RecType – ein System zur Erkennung von Schreibmaschinendokumenten

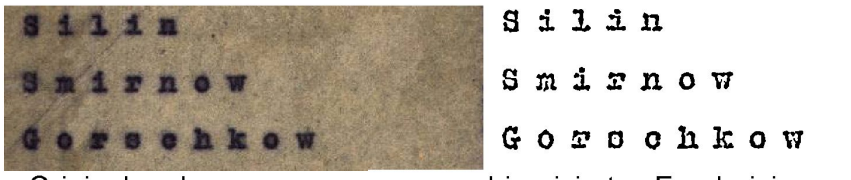
Gesellschaft zur Förderung angewandter Informatik e.V. (GFaI)
Rudower Chaussee 30, 12489 Berlin
Kontakt: Dr. Wolfgang Schade,
Tel./Fax :+49 30 6392 1605/02
e-mail: schade@gfai.de

In allen Archiven lagern Dokumente mit Informationen, um zum Beispiel für die Dokumentation der Stadt- und Landesgeschichte oder für historische Forschungen der interessierten Öffentlichkeit zur Verfügung gestellt zu werden. Um diese Dokumente einem größeren Personenkreis zugänglich zu machen oder den Zugang zu erleichtern, ist neben der Digitalisierung auch eine Inhaltserfassung, zumindest eine Indexierung, von Vorteil.

Am Stand wird ein System vorgestellt, mit dem die Inhaltserfassung und Indexierung von Dokumenten, insbesondere auch von mit Schreibmaschine erstellten, erfolgen kann.

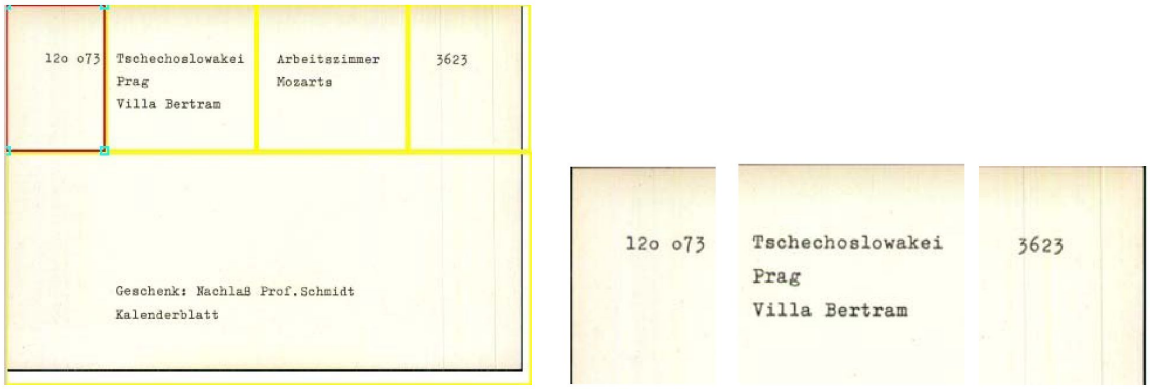
Generell liefern handelsübliche OCR-Systeme sehr gute Ergebnisse bei der Retrokonversion von auf weißem Papier gedrucktem Text. Schwierigkeiten entstehen beim Auftreten von gestörtem Hintergrund (z.B. bei älteren Dokumenten) und bei Schreibmaschinendokumenten auch bei verschmutzten Typen, ungleichmäßigem Anschlag oder bei Verarbeitung eines Durchschlags. Speziell bei der Erfassung von Karteikarten tritt ebenfalls das Problem einer richtigen Strukturanalyse auf. Diese Probleme werden mit unserem System in der Vorverarbeitung angegangen, so dass die nachfolgende OCR wesentlich bessere Ergebnisse liefert.

Für die Beseitigung von Störungen und zur Verbesserung der Lesbarkeit wird in **RecType** eine Kombination verschiedener Verfahren (Ortsfrequenzanalyse, Analyse der Strukturentropie mit nachfolgendem Tresholding und Histogramm-Equalizing) eingesetzt.



Original und binarisiertes Ergebnisimage

Für die Unterstützung der Strukturanalyse einer Dokumentenklasse wird in **RecType** über einen graphischen Editor ein Klassentemplate beschrieben. Insbesondere werden damit der Verlauf und die ungefähre Lage vorhandener Trennbereiche (waagerechte und senkrechte Linien, Trennzeilen bzw. –spalten) beschrieben, die es bei der Analyse eines aktuellen Images der Klasse erlauben, die interessierenden Bereiche zu separieren und die darin stehenden Informationen nach Anwendung der OCR den entsprechenden Feldern in der Datenbank zuzuordnen.

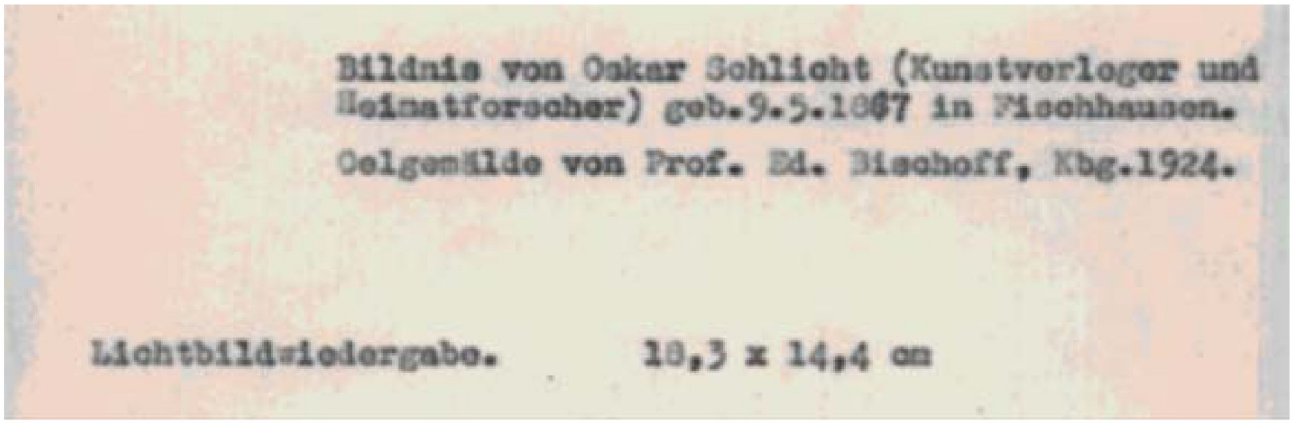


Zur weiteren Verbesserung der Worterkennung wird der OCR eine unscharfe Suche nachgeschaltet.

Texterkennungsprogramme bieten einen Wörterbuchabgleich an, um die Erkennung zu verbessern. Bislang scheitern sie aber noch an Wörtern mit hoher Fehlerrate. Ziel des eingesetzten Verfahrens ist es, diese fehlerhaften Wörter durch eine unscharfe Suche im Wörterbuch nachträglich zu korrigieren.

Bei einem Anteil falsch erkannter Buchstaben zwischen 30-50% wächst die Anzahl der potentiellen Treffer, je nach Größe des Wörterbuchs, sehr schnell an. Mit der hier eingesetzten Methode gelingt es, mit Hilfe von Hashwerten die Menge der Kandidaten mit wenigen Vergleichen einzugrenzen.

Um das ähnlichste Wort zu finden, berücksichtigt **RecType** außerdem von der OCR häufig verwechselte Buchstaben wie „e“, „c“ und „o“ oder „i“ und „l“. Beste Ergebnisse für hohe Fehlerraten werden erzielt, wenn sich der Wortschatz einschränken lässt. Wörter mit großer Wortlänge können bei gleicher oder sogar höherer Fehlerrate eindeutiger erkannt werden als kurze Wörter.



Ohne Nachkorrektur:	**Nachkorrektur:**
Bildnis von 0~ar 30h1ioht (Kunotvorlocor und Üe10atforochor) cob.9.5.1007 in 2icchh~oon Oolge~lldo von I>rof. rA. 310050ff, Kbg.1924. Llchtblldslc.lorcabc. 1O,3 x 14,4 Ctl	Bildnis von Oskar Schlicht (Kunstverleger und Heimatforscher) cob.9.5.1007 in Fischhausen. Oelgemälde von Prof. rA. 310050ff, Kbg.1924. Lichtbildwiedergabe. 1O,3 x 14,4 Ctl