RecType - ein System zur Erkennung von Schreibmaschinendokumenten

RecType – a system for recognition of typewritten documents

Dr. Wolfgang Schade
Gesellschaft zur Förderung angewandter Informatik e. V.
Rudower Chaussee 30, 12489 Berlin
Tel.: +49 (0) 30 -6392 1605, Fax: +49 (0) 30 -6392 1601
E-mail: schade@gfai.de, Internet: www.gfai.de

Zusammenfassung:
Für gedruckte, neuzeitliche Dokumente liefern kommerzielle OCR-Systeme eindrucksvolle Resultate (98% Erkennungsrate). Diese Systeme versagen jedoch bereits für einen Großteil von Schreibmaschinendokumenten (60% Erkennungsrate): Niedrige Papierqualität, geringer Zeichenkontrast, Ungenauigkeiten beim Setzen der Schriftzeichen beeinträchtigen die automatische Erkennung des Textes.
In nationalen Forschungsprojekten* wurden deshalb neue Algorithmen und Verfahren für Bildvorverarbeitung (Hintergrund-Säuberung für schlecht konditionierte Dokumente), für die Heraustrennung von Teilregionen aus dem Dokument und für die Verbesserung der OCR-Erkennung entwickelt.
In RecType wurden diese Algorithmen mit denen eines kommerziellen OCR-Systems kombiniert, wodurch eine Erkennungsrate von 90% erreicht wird.

Abstract:
For printed, modern documents commercial OCR (Optical character recognition)-systems provide striking results (98% recognition rate). But these systems fail even for a majority of typewriter documents (60% recognition rate). Low paper quality, low print contrast, inaccuracies when setting the characters impair the automatic recognition of the text.
Therefore, during national research projects* were developed new algorithms and procedures for image pre-processing (background cleaning for badly conditioned documents), separation of interesting regions, and improvement of OCR-recognition
In RecType, these algorithms are combined with those of a commercial OCR-system to increase the recognition rate up to 90%

RecType verwendet als Ausgangsmaterial unkomprimierte Farb-Images (TIFF).
In einem ersten Schritt der Bildvorverarbeitung erfolgen eine Trennung von Vorder- und Hintergrund und eine „Reinigung" des Hintergrundes. Hierfür werden Verfahren der Tiefpassfilterung, Ortsfrequenz- und Entropie-Analyse kombiniert:
Anhand eines Beispieles sollen die Ergebnisse der einzelnen Zwischenschritte dargestellt werden. Das Beispiel soll die Möglichkeiten und auch die Grenzen des entwickelten Algorithmus verdeutlichen. Bisher war es nicht möglich, in diesem Dokument einzelne Buchstaben zu erkennen.

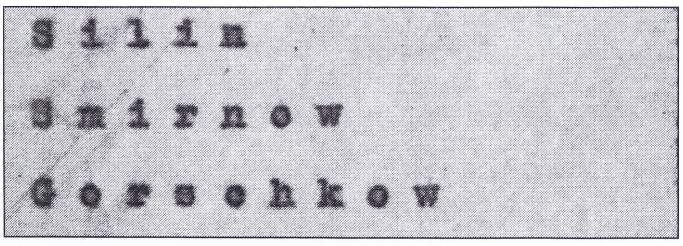


Abb: 1: Ausschnitt aus einer Namensliste

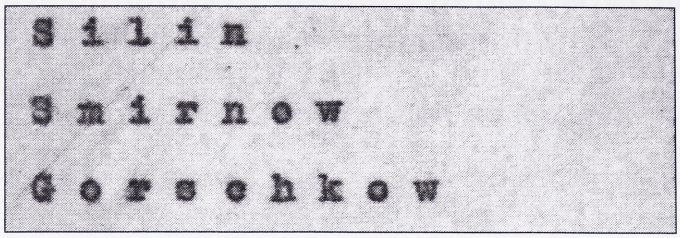


Abb: 2.: Ergebnis der Farbreduktion

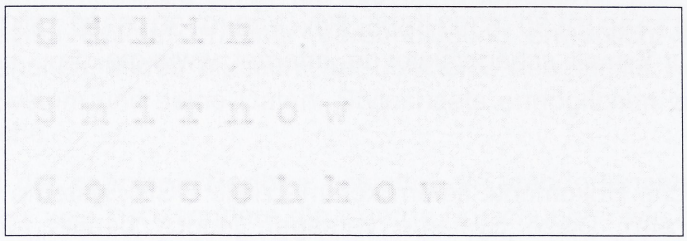


Abb. 3: Ergebnis der Ortsfrequenz-Analyse

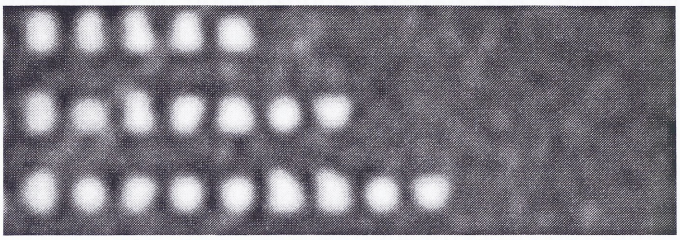


Abb. 4: Ergebnis der Entropie-Analyse

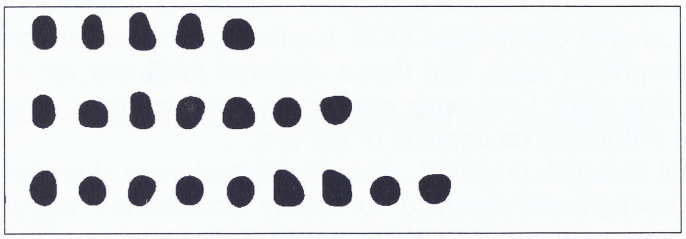


Abb. 5: Binär-Maske aus der Entropie-Analyse

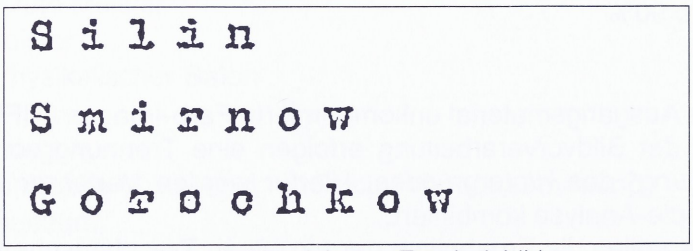


Abb. 6: Binarisiertes Ergebnisbild

Anhand dieses Ergebnisbildes werden von der OCR folgende Zeichen erkannt:

S i 1 1 n
S m i r n o w
G ö c o c h k o w

Es werden 16 von 21 Zeichen korrekt erkannt. Dies ist eine deutliche Verbesserung.

Damit bestimmte Regionen des Dokuments automatisch erkannt und die darin enthaltenen Informationen in die richtigen Felder einer Datenbank abgespeichert werden können (insbesondere bei Karteikarten), wird mittels eines Editors eine Beschreibung der Dokumentenklasse (Klassentemplate) erzeugt. Anhand dieser Beschreibung werden die einzelnen Regionen, die im allgemeinen auf den Einzeldokumenten nicht vollständig übereinstimmen, dann auf jedem einzelnen Dokument bestimmt und an das kommerzielle OCR-System übergeben. Beschreibungselemente des Templates sind insbesondere die „Regionentrenner": waagerechte und senkrechte Linien (Hintergrundaufdrucke) oder Leerzeilen bzw. –spalten. Durch die bei der Templatebeschreibung erfolgte Reihenfolge der Eingabe wird die Suche nach den „Trennern" im Image erleichtert.
Eine weitere Ergänzung ist die Einführung zweier Prüfalgorithmen auf das Ergebnis der OCR.

Der erste umgesetzte Prüfalgorithmus basiert auf der Voraussetzung, dass die zu erkennenden Dokumente Nichtproportionalschriften enthalten. Nichtproportionalschriften sind Schriften, bei dem jedes Zeichen eine – in der Breite konstante – Zelle belegt. Dabei werden in der Datenstruktur des OCR Erkennungsergebnisses zwei aufeinander folgende segmentierte Bildbereiche gesucht, die jeweils in der Größe von einem Zeichen der Nichtproportionalschrift abweichen. Bei der Suche werden diejenigen Zeichen ausgewählt, die zu schmal für ein Zeichen sind und deren Abstand zueinander zu gering ist.

Bei gefundenen Zeichen wird überprüft, welche Zeichen die OCR erkannt hat. Sofern die erkannte Zeichenfolge einer der in der folgenden Tabelle aufgeführten Ausgangszeichenfolgen entspricht, wird eine Ersetzung durchgeführt. Die Zeichenfolge „rn" wird zum Beispiel dann durch ein „m" ersetzt.

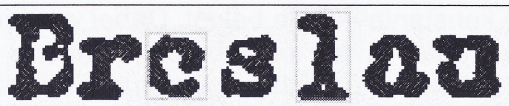
Ausgangszeichen	Ersetzt durch:
()	0
ii	u
rn	m
ni	m
cl	d
li	h
oi	a
vv	w
IC	K
VV	W
NI	M
LI	U
I)	D
I(	K

Der zweite Prüfalgorithmus vergleicht den Bildausschnitt, den die OCR als Zeichen erkannt hat, mit bekannten Referenzbildern. Hierbei werden die Kenntnisse über den benutzten Zeichenfont ausgenutzt. Zum Vergleichen der Bilder wird die Korrelation der Bilder genutzt.

In der Datenstruktur der OCR werden für jedes erkannte Zeichen bis zu 8 mögliche Klassifizierungsergebnisse zurückgeliefert. Außerdem bestimmt die OCR für jedes mögliche Klassifizierungsergebnis einen Faktor, der beschreibt, wie wahrscheinlich die Korrektheit des Klassifikationsergebnisses ist. Die 8 möglichen Klassifizierungsergebnisse sind nach der Wahrscheinlichkeit sortiert.

Zur Überprüfung werden nun die zu den 8 Ergebnissen gehörenden Referenzbilder jeweils mit dem zu erkennenden Bildausschnitt verglichen. Die dazu errechneten Korrelationswerte werden sortiert. Stammt der kleinste Korrelationswert von einem anderen Zeichen als von dem, das die OCR als wahrscheinlichstes Ergebnis erkannt hat, werden die Zeichen im OCR Ergebnis getauscht.

Die Bilder in der folgenden Tabelle zeigen einige Ergebnisse dieses Matchingalgorithmus. In der Tabelle sind jeweils das Ausgangs-OCR-Ergebnis und das Ergebnis nach Prüfung mit dem Matchingmodul dargestellt.

OCR-Ergebnis vor Prüfung	**Bild**	**OCR-Ergebnis nach Prüfung**
Brcslau		Breslau
Schlosien		Schlesien
V~mählun~		Vnmählung

Das System wurde anhand von 5000 Karteikarten des Herder-Instituts Marburg getestet.
Die Erkennungsrate konnte dabei von 60% auf 90 % gesteigert werden.
Damit sank die Nachbearbeitungszeit für diese Dokumente auf ein Viertel.

---

* Die hier vorgestellten Ergebnisse sind Resultate der Forschungsprojekte DOVER und EvA4, finanziert mit Mitteln des BMWi (INNO-WATT und ProInno2).