Probleme und Lösungsansätze bei der Dokumentenerfassung

Problems and Methods of Resolution for Document Recognition

Dr. Wolfgang Schade, Karola Witschurke, Karsten Neß, Mark Alinski
Gesellschaft zur Förderung angewandter Informatik e.V. (GFaI)
Rudower Chaussee 30, 12489 Berlin
Tel.: +49 30 6392 1600 Fax: +49 30 6392 1602
E-mail: schade@gfai.de Internet: www.gfai.de

Die Entwicklungen wurden durch das BMWi gefördert.

Zusammenfassung:
In der GFaI werden seit etlichen Jahren Anstrengungen unternommen, um die Erkennung von historischen Dokumenten, speziell Schreibmaschinendokumenten, zu verbessern. Ausgehend von unzureichenden Ergebnissen kommerzieller Schrifterkennungssysteme für derartige Dokumente wurden neue Verfahren entwickelt :

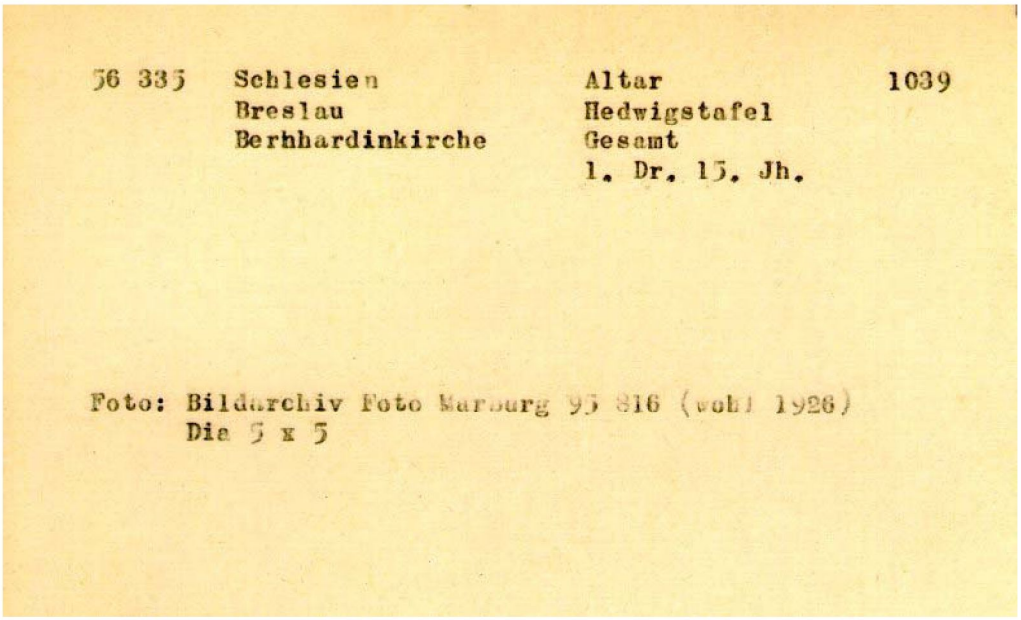
1. zur Beseitigung von Verunreinigungen auf den Dokumenten und Möglichkeiten zur Verbesserung der Typenschrift
2. zur Einbringung zusätzlicher Informationen in den Erkennungsprozess durch Einführung von Templates für Dokumentenklassen sowie
3. Verfahren zum Auffinden spezieller Bereiche zur Separierung der zu erkennenden Informationen auf strukturierten und unstrukturierten Karteikarten

Abstract
This paper describes the progress in recognition of historical documents, especially typewritten ones. In order to improve insufficient results from commercial OCR software new algorithms have been developed for:

1. removal of noise and levelling of typewritten characters
2. layout analysis by defining templates for different classes of documents
3. detection of regions of interest on structured and non-structured file cards in order to separate the information to be retrieved

1. Ausgangslage

Kommerzielle Schrifterkennungssysteme (OCR-Optical Character Recognition) beherrschen das Auslesen der Information aus neuzeitlichen Druckerzeugnissen (ab Schriftgröße 6 pt). Eine Erkennungsrate von mindestens 98% ist nichts Ungewöhnliches. Damit verbunden ist die Fähigkeit, in dem Dokumentenimage bei entsprechender Scanqualität (etwa 300 dpi) Bilder, Bildunterschriften, Artikelüberschriften, Zeitungsspalten und Tabellen zu erkennen und in einem virtuellen Dokument 1:1 abzubilden. Ausgangspunkt sind dabei bitonale Images. Als Beispiel derartiger OCR-Systeme seien Finereader von ABBYY, DokuStar/RekoStar von OCE/ODT oder KADMOS mit der Erkennungsmaschine von re-Recognition genannt. Erkannt werden auch handschriftliche Eintragungen, soweit es sich um Ziffern beziehungsweise getrennte Schriftzeichen (z.B. Blockschrift) handelt, wenn auch nicht mit dieser hohen Erkennungsrate (Beispiel: Banküberweisungen, Postsysteme). Schreibschrift oder gar Sütterlin werden nicht erkannt.
Auch mit Schreibmaschine erstellte Dokumente bereiten den OCR-Systemen Schwierigkeiten. Der Grund dafür sind undeutliche Buchstaben. Die Ursachen sind vielfältig, z.B. verschmutzte Typen oder ein ungleichmäßiger Anschlag beim Ausstellen des Dokuments.




```
j6 333 Schlesie'i Altar 1039 Breslau
Hedwigstafel Berhhardinkirche
fitesamt
l. Dr. 13, Jh.
Foto; ßiiü.^rcl;iv i'üt.»- '>ui.,i.r^
/.» .ii; •, > uE. J ••,•:-.:.•,. Die /•
x 5
```

Abb.1: Beispiel: „Erkennung“ mit Finereader, einem durchaus anerkannt guten OCR-System

Eine weitere Schwierigkeit entsteht, wenn nicht die Originalseiten, sondern nur die Durchschläge zur Verfügung stehen:

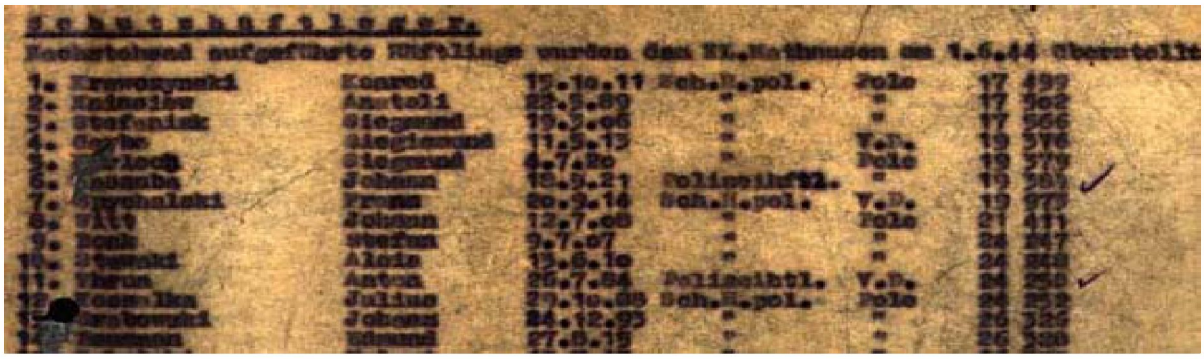


Abb.2: Ausschnitt aus einer Transportliste der Gedenkstätte Stutthof/Polen

Ein weiteres Problemfeld, dem wir uns in den letzten Jahren zuwandten, war das Aufsuchen interessierender Regionen auf Dokumenten. Hierbei ging es um Karteikarten, deren Aufbau zwar „im Prinzip“ festgelegt ist, die sich jedoch im konkreten Fall sehr unterscheiden:

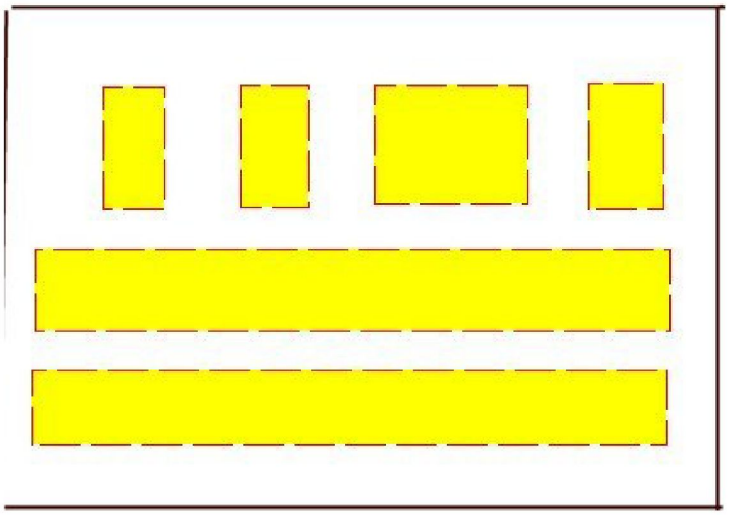
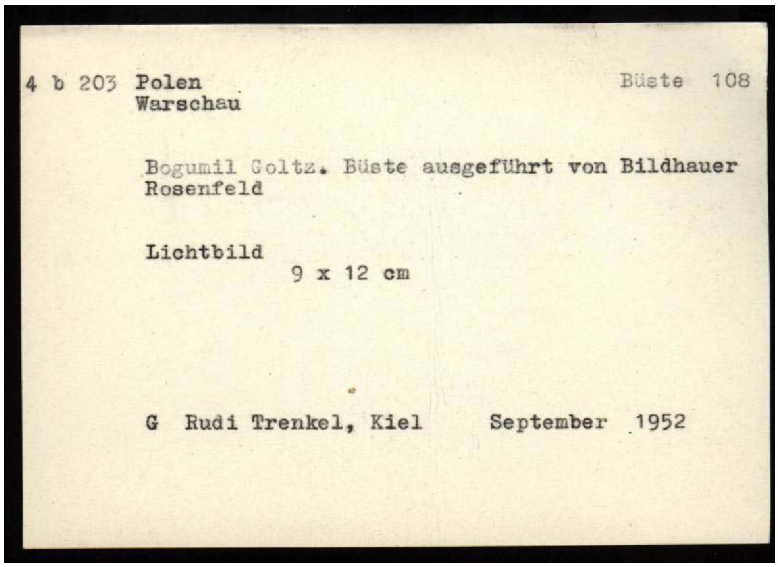


Abb.3: prinzipieller Aufbau der Karteikarte



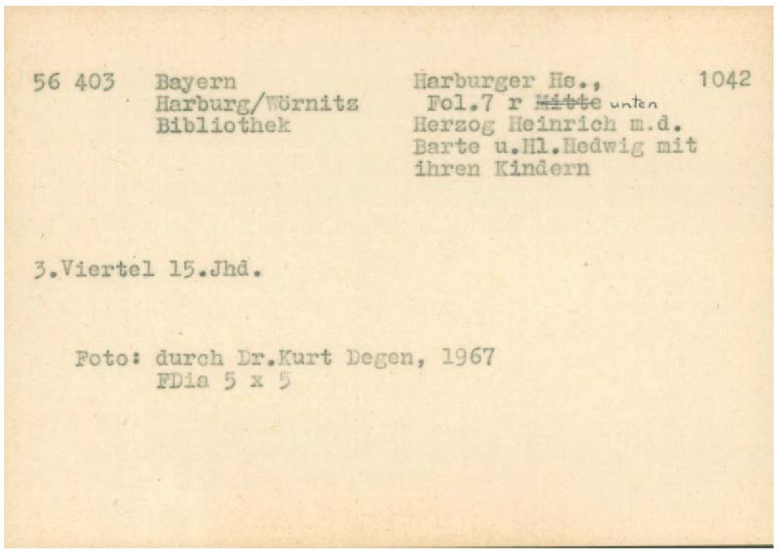


Abb.4: Unterschiedliche konkrete Umsetzung

2. Verfahren zur Beseitigung von Störungen und Verbesserung der Lesbarkeit

2.1 Farbraumquantisierung

Die vorliegenden Dokumente sind dreikanalige Farbbilder mit 24 Bit Farbtiefe. Da eine Reduktion auf ca. 128 Farben (7 Bit Farbtiefe) für die nachfolgenden Auswertungen optimal ist, wurden drei verschiedene Farbreduktionsverfahren untersucht. Als Ergebnis wurde festgestellt, dass die Farbreduktion anhand folgender Formel erfolgen kann:

$$GW = \frac{R - 32 + 0.0635 * G + 0.0625 * B}{2}$$

Diese Variante der Farbreduktion kann sehr schnell berechnet werden, da die aufwendige Berechnung von Clustern entfällt

2.2 Analysemethoden zur Trennung von Vorder- und Hintergrund

Die Analyse der fraktalen Eigenschaften bot nur unzureichende Ansätze, die Trennung von Vorder- und Hintergrund zu verbessern.
Wesentlich bessere Ergebnisse wurden mit der Ortsfrequenzanalyse erzielt. Drei Operatoren aus der Ortsfrequenzanalyse wurden genutzt, um einerseits die charakteristischen Eigenschaften der mit Schreibmaschine geschriebenen Buchstaben und Zahlen hervorzuheben und andererseits die Textur des Hintergrundes zu unterdrücken:

1. **Reduce-Operator**: Hohe Frequenzen werden aus dem Bild entfernt. Die verbleibenden Informationen können in einem Bild mit halber Breite und Höhe und somit ¼ der Pixel dargestellt werden. Diese Bildverkleinerung führt zu einem deutlichen Gewinn an Performance für folgende Berechnungen.
2. **Expand-Operator**: Es wird die ursprüngliche Bildgröße wieder hergestellt.
3. **HDC-Operator**: Auf der Basis lokaler Pixelfrequenzen werden Texturen unterdrückt und vom Vordergrund getrennt.

Diese drei Operatoren wurden zur Verbesserung der Trennung von Vorder- und Hintergrund zu einem effektiven Algorithmus zusammengesetzt.

2.3 Statistische Charakterisierung von Vorder- und Hintergrund

Ausgangspunkt für statistische Untersuchungen bildet die Überlegung, dass Vordergrundpixel (Schreibmaschinenanschläge) wesentlich seltener auftreten als Hintergrundpixel. Außerdem treten Vordergrundpixel nicht einzeln auf, sondern immer zusammengefasst in lokalen Häufungen. Ein geeignetes Verfahren, um diese Eigenschaften hervorzuheben, ist die Analyse der Struktur-Entropie. Mit geringer Wahrscheinlichkeit (< 10%) im Bild vorhandende Farbwerte, die zudem mit lokalen Häufungen auftreten, haben eine hohe Entropie. Dem Hintergrund zuzuordnende Farbwerte haben eine geringe Entropie. Damit können die Bereiche, in denen Buchstaben oder Zahlen zu finden sind, mit hoher Genauigkeit extrahiert werden. Bild 5 zeigt das Ergebnis der Entropie-Analyse mit nachfolgendem Tresholding und morphologischer Nachbearbeitung. Der Schwellwert für das Thresholding wird adaptiv anhand des Histogramms berechnet, als morphologischer Operator kommt Dilate mit einem 3x3 Pixel großem Kern zum Einsatz.

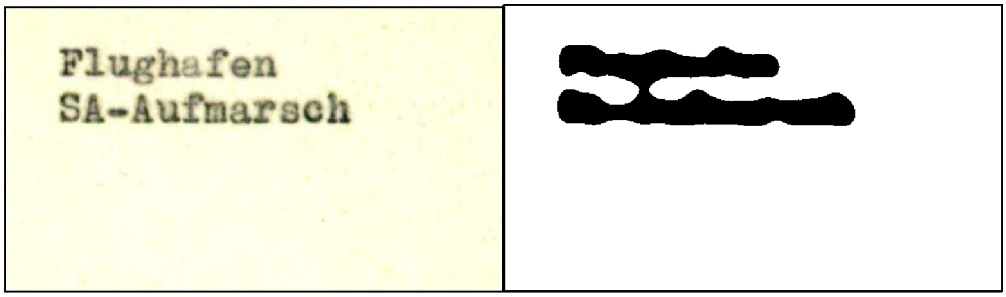


Abb.5: Ergebnis der Entropie-Analyse

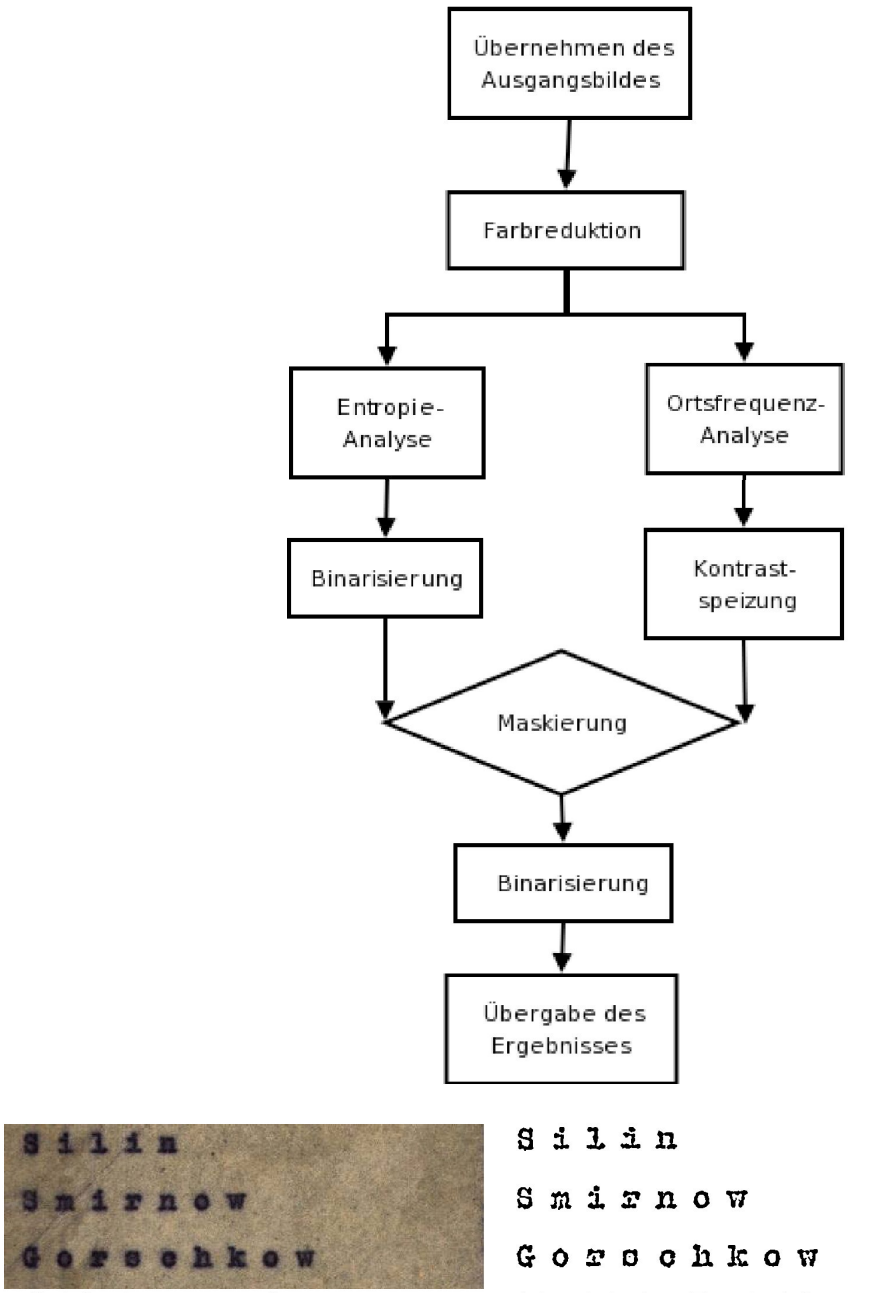
Das Beispiel zeigt, dass mit Hilfe des Struktur-Entropie-Algorithmus eine effektive Erkennung des Hintergrundes möglich ist und Bereiche der Schreibmaschinenanschläge mit hoher Genauigkeit erkannt werden können. Das Ergebnis wird zur Maskierung des Hintergrundes genutzt.

2.4 Binarisierung

Das Ergebnis der Ortsfrequenz-Analyse wird zur Verbesserung der Trennung von Schreibmaschinenanschlägen und Hintergrund mittels Histogramm-Equalize-Operator nachbearbeitet und mit dem Ergebnisbild der Entropie-Analyse maskiert. Nicht maskierte Bereiche werden anschließend mit einem lokal-adaptiven Schwellwert binarisiert.

2.5 Kombination der Verfahren

Durch Kombination der verschiedenen Methoden konnte eine deutliche Verbesserung der Lesbarkeit der Images erreicht werden:



Original und binarisiertes Ergebnisimage

Abb.6: Verfahrenskombination und Ergebnis

3. Templates zur Beschreibung der Dokumentenklasse

Durch die Einführung eines Templates zur Beschreibung einer Dokumentenklasse können der Layoutanalyse zusätzliche Hilfsmittel, insbesondere zur Strukturerkennung, übermittelt werden. Definiert werden hierbei u.a. die „Trenner“-Art (line / no line), die Rechteckbereiche, in denen der Trenner liegen kann, die „Trenner“-Breite und die „Trenner“-Länge.

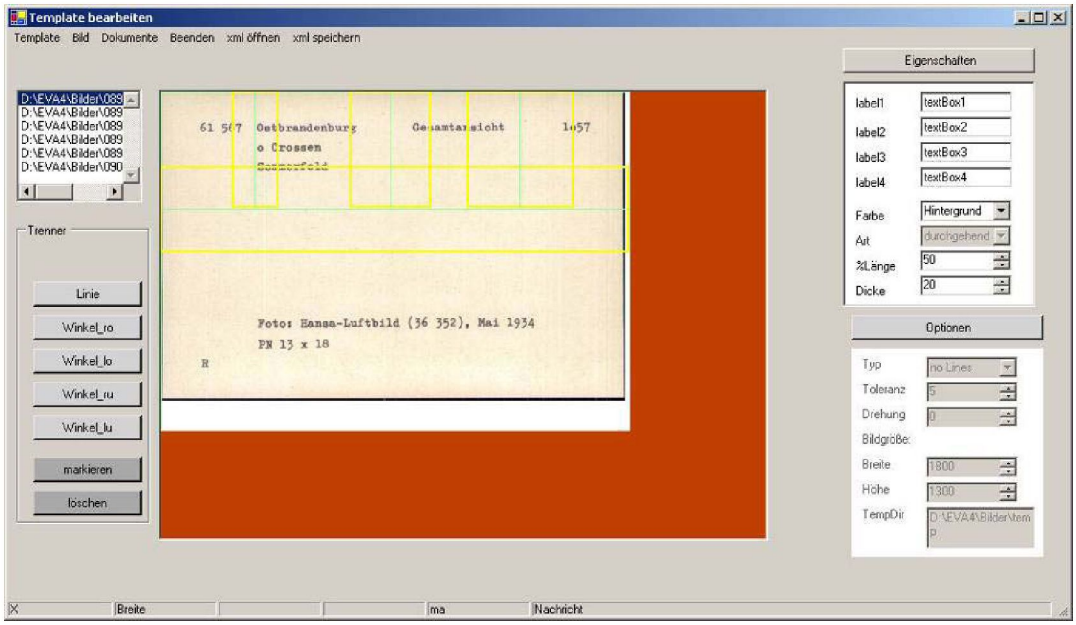
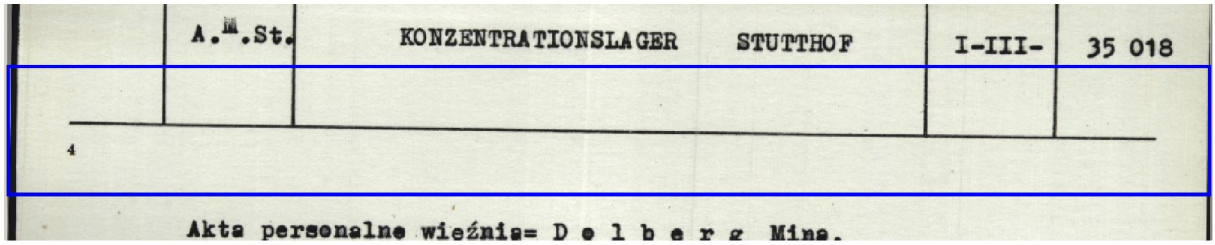


Abb.7: Oberfläche zur Templatebeschreibung

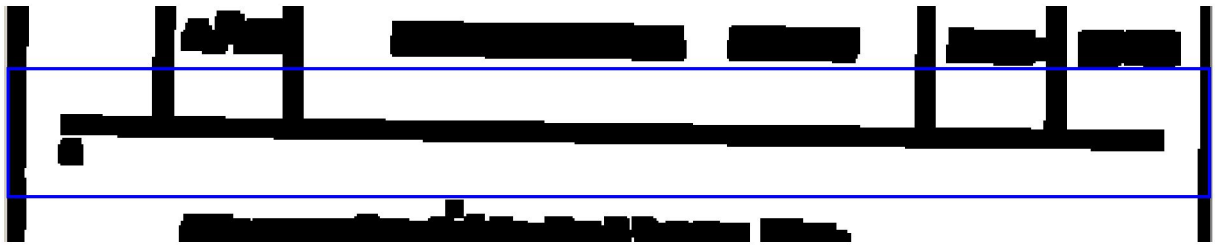
4. Strukturanalyse

Über den graphischen Editor zur Templatebeschreibung werden nur die Trennerbereiche eingegeben. Durch die Eingabereihenfolge wird intern ein Binärbaum erzeugt, der die Bipartitionierung des Dokuments widerspiegelt. Image und entstehende Teilbilder werden durch Auffinden der realen Trennerpositionen durch die Abarbeitung des Binärbaumes sukzessiv bipartitioniert. Somit wird nur die Kenntnis über die **relative** Position der Bereiche zueinander ausgenutzt.
Das folgende Beispiel soll die Erkennung einer waagerechten Trennlinie im Detail zeigen (Karteikarte mit aufgedruckter Strukturierung). Die Erkennung einer senkrechten Linie verläuft nach einem ähnlichen Schema.

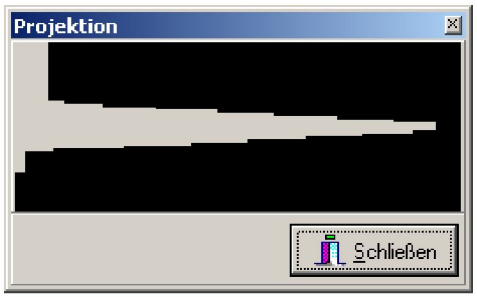
1. Zuerst werden der Suchbereich und die Informationen über die zu findende Trennlinie aus der Templatedatei ausgelesen. Das folgende Bild zeigt den Suchbereich und die darin enthaltene Trennlinie.



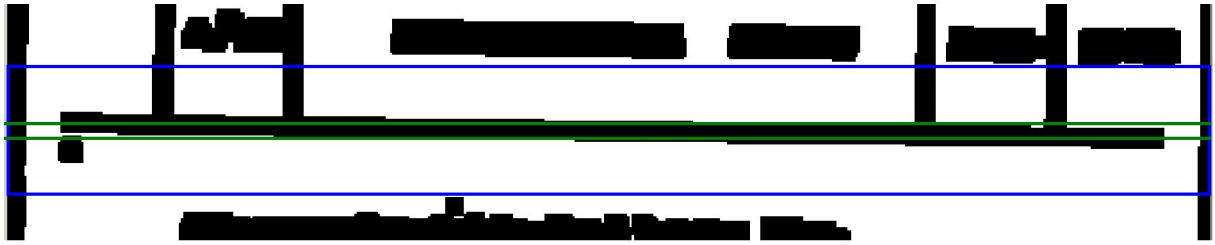
2. Auf dem Bild ist deutlich zu erkennen, dass die Trennlinie leicht verdreht auf der Vorlage liegt (Verrutschen der Karte beim Scanvorgang). Da das Finden der Trennlinie eine waagerechte Linie voraussetzt, erfolgt im nächsten Schritt eine Korrektur der Vorlage. Hierbei werden die dunklen Bildanteile um einen festen Wert vergrößert.



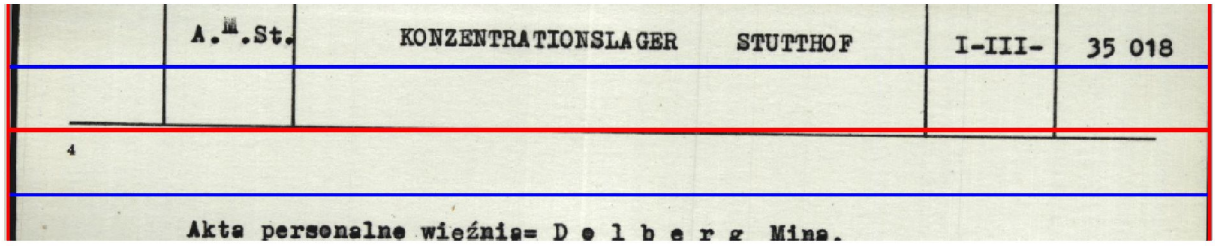
3. Die Suche nach der Trennlinie erfolgt mit Hilfe einer waagerechten Projektion. Für eine waagerechte Projektion werden für jede Bildzeile die vorkommenden gefärbten Bildpunkte aufaddiert. Für ein 100 Zeilen hohes Bild erhält man also 100 einzelne Projektionswerte. Wenn man diese graphisch darstellt, entsteht folgendes Bild:



4. Nun wird für jede Zeile (von oben beginnend) geprüft, ob die Zeile genügend gefärbte Bildpunkte für eine Trennlinie besitzt. Die ungefähre Länge der Trennlinie ist durch ein Attribut in der Templatedatei bekannt. Sobald eine Zeile gefunden wurde, die genügend gefärbte Bildpunkte besitzt, stoppt die Suche, und der Index der gefundenen Zeile wird abgespeichert. Nun startet eine zweite Suche, ausgehend vom gefundenen Index. Diesmal wird solange gesucht, bis eine Zeile gefunden wird, die nicht mehr ausreichend gefärbte Bildpunkte besitzt. Auch diesmal wird der gefundene Index abgespeichert. In der folgenden Abbildung sind die beiden gefundenen Zeilen durch eine Linie gekennzeichnet.



5. Zwischen den beiden gefundenen Zeilen wird der geometrische Schwerpunkt bestimmt. Dieser wird dann als gefundene Trennposition zurückgegeben. Die letzte Abbildung zeigt die markierte Trennposition.



Durch diese Verfahren der Bildvorverarbeitung und der Unterstützung der Layoutanalyse konnte die Erkennungsrate der eingesetzten OCR in unserem Fall von 60% auf 90% erhöht werden, wodurch die (notwendige) Nachbearbeitung erheblich gesenkt wurde.