Erfahrungen bei der Erfassung eines Zeitungsarchivs

Experiences in Capturing a Paper Archive

Dr. Wolfgang Schade, Melanie Irrgang, Martin Tölle
Gesellschaft zur Förderung angewandter Informatik e.V. (GFaI)
Volmerstr. 3, 12489 Berlin
Tel.: 030 814 563 470, Fax: 030 814 563 302
E-Mail: schade@gfai.de, Internet: www.gfai.de

Zusammenfassung:
Das Herder-Institut Marburg ist eine der zentralen Einrichtungen der historischen Ostmitteleuropa-Forschung in Deutschland. Es beschäftigt sich intensiv mit der Geschichte und Kultur Polens, Estlands, Lettlands, Litauens, Tschechiens, der Slowakei und der Region Kaliningrad. Das Institut bietet eine der besten Spezialbibliotheken und verfügt über umfangreiche Sammlungen (Bilder, Karten, archivalische Materialien) mit Schwerpunkt Baltikum sowie der historischen deutschen Ostgebiete. Ein bedeutender Anteil besteht in der Sammlung von Zeitungsausschnitten. Im Rahmen eines Forschungsprojektes des BMWi-INNO-KOM-Programms werden von der GFaI Untersuchungen durchgeführt, um die Erfassung der interessierenden Informationen möglichst automatisch und damit effektiver zu gestalten.

Abstract:
The Herder Institute in Marburg is one of the central institutions of historical research on East Central Europe in Germany. It works intensively with the history and culture of Poland, Estonia, Latvia, Lithuania, the Czech Republic, Slovakia and the Kaliningrad region. The Institute offers one of the best special libraries and has extensive collections (images, maps, archival materials) with emphasis Baltic States as well as historical German East regions. A significant part of it is a collection of newspaper clippings. In a project of the BMWi-INNO-KOM program GFaI investigates capturing the interesting information as automatically as possible and thus for a more effective procedure.

1. Material

Die insgesamt über 5 Millionen Seiten Material sind in Ordner abgelegt, in unterschiedliche Kategorien unterteilt und dort alphabetisch geordnet. Kategorien sind z.B. deutsche Persönlichkeiten, polnische Persönlichkeiten, Orte. Bei dem uns zur Zeit vorliegenden Material handelt es sich um ganze Zeitungsseiten, aufgeklebte Zeitungsausschnitte, teilweise auch Schreibmaschinenseiten von zumeist deutschen Quellen, aber auch polnische, tschechische, lettische, litauische und russische Ausschnitte sind vorhanden.

2. Ergebnisse der kommerziellen OCR (Finereader)

Von Schreibmaschinendokumenten abgesehen, liefert das kommerzielle Schrifterkennungssystem auf dem uns vorliegenden Material i.a. gute Ergebnisse, davon abweichend sind nur Artikel aus polnischen und tschechischen Zeitungen (Grund: Zeitungspapier und Druckbild der 70er Jahre). Die Software erkennt existierende Regionen und ist in der Lage, die Ergebnisse positionsgetreu wieder abzubilden. Bereiche, die nicht als Textbereich erkannt werden, werden dabei als Bildbereich an die entsprechende Position eingefügt.
Allerdings gibt es auch Anforderungen, die von der Software nicht gelöst werden. Zwar erkennt die OCR i.a. Zeitungsspalten. Aber bei Vorhandensein einer Abbildung oder Überschrift, die diese Spalten unterbricht, ist nicht festgelegt, ob der Artikel unterhalb der Abbildung oder neben der gerade behandelten Teilspalte fortgesetzt wird. So werden bei mehrspaltigen Zeitungsartikeln, die von Bildern unterbrochen werden, die einzelnen Spaltenteile nicht in der richtigen Reihenfolge geliefert, auch stehen dann die Bildunterschriften zwischen dem Artikeltext, und u.U. die Artikelüberschrift nicht am Anfang des Artikels.

Stempel mit dem Zeitungsnamen werden selten erkannt, erst recht nicht, wenn der Handstempel eine andere Ausrichtung hat als der aufgeklebte Zeitungsausschnitt.

3. erwünschte Struktur der erfassten Informationen

Um die späteren Abfragemöglichkeiten zu unterstützen, sollen u.a. folgende Informationen gesondert erfasst werden;
Sprache des Artikels
Name der Zeitung/Zeitschrift
Erscheinungsdatum
Name der betreffenden Persönlichkeit, ergänzbar durch Geburts- und Sterbedatum sowie seine PND-Nummer.
Beschreibung des Artikels
Name des Artikelautors
Überschrift des Artikels
Artikeltext, Abschnitte in der richtigen Reihenfolge
Kennzeichnung, ob vollständige Freigabe im Internet möglich
Erfassung der Abbildungen
Erfassung der Unterschriften zu den Abbildungen
Erfassung des Inhalts der Abbildungen
Autor(Fotograf) der Abbildung
Kennzeichnung, ob Freigabe der Abbildung im Internet möglich

4. Einfluss der Scanparameter und Bildverbesserung

Unsere Untersuchungen ergaben, dass schon die Einstellung der Scanparameter die OCR-Erkennungsrate beeinflusst.
Bei der Verwendung des ABBYY FineReader 9.0 (latest version) wird ein Bild in Bereiche zerlegt. Diese Bereiche werden von FineReader dann nach ihren Inhalten analysiert und gegebenenfalls der Text ermittelt. In vielen Fällen kommt es jedoch vor, dass Bereiche wegen ihrer Beschaffenheit nicht als Textfeld deklariert werden und folglich nicht der Texterkennung zugeführt werden.

Mit dem an der GFal entwickelten adaptiven Verfahren der Bildvorverarbeitung konnte die Identifizierung von Textfeldern durch FineReader wesentlich verbessert werden:

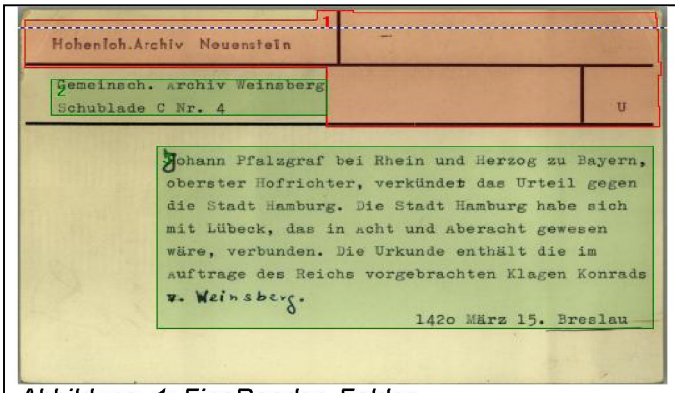


Abbildung 1: FineReader- Felder

Das nebenstehende Bild zeigt beispielhaft eine automatisierte Einteilung von ABBYY FineReader 9.0. Die **grünen** Bereiche sind von der Software als Textfelder deklariert worden, die roten Felder als Bildbereich. Man sieht, dass Abby nicht alle Schriftfelder erkennt.

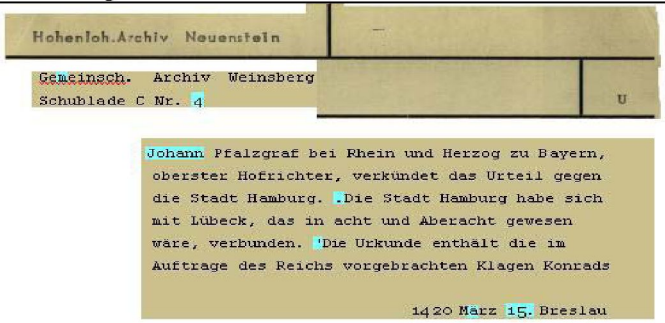
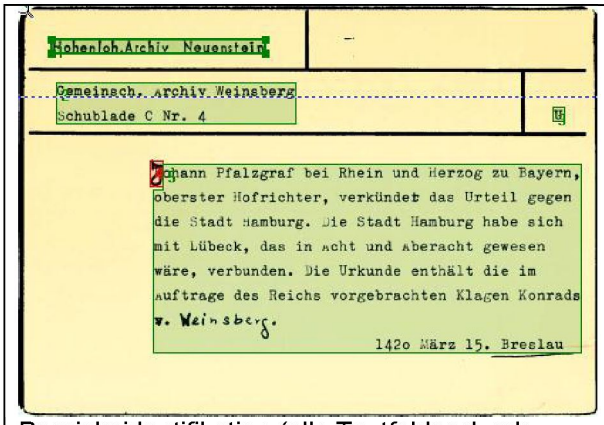
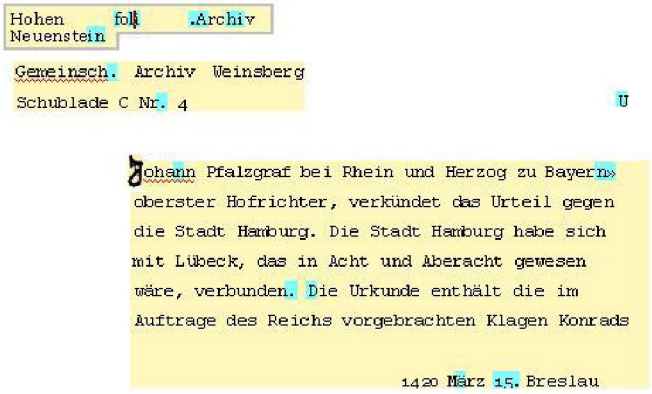


Abbildung 2 : FineReader OCR-Ergebnis

Das Ergebnis zeigt die gute Texterkennung in den oben grün gekennzeichneten Textbereichen. Unerkannte (rote) Bereiche werden als Bild wiedergegeben. („Hohenloh. Archiv Neuenstein“ und das „U“ wurden nicht als Schriftbereich deklariert).



Bereichsidentifikation (alle Textfelder durch FineReader nach Bildvorverarbeitung erkannt)

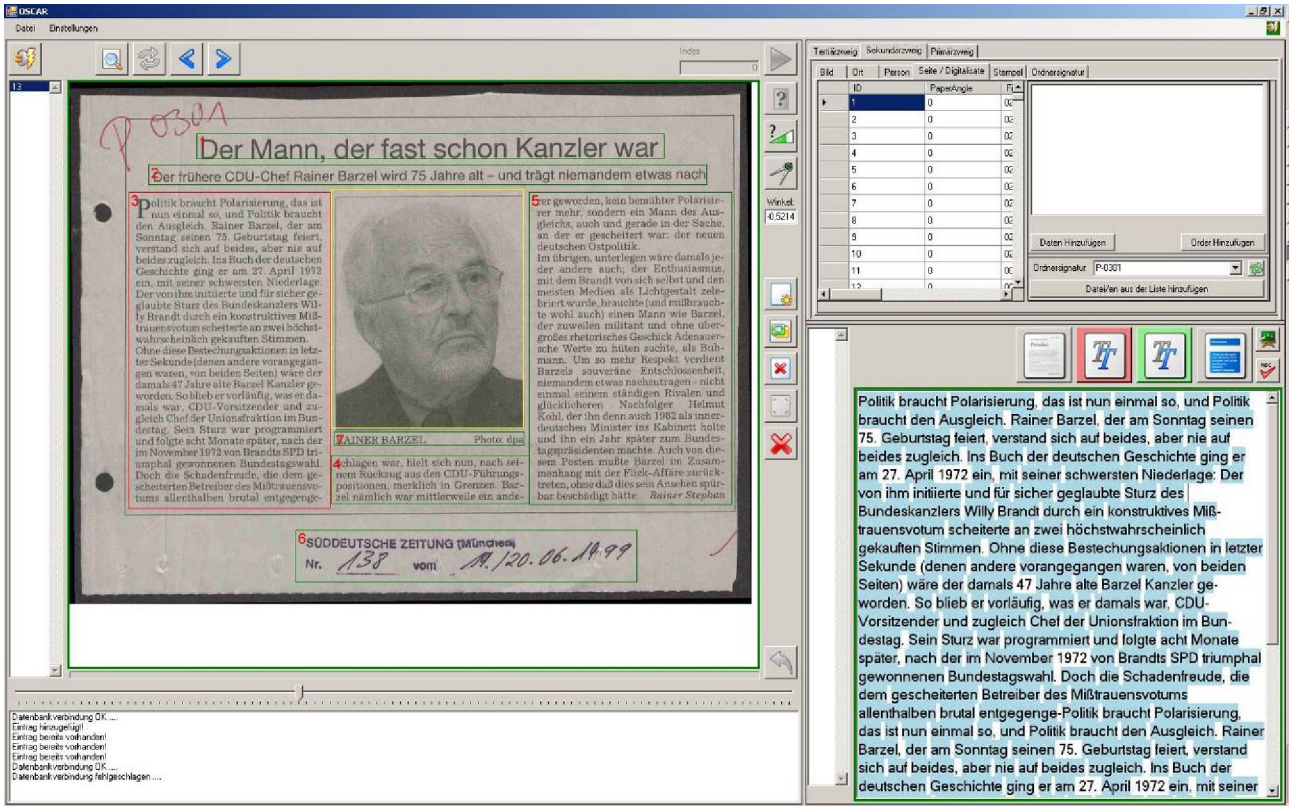


Ergebnis Texterkennung durch FineReader nach Bildvorverarbeitung

Imagebearbeitungsalgorithmen wurden von der GFal auch für andere Fälle entwickelt, so zum Beispiel zur Beseitigung von Grauschleiern bei mit Buchscannern aufgenommenen Kupferstichen.

5. Layoutanalyse, Separierung von interessierenden Regionen

Es wurde eine Oberfläche entwickelt, die die Möglichkeit einer interaktiven Unterstützung bei der Bereichsfindung und –klassifikation unterstützt. So können in dem Image der Zeitungsseite die interessierenden Bereiche gekennzeichnet und für die Erkennung in die gewünschte Reihenfolge gebracht werden Abbildungen und Unterschriften können extrahiert sowie Stempel- und Handschriftregionen markiert werden, solange noch keine Programme existieren, die diese Differenzierung selbstständig vornehmen.



Für den Anwender ist weiterhin wichtig, ob der entsprechende Artikel ohne Verletzung von Eigentumsrechten vollständig ins Netz gestellt werden kann oder nicht. Diese Entscheidung muss - in Abhängigkeit von der Art der Information - von dem Archiv gefällt werden. Die Oberfläche ermöglicht auch die Ergänzung gewünschter Angaben und – nach Durchlauf durch die OCR – eine Korrektur der Ergebnisse.
Die Resultate werden in einer MySQL-Datenbank abgelegt.

6. Methoden zur Korrektur der gelieferten Ergebnisse

Zunächst bietet das System die Möglichkeit, bekannte Wörter in einem Wörterbuch abzulegen. In der Anzeige des von der OCR erkannten Textes werden dann alle Wörter, die nicht im Wörterbuch abgelegt sind, blau untersetzt, so dass diese „verdächtigen" Wörter durch den Nutzer schnell erfasst und – falls erforderlich und gewünscht – korrigiert werden können. Blau untersetzte, aber orthographisch richtige Wörter können zum vorhandenen Wörterbuch hinzugefügt werden.

Eine weitere Methode, den „rohen" OCR-Text zu korrigieren, besteht in der Anwendung einer unscharfen Suche. Ausgehend von einem Wörterbuch, sucht man im OCR-Text nach Wörtern, die diesen möglichst ähnlich sind. Diese Ähnlichkeit lässt sich beispielsweise mit der „Levenshtein distance" messen. Dabei berechnet man, wie „teuer" die Umwandlung eines Strings in einen anderen ist. Kosten entstehen durch das Löschen eines Zeichens (1), durch das Einfügen eines Zeichens (1), durch das Ersetzen eines Zeichens (1) und durch das Vertauschen zweier Zeichen (2). Die Umwandlung von „Blld" in „Bild" kostet also eine Ersetzung (1). Je nachdem, wie groß man die Toleranz, bzw. die Entfernung, in der man noch sucht, setzt, steigt auch die Rechenzeit. In diesem Beispiel hätte eine Entfernung von Eins gereicht. Längere Zeichenketten können aber potentiell mehr Fehler enthalten und benötigen eine größere Toleranz. Es ist eventuell empfehlenswert, die Toleranz an die Wortlänge anzupassen.

Bei einer unscharfen Suche versucht man, die Anzahl der nötigen Vergleiche auf ein Minimum zu reduzieren, um Rechenzeit zu sparen. Ein Verfahren arbeitet mit Hashwerten, die sowohl eine scharfe als auch eine unscharfe Suche ermöglichen (Rönnblom 2006). Um diese Hashwerte zu berechnen, wird zunächst eine Wörterbuchanalyse vorgenommen, die für jeden Buchstaben berechnet, mit welcher Wahrscheinlichkeit er einmal, zweimal, dreimal usw. in einem Wort vorkommt.

	1	2	3	4	5
a	54%	15%	2%	0%	0%
b	26%	3%	0%	0%	0%
c	30%	4%	0%	0%	0%
d	24%	2%	0%	0%	0%
e	85%	52%	20%	5%	1%
f	23%	3%	0%	0%	0%
g	40%	8%	1%	0%	0%
h	42%	8%	1%	0%	0%
i	54%	16%	3%	1%	0%
j	1%	0%	0%	0%	0%
k	25%	3%	0%	0%	0%
l	43%	10%	1%	0%	0%
m	23%	4%	0%	0%	0%
n	67%	25%	6%	1%	0%
o	27%	5%	1%	0%	0%
p	17%	2%	0%	0%	0%
q	1%	0%	0%	0%	0%
r	64%	21%	4%	0%	0%
s	58%	20%	5%	1%	0%
t	58%	19%	4%	0%	0%
u	41%	8%	1%	0%	0%
v	9%	0%	0%	0%	0%
w	12%	0%	0%	0%	0%
x	1%	0%	0%	0%	0%
y	1%	0%	0%	0%	0%
z	15%	1%	0%	0%	0%
ä, ö, ü, ß	0%	0%	0%	0%	0%

Buchstabenhäufigkeit in der deutschen Sprache

Diese Informationen fließen in den entwickelten Algorithmus ein. Weitere Bewertungskriterien für die Ähnlichkeit von Wörtern sind u.a.
Anzahl der übereinstimmenden Buchstaben
Anzahl der übereinstimmenden Position/Reihenfolge der Buchstaben
Wortlängendifferenz.
Die Erfolgsaussichten, durch das Matching das richtige Wort zu finden, hängen dann natürlich vom Umfang des verwendeten Wörterbuchs und der (einstellbaren) geforderten Übereinstimmung ab.

Aufgrund der guten Erkennungsrate der OCR bei dem bisherigen Material gehen wir jedoch zur Zeit davon aus, dass eine vollständige Korrektur des Artikelinhalts nicht notwendig ist. Für das Auffinden von Artikeln bei Eingabe von Schlagwörtern genügt es vermutlich, die einzelnen Artikel mit den darin vorhandenen Substantiven zu indexieren und für den Abgleich mit dem eingegebenem Schlagwort bereits existierende Verfahren einzusetzen (z.B. Soundslike von MySQL).

7. Ausblick

Einige der gewünschten zu extrahierenden Informationen sollen in der Folgezeit noch automatisiert werden. Dazu gehören:
Stempelseparierung und –erkennung
Separierung von Überschriftsregionen
Separierung von Abbildungen und Skizzen

Folgende weitere Probleme scheinen lösbar, die einer „intelligenten Erschließung“ nahe kommen:
Extraktion des Artikelautors
Extraktion des Abbildungsautors(Fotografen)
Klassifikation des Artikels (z.B. Roman, Interview, Todesanzeige)
Unterstützung bei der Anordnung der Textteilregionen in der richtigen Reihenfolge

Die Untersuchungen sind Ergebnisse von Forschungsprojekten, gefördert durch das BMWi im Rahmen des INNOWATT- bzw, INNO-KOM-Programms