da hier zunächst durch langwieriges (und meist manuelles) Erstellen eines Zeichenvorrat, die sogenannte ‘ground truth’, angelegt werden muss, auf deren

[1] Thomas Haye und Stephan Müller: Mittelalter-Philologie im Internet. 38. Beitrag: Digitalisierung mittelalterlicher Handschriften aus Sicht der Forschung, in: Zeitschrift für deutsches Altertum un deutsche Literatur 140 (2011), S. 416–420, hier S. 416.

Grundlage das Training eines Klassifikators erfolgt. Dieser Klassifikator ist Kern des OCR-Systems und ermöglicht es, Muster (pattern) in der Vorlage Zeichen aus dem Zeichenvorrat zuzuordnen. Diese auf einer festgesetzten Wahrheit beruhende Beziehung zwischen den Bildmustern und den kodierten Zeichen ist nur für die Typen von Mustern gültig, die zum Training verwendet wurden, also z. B. für eine Schrifttype oder eine bestimmte Schriftform. Da die Handschrift jedes Schreibers eigene Charakteristika aufweist, welche sie zumindest von anderen Händen unterscheidbar macht, können die in den meisten Handschriftendigitalisaten aufgefundenen Muster nicht ohne weiteres eindeutig kodierten Zeichen zugewiesen werden, so dass die automatische Erstellung maschinenlesbaren Textes anhand von Digitalisaten in den allermeisten Fällen zu unbrauchbaren, weil fehlerbehafteten Ergebnissen führt. Sollte es jedoch gelingen, anhand spezifischer Merkmale eine Schreiberhand (unter Abstraktion von nicht mehr quantifizierbaren Abweichungsquellen wie Lebensalter und Tagesform des Amanuensis oder dem Zustand der Schreibmaterialien und -utensilien) von anderen Schreiberhänden abzugrenzen, so könnte damit die wichtige Fragestellung nach dem Schreiber automatisiert werden.

In unserem Projekt haben wir also das Untersuchungsziel umgekehrt: Als Ergebnis der Analyse von Handschriftenabbildungen steht nicht ein elektronischer Text, sondern die Identifikation der schreibenden Hand bzw. Hände. Eine mögliche Vorgehensweise zur Lösung dieser Aufgabe ist es, anhand einer von einem sicher zugeordneten Schreiber angefertigten Handschrift die Charakteristika dieser Schrift als ‘ground truth’ anzutrainieren. Hierzu gehören Buchstabengröße und –abstand, Dichte des Schriftbildes, Neigung u. a., aber nicht notwendig, wie in der klassischen, vom forschenden menschlichen Auge ausgehenden Paläographie, einzelne Buchstabenformen. Basierend auf diesen Charakteristika wird überprüft, ob es möglich ist, diese Hand in anderen Handschriften nachzuweisen. Die Erkennungsgenauigkeit muss dazu aufgrund des Trainings mit einer Handschrift in einem anderen Codex über einem zu definierenden Schwellwert (threshold) liegen, um als Indiz gewertet zu werden, dass derselbe Schreiber die Handschrift geschrieben haben könnte. Der angestrebte Algorithmus würde damit sonstige paläographische oder kodikologische Befunde unterstützende bzw. ergänzende Argumente zur Verifikation von unsicheren Zuschreibungen liefern. Im Gegenzug müsste das Unterschreiten dieses Schwellwertes Argumente für Falsifikationen solcher Zuschreibungen ermöglichen. Darüber hinaus sollte es bei entsprechender Materialbasis und angepassten Schwellwerten möglich sein, Schriftfamilien zu unterscheiden. Würde man die Eigenschaften der wichtigsten Schriftfamilien und räumliche Zusammenhänge, wie etwa insulare Schriftformen in entsprechenden Zeichenvorräten sammeln, wäre es denkbar, Hinweise auf die Datierung und Lokalisierung einer Handschrift aus diesen Vergleichsdaten zu ermitteln.

Ein alternativer Ansatz wäre eine Betrachtung des Schriftbildes als Ganzes und die Bestimmung der Ähnlichkeit von Texten. So könnte ein Grad der Ähnlichkeit eine

Aussage zulassen, ob ein Text eher vom selben oder von einem anderen Schreiber stammt.

Um die aufgestellten Thesen über die Nützlichkeit solcher Erkennungsalgorithmen zu testen, musste zunächst passendes digitalisiertes Material ausgewählt werden. Dabei fiel die Entscheidung auf die wichtigste frühmittelalterliche Buchschrift, die karolingischen Minuskel, die sich durch einheitliche, relativ stark standardisierte Formen und eine meist geringe individuelle Varianz auszeichnet. Diese allgemeinen Spezifika unterstützen die Brauchbarkeit von Digitalisten karolingischer Handschriften ebenso wie deren meist hohes kodikologisches Niveau, das bei entsprechender Fotoaustattung nur wenig Nachbearbeitung erforderlich macht. Mit den im Rahmen des "Europeana Regia"-Projekts digitalisierten Codices Weissenburgenses der HAB steht eine breite Materialbasis zur Verfügung, die weitere Vorteile aufweist: Die weitaus meisten Codices stammen aus dem Skriptorium des Klosters Weißenburg im Elsass, sind also regional und zeitlich gut einzuordnen. Mit dem den DFG-Richtlinien entsprechenden Katalog von Hans Butzmann[2] sind diese Handschriften außerdem kodikologisch gut erschlossen. Die Beschreibungen liefern die Vorlagen der Schreiberidentifikation, die es zu verifizieren (oder falsifizieren) galt.

Aus diesem Bestand haben wir in einem ersten Schritt eine Handschrift gewählt, die von einem identifizierten und uns bekannten Schreiber geschrieben worden ist. Die Entscheidung fiel auf Cod. Guelf. 62 Weiss., der nach den Angaben des Kolophons zwischen 819 und 826 von dem Mönch Waldmann während des Abbatiats von Gerhoh geschrieben worden ist.[3] Auf diese Handschrift wurde der Algorithmus angewendet. Weiterhin musste es eine zweite Handschrift geben, die teilweise, aber nicht vollständig von demselben Schreiber angefertigt worden ist, um festzustellen, ob das Verfahren die Handwechsel und damit die Anteile der einzelnen Schreiber erkennt. Hier bot sich Cod. Guelf. 63 Weiss. an, an dem wiederum Waldmann selbst mit zwei weiteren Kopisten, vielleicht seinen Schülern, gearbeitet hat.[4]

Ansätze

In unseren bisherigen Arbeiten haben wir uns zunächst überwiegend mit dem alternativen Ansatz beschäftigt, bei dem die Merkmale zur Schreibererkennung von dem Textblock eines Blattes abgenommen werden. Wie erwähnt, wird nicht versucht, einzelne Zeichen als pattern zu identifizieren, sondern die Handschriftenseite wird als Ganzes interpretiert. Der hochdimensionale Merkmalsvektor einer Seite wird dann mit anderen Seiten verglichen. Je ähnlicher die Merkmalsvektoren sind, desto wahrscheinlicher ist es, dass diese von der selben Hand stammen. Um die Extraktion

[2] Hans Butzmann: Die Weissenburger Handschriften,Frankfurt am Main 1964 (Kataloge der Herzog August Bibliothek Wolfenbüttel: Neue Reihe, Bd. 10)
[3] Ebd., S. 202.
[4] Ebd., S. 203.

des Textes möglichst einfach zu gestalten, wurden zunächst „typische“ Seiten, vor allem solche ohne Buchschmuck, ausgewählt und für die Gewinnung der Eigenschaften zu Grunde gelegt.

Bei geeigneter Auswahl von Merkmalen konnte gezeigt werden, dass der Schreiber der Handschrift 62 Weiss. auch in 63 Weiss. tätig war. Es konnten die Seiten auf denen er schreibt von denjenigen Seiten geschieden werden, auf denen andere Schreiber tätig sind.

Ergebnisse pro Handschrift

Nach diesem erfolgreichen ersten Test wurden die Algorithmen auf weitere Beispiele angewendet. Diese waren die Handschriften Cod. Guelf. 18 Weiss., 10 Weiss. und 14 Weiss. Dort soll es Überschneidungen der Haupthände geben, die in den Katalogen aber nicht näher lokalisiert sind.

Zusammenfassung / Ausblick

Die ersten Tests haben gezeigt, dass eine automatisierte Schreibererkennung basierend auf einem hochdimensionalen Merkmalsvektor einer Textseite möglich ist. Umfangreichere Tests sind nötig, um die universelle Anwendbarkeit zu überprüfen. Als Alternative wird parallel dazu an einem Ansatz gearbeitet, der zeilenorientiert auf völlig anderen Merkmalen arbeitet und eine Klassifikation basierend auf trainierten Klassifikatoren ermöglicht.