Automatische Typenbestimmung in historischen Drucken

,
weichsel@uni-mainz.de
Johannes Gutenberg Universität Mainz, Deutschland

,
vincent.christlein@fau.de
Friedrich-Alexander-Universität Erlangen-Nürnberg, Deutschland

Eine zentrale Methode der Analytical Bibliography ist die Bestimmung der Drucktype einer Inkunabel. Da im Frühdruck nur sehr beschränkt mit Schriften gehandelt wurde, ermöglicht diese Bestimmung oft die (näherungsweise) Datierung und Firmierung eines Drucks. Als Hilfsmittel dafür steht seit Jahrzehnten das Typenrepertorium zur Verfügung, bei dem der Benutzer mithilfe der Form von ›M‹(gebrochene Schriften) und ›Qu‹(Antiqua) sowie der Schriftgröße eine Vorauswahl aus den mehreren Tausend bekannten Typen trifft und dann Tafeln mit dem vollständigen Zeichensatz einer Schrift mit dem zu bestimmenden Druck abgleicht. Diese Methode ist erprobt, aber langwierig und nur für Spezialisten zu handhaben, auch wenn das umständliche Wälzen des Tafelwerks inzwischen einer Datenbank (Staatsbibliothek zu Berlin 2014) gewichen ist.

Diese Methode soll durch ein Werkzeug ergänzt werden, das auf die inzwischen in großem Umfang zur Verfügung stehenden Volldigitalisate von Inkunabeln mit Mitteln der Mustererkennung analysiert und dem Nutzer vollautomatisch die in einem vorliegenden Digitalisat verwendete Type bzw. eine Liste der wahrscheinlichsten Typen ausgibt.

Die automatische Identifikation historischer Druckschriften ist bisher nicht bearbeitet. Es kann aber auf Ansätze zur Identifikation von Schreiberhänden zurückgegriffen werden, die zwar nicht 1:1 übertragbar sind, aber viele Probleme vorwegnehmen. Konkret bedeutet das eine Merkmalsextraktion an den Konturen der Schrift, die so gewonnenen lokale Merkmale werden anschließend zu einem globalen Merkmalsvektor zusammengefasst und klassifiziert.

Ein großer Vorteil für die Umsetzung dieses Ansatzes ist die bestehende im Gesamtkatalog der Wiegendrucke vorliegende Zuordnung von ca. 15.000 digitalisierten Inkunabelausgaben den jeweiligen Schriften des Typenrepertoriums. Dieser große Bestand an Ground-Truth-Daten ermöglicht es, für bestimmte Typen zu trainieren und so die Gesamterkennungsrate signifikant zu erhöhen. Dabei sinkt durch die große Datenmenge die Abhängigkeit von variierenden Aufnahmebedingungen bei Digitalisaten. Außerdem verbessert sich die Unterscheidung von Type und Hintergrund (Bedruckstoff), der bei kleineren Beständen an Ground-Truth-Daten leicht mitklassifiziert wird.

Das fertige Werkzeug soll einerseits die Arbeit von Inkunabelforschern erheblich beschleunigen und auch nahestehenden Disziplinen die Bestimmung von Typen ermöglichen. Gleichzeitig könnte diese Methode auch die Fortschreibung des Typenrepertoriums ins 16. Jahrhundert ermöglichen, für das es bisher kein derartiges Verzeichnis gibt, weil die händische Bestimmung angesichts der im Lauf des 16. Jahrhunderts exponentiell steigenden Druckproduktion ausgeschlossen ist. Je nach erreichbarer Präzision der vollautomatischen Erkennung soll eine halbautomatische Erkennung ergänzend implementiert werden, bei der der Benutzer wenige typische Zeichen markiert und so die Zuordnungsgenauigkeit wenigstens in den Bereich der analogen Bestimmungstechnik bringt, dabei aber immer noch wesentlich schneller und einfacher zu bedienen ist.

Der Vortrag präsentiert die Ergebnisse eines Proof of Concept, der mit Einzelseiten aus 100 Inkunabeldigitalisaten, von denen jeweils zwei aus derselben Schrift gesetzt waren, die Gangbarkeit dieses Ansatzes untersucht hat. Es wurde dazu eine Methode zur Schreiberidentifizierung adaptiert, die sehr erfolgreich zeitgenössische Schriften dem richtigen Schreiber zuordnen kann (Christlein / Bernecker / Angelopoulou 2015). Dabei handelt es sich um einen ganzheitlichen Ansatz, bei dem auf Basis einer Bilddatei Merkmale an der Schrift berechnet werden. Für diese Studie wurde der Textbereich im Bild manuell markiert und anschließend binarisiert (Sauvola / Pietkäinen 2000). Zusammenhängende Komponenten wurden anhand ihrer Flächengröße und Breiten-Höhenverhältnisses gefiltert, so dass möglichst nur Schrift extrahiert wird. An den Konturen der Schrift werden anschließend Zernike-Momente berechnet, welche anschließend mittels VLAD zu einem globalen Merkmalsvektor aggregiert werden. Diese Merkmalsvektoren dienen anschließend zur Typenbestimmung.

Erste Ergebnisse zeigen, dass diese Methode einen möglichen Weg zur Typenbestimmung darstellt: Testete man eines der Dokumente und verglich es mit den restlichen 99 Dokumenten so wurde in 45% der Fälle das Dokument mit derselben Type als wahrscheinlichstes Dokument zurückgeliefert. Betrachtet man die Liste der zehn wahrscheinlichsten Dokumente, so befand sich die richtige Type mit 77% Wahrscheinlichkeit unter ihnen. Anstatt also händisch jede mögliche Type zu überprüfen, kann eine nach Wahrscheinlichkeit sortierte Liste zur Bestimmung benutzt werden und somit den Aufwand der Typenbestimmung drastisch verringern. Dabei ist zu berücksichtigen, dass bei der Durchführung dieses Experiments keinerlei Rücksicht auf unterschiedliche Auflösungen oder anderen Störelemente (Artefakte, schlechte Bildqualität, etc.) genommen wurde und die

Datenbasis noch sehr gering war. Bei Berücksichtigung von Störfaktoren und insbesondere bei höheren Datenmengen ist eine signifikante Steigerung der Erkennungsrate zu erwarten.

Damit lässt sich bereits an diesem Proof of Concept zeigen, dass die Methode im Prinzip funktioniert und eine wertvolle Ergänzung zur konventionellen Vorgehensweise darstellen kann, auch wenn noch zu klären bleibt, ob sie die analoge Bestimmung mittelfristig ersetzen kann.

Bibliographie

Christlein, Vincent / Bernecker, David / Angelopoulou, Elli (2015): "Writer Identification using VLAD encoded Contour-Zernike Moments", in: *Document Analysis and Recognition (ICDAR) 2015*. 13th International Conference 23–26 August 2015, Nancy 906-910.

Eisermann, Falk / Duntze, Oliver (2014): "Auf der Spur der seltsamen Typen. Das digitale Typenrepertorium der Wiegendrucke", in: *Bibliotheksmagazin* 3: 41–48 https://www.bsb-muenchen.de/fileadmin/imageswww/pdf-dateien/bibliotheksmagazin/BM2014-3.pdf [letzter Zugriff 29. Dezember 2015].

Haebler, Konrad (1905): *Typenrepertorium der Wiegendrucke*. Abt. I. Deutschland und seine Nachbarländer. Halle: Haupt.

Sauvola, Jaakko / Pietikäinen, Matti (2000): "Adaptive document image binarization", in: *Pattern Recognition* 33, 2: 225-236.

Staatsbibliothek zu Berlin (2014): *Typenrepertorium der Wiegendrucke digital* http://tw.staatsbibliothek-berlin.de [letzter Zugriff 29. Dezember 2015].