Verbesserte Schrifterkennung von maschinengeschriebenen Archivdokumenten mittlerer und niederer Qualität auf der Basis anwendungsspezifischer Tools Verfahren, Ergebnisse, Vergleich

Improved character recognition of typed documents from middling and lower quality based on application depending tools Processes, results, comparison

Wolfgang Schade, Karola Witschurke, Cornelia Rataj
Gesellschaft zur Förderung angewandter Informatik e.V.
Rudower Chaussee 30, D-12489 Berlin
Tel./Fax: 030 6392 1605/02; email: schade@gfai.de

Zusammenfassung:
Das EU-Projekt MEMORIAL, an dem sich neben der GFaI auch die Technische Universität Gdansk, die Universität Liverpool, die Firma ID-K (Israel) sowie die beiden Museen „Gedenkstätte Stutthof" (Polen) und „Moreshet" (Israel) beteiligen, hat als wesentlichstes Ziel die verbesserte automatisierte Erfassung personengebundener Daten aus Schreibmaschinendokumenten und Karteikarten. In diesem Beitrag werden Methoden und Verfahrensschritte vorgestellt, die zu einer Steigerung der Erkennungsrate durch OCR-Systeme führen.

Abstract:
MEMORIAL is a project of the IST 5-Programme of the European Commission. Partners of MEMORIAL are Technical University of Gdansk, University of Liverpool, an Israelian company (ID-K), two archives (Memorial place Stutthof/Poland and Moreshet/Israel), and GFaI. The most important objective of MEMORIAL is to improve recognition rate of OCR systems for digitization of typewritten documents and file cards. The paper offers the developed methods and process steps to reach this.

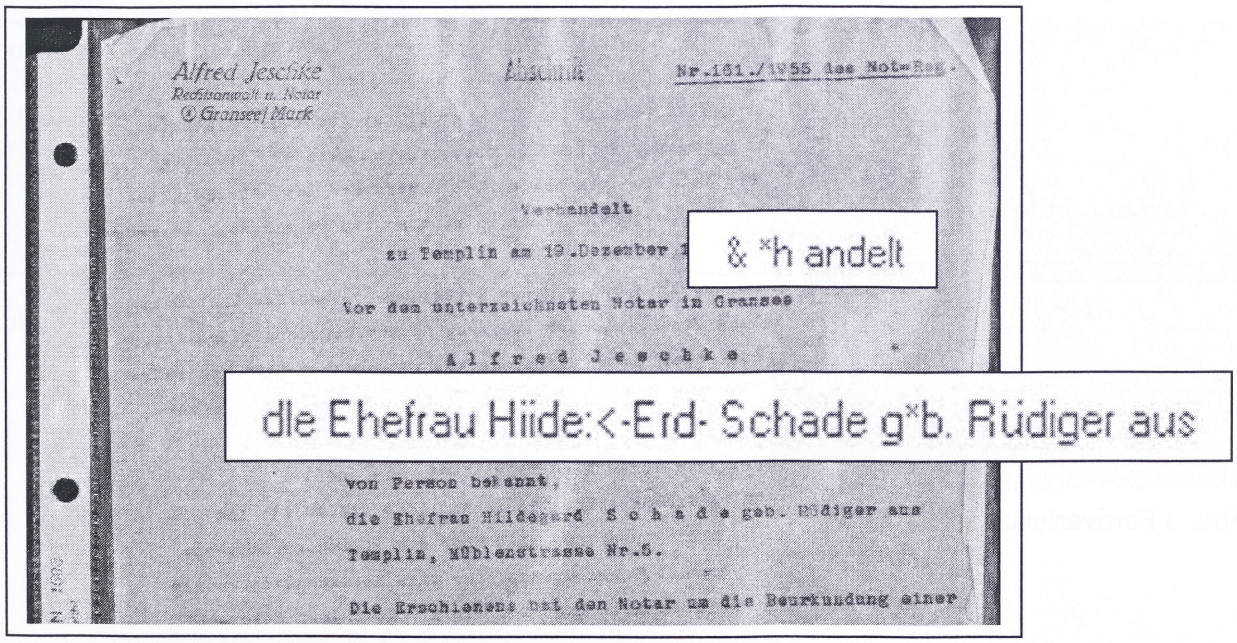
Einleitung

In allen Archiven lagern Dokumente mit Informationen, die auf Grund Ihres Inhalts für die Nachwelt erhalten werden sollen, um zum Beispiel für die Dokumentation der Stadt- oder Landesgeschichte oder für historische Forschungen der interessierten Öffentlichkeit zur Verfügung gestellt zu werden. Bedingt durch die inzwischen eingetretenen Fortschritt in der Entwicklung der Computer- und Archivierungstechnik einerseits und dem zunehmenden Verfall alter Originaldokumente andererseits ist eine starker **Trend zur Digitalisierung vorhandener archivischer Dokumente** zu beobachten, wie zahlreiche nationale und internationale Tagungen beweisen. Damit wird erreicht, dass die Dokumente nicht mehr direkt ausgehändigt werden müssen bzw. die interessierten Nutzer nicht mehr ins Archiv reisen müssen, um z.B. den entsprechenden Mikrofilm auswerten zu können, sondern dazu (wenigstens perspektivisch) den PC am eigenen Arbeitsplatz und das Internet benutzen können, um die entsprechenden Informationen zu erhalten.
Damit verbunden sind auch Bestrebungen, den **Inhalt** von Papierdokumenten (wenigstens teilweise) **zu erfassen**. Gegenwärtige Forschungsprojekte in der EU beschäftigen sich zum Beispiel mit der Inhaltserfassung von Artikeln moderner Zeitungen. Sie können sich hierbei auf leistungsfähige Schrifterkennungssysteme(Optical Character Recognition—[OCR]-Systeme) stützen, die unter gewissen Voraussetzungen (reiner Untergrund, akzeptable Schriftgröße) mit gedruckten Schriftstücken gute Ergebnisse vorweisen.
Schwieriger ist es schon mit älteren Druckwerken, da das Erkennen von Frakturschrift bislang nur unzureichend in die Erkennungsmaschinen integriert ist, jedoch laufen hier Forschungsprojekte (z.B. META-E der EU).

Ebenso schwierig ist aber die Erkennung von Schreibmaschinenschrift für die OCR wegen des ungleich kontrastierten Schriftbildes, Hintergrundverunreinigungen des Papiers oder störende Linien z.B. auf strukturierte Papier-Karteikarten). Damit ist die Digitalisierung des Inhalts zur Zeit noch mit großem Personaleinsatz verbunden und hängt ab von den Bedürfnissen der Öffentlichkeit, die sich durch Anfragen an die Archive bzw. aus den in den Archiven selbst vorhandenen Arbeitsaufgaben ergeben.
Im Projekt MEMORIAL wird der Ansatz verfolgt, die Verarbeitungsschritte, die **vor** der eigentlichen Anwendung der OCR liegen, unter Verwendung neuer und heute möglicher Technologien so zu erweitern, dass eine Verbesserung der OCR-Resultate erzielt wird.

Erarbeitete Teilschritte
Generell allen kommerziellen Erkennungsmaschinen gemein ist, dass sie auf Binärbildern (bestenfalls Graustufen-Bildern) operieren und eine Bildqualität von ca. 300 dpi erwarten. Derartige images ließen sich lange Zeit – insbesondere bei einer großen Anzahl der zu behandelnden Dokumente – durch den Scanvorgang nur binär sinnvoll erfassen (Speicherkapazität der Rechner, Scan-Kosten). Darüber hinaus liegt das Hauptanwendungsgebiet (Verarbeiten von Büropost: Rechnungen, Lieferscheine, Korrespondenz, Erfassung von Buch- und Zeitungsinhalten) wegen des weißen Papieruntergrundes natürlich im Bereich von Binärbildern, so dass eine Farberfassung der Dokumente keinen Nutzen erbringen würde.
Bei den Dokumenten, für die in MEMORIAL Verfahren entwickelt werden, sind die Voraussetzungen andere.
Zum einen ist der Dokumentenuntergrund zumeist farbig. Bei einem gleichmäßigen Farbton des Untergrundes würden auch die heute gängigen OCR-Systeme durch Binarisierung erfolgreich arbeiten können, diese Voraussetzung ist jedoch nicht gegeben. Die Karteikarten haben unterschiedlichste und nicht immer gleichartige Färbung, sind am Rande vergilbt oder gar verschmutzt, und auch das früher genutzte Schreibmaschinen - Durchschlagpapier ist nicht weiß, wodurch die Ergebnisse der OCR geschmälert werden.
Zum anderen bieten die inzwischen existierenden Geräte heute die Möglichkeit, Farb-Images kostengünstig, auch mit transportabler Technik, herzustellen und abzuspeichern, wie im Vortrag „Scantechnologien und ihre Grenzen bei der Erfassung unterschiedlicher Archivdokumente“ von A. Geschke dargelegt. Damit hat man die Chance, die Farbinformationen des Dokuments für die Aufbereitung des Bildes mit hinzuzuziehen. Folgende Teilprozesse wurden erarbeitet:

1. Als ersten Schritt gilt es dabei, für die Dokumente ***geeignete Scanmethoden*** auszuwählen. Sind Dokumente z.B. in Folien eingeschlossen, so treten Spiegelungen auf, die für die OCR nicht zu behandeln sind (Abb.1)



2. Als nächstes wird versucht, den **Dokumentenhintergrund** zu **bereinigen**. Dazu nutzt man die auch die Beschreibung der Dokumentenklasse (Vortrag „Definition gleichartiger Dokumententypen zur Verbesserung der Erkennbarkeit und ihre XML-Beschreibung"). Verschiedene Spezifika werden durch adaptive Bildverarbeitungstechniken identifiziert und markiert, wie zum Beispiel Dokumentenhintergrund; Flächen, in denen das Papier rekonstruiert wurde, oder Außenkanten, die durch den Scan-Vorgang entstanden sind (Abb.2)

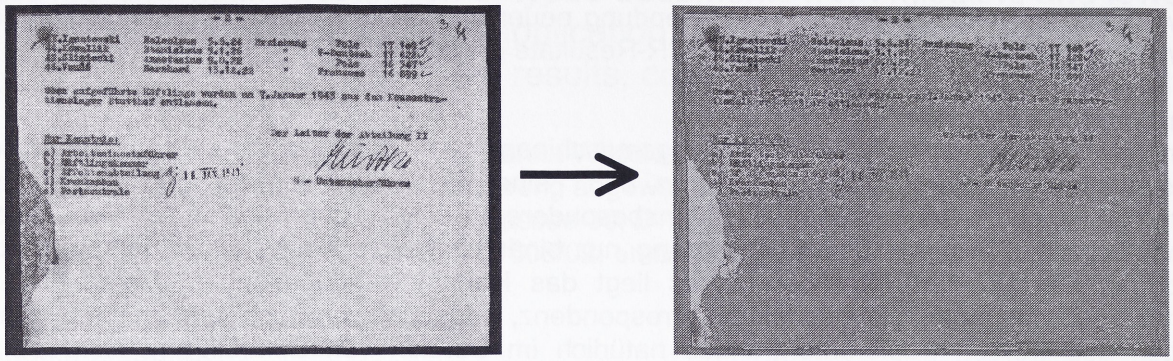
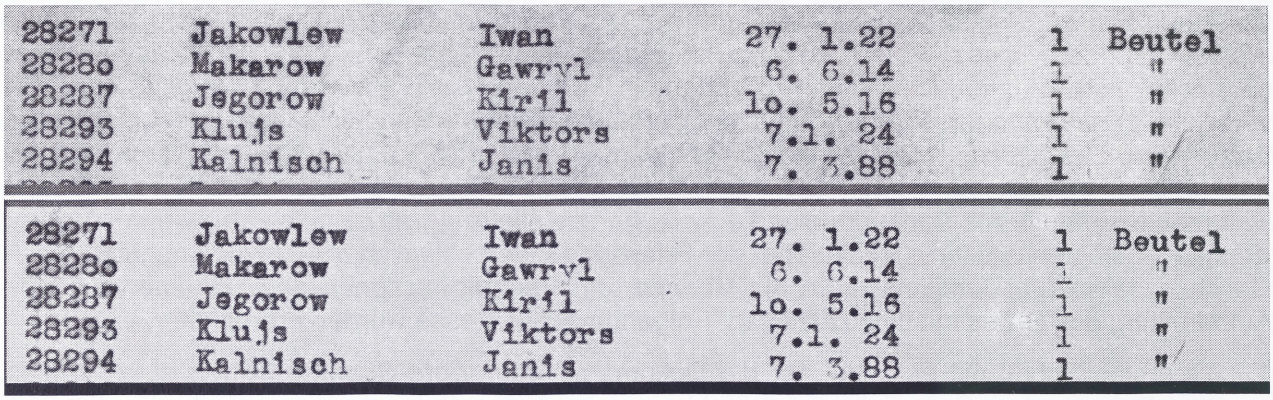


Abb. 2a Markierung von Hintergrundspezifika



28271	Jakowlew	Iwan	27. 1.22	1	Beutel
28280	Makarow	Gawryl	6. 6.14	1	"
28287	Jegorow	Kiril	10. 5.16	1	"
28293	Klujs	Viktors	7.1. 24	1	"
28294	Kalnisch	Janis	7. 3.88	1	"

Abb. 2 b Hintergrundbereinigung

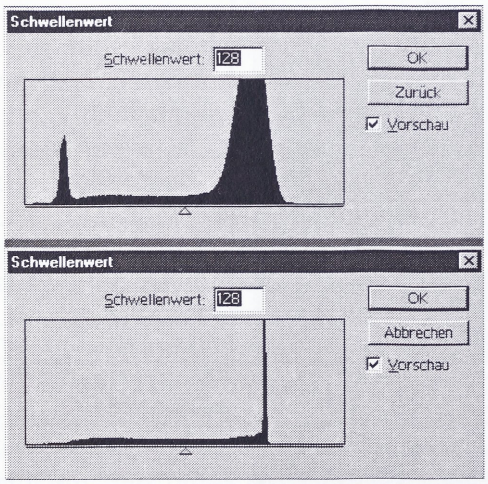
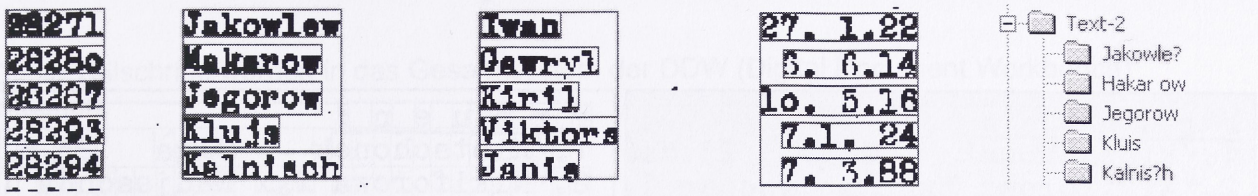


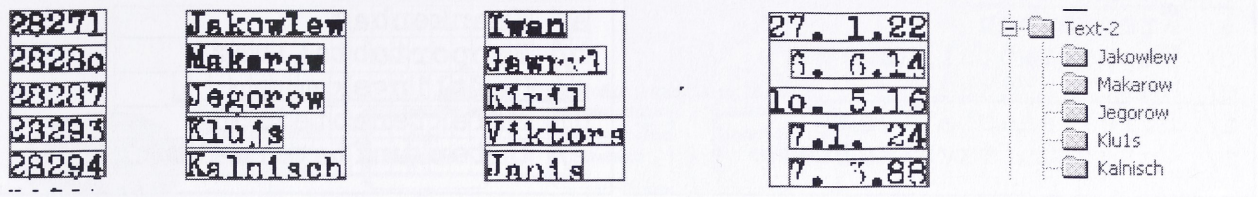
Abb. 3 Farbverteilung

Anhand der Charakteristiken der Farbverteilung im Original und im bearbeiteten Bild (s. Abb. 3) ist zu erkennen, dass für das bearbeitete Bild eine exaktere Bestimmung des Schwellenwertes zur Binarisierung erfolgen kann.

Daraus resultiert eine erste Verbesserung der OCR-Ergebnisse:
Die folgenden Abbildungen zeigen die Resultate der OCR (rechts) ohne:



und mit Hintergrundbereinigung:



Die Hauptprobleme, welche die OCR an einer guten Erkennung hindern, sind:

- Typen, die mit unterschiedlicher Stärke auf der Maschine angeschlagen wurden
- Sich berührende oder überlagernde Schriftzeichen

Zur Bewältigung dieser Schwierigkeiten werden folgende Schritte unternommen:

3. Es werden die **Text-Linien** in den Text-Feldern gewonnen. Als Hilfsmittel zur Identifizierung der Textfelder dient dazu die allgemeine **Beschreibung der Dokumentenklasse in den Templates.**

4. Innerhalb der so gefundenen Textlinien werden die **Schriftzeichen lokalisiert**. Gleichzeitig werden sich berührende **Schriftzeichen** voneinander **getrennt**. Schräg verlaufende Textlinien (z.B. bei nachträglich vorgenommenen Eintragungen in Formularen) werden ausgerichtet.

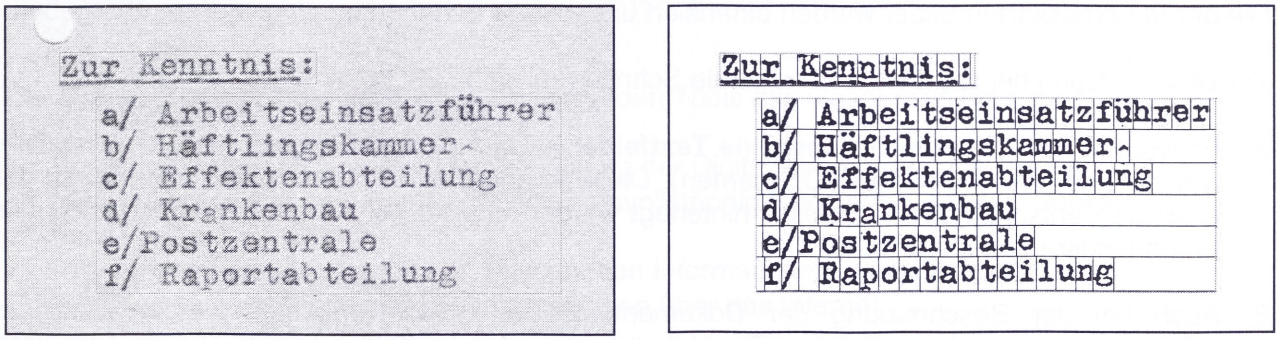
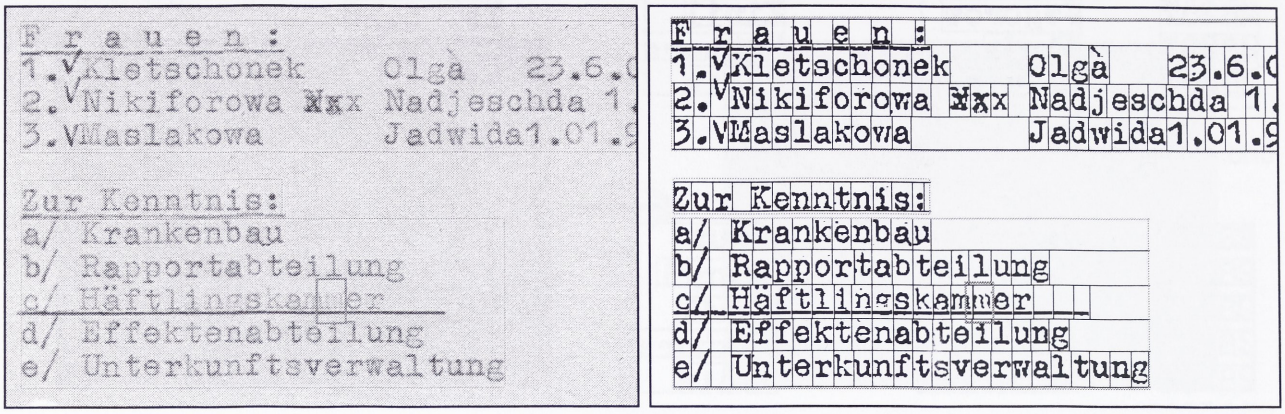
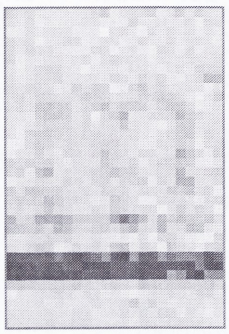
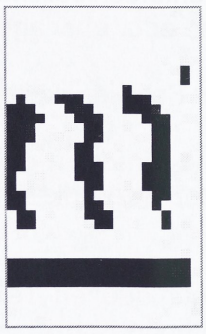


Abb. 5: Erkennung der Schriftzeichenboxen

5. Für jedes Zeichen innerhalb einer Box wird **individuell** eine **Kontrast-Anpassung** ermittelt. Dadurch können (insbesondere bei bekanntem Schriftsatz) „verlorene“ Zeichen wiederhergestellt werden (Abb.6).







Die derart verbesserten Bilder werden binarisiert und an das Schrifterkennungssystem übergeben.

Zur Unterstützung der OCR wurden folgende Schritte realisiert:

5. Es wurden **Bibliotheken für spezielle Textfelder** entwickelt, die für unsere Anwendungsfälle typisch sind (z.B. Vornamen; Ortsnamen). Diese können zum Erkennungsprozess durch die OCR den entsprechenden Feldern hinterlegt werden und so zur weiteren Verbesserung der Erkennungsrate beitragen.

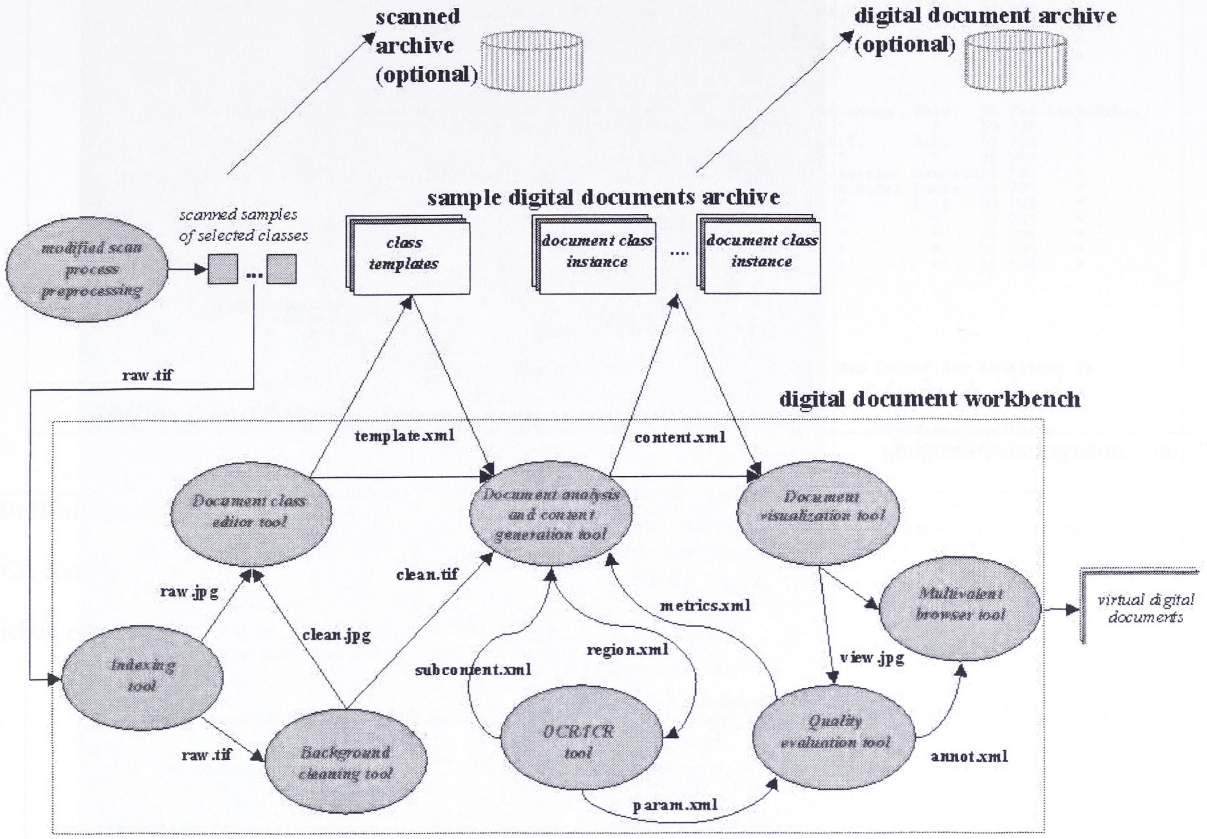
6. Auch bei der Beschreibung der Dokumentenklasse durch Templates werden derartige Informationen über den möglichen Text-Inhalt bestimmter Textfelder erfasst. **Die im Template aufgenommenen Informationen werden an das Erkennungssystem weitergeleitet** und können, soweit es das Erkennungssystem zulässt, zur Verbesserung der Erkennungsrate genutzt werden.

   Zu den weiterhin zur Verfügung gestellten Informationen gehören:
   - Schriftstärke
   - Zeichensatz (Sprache)

Vonseiten der Layoutanalyse stehen noch zur Verfügung:

- Position des Beginns einer Textzeile
- Drehwinkel von Textfeldern zur Normalen

Diese Teilschritte wurden in das Gesamtsystem der DDW (Digital Document Workbench) integriert.



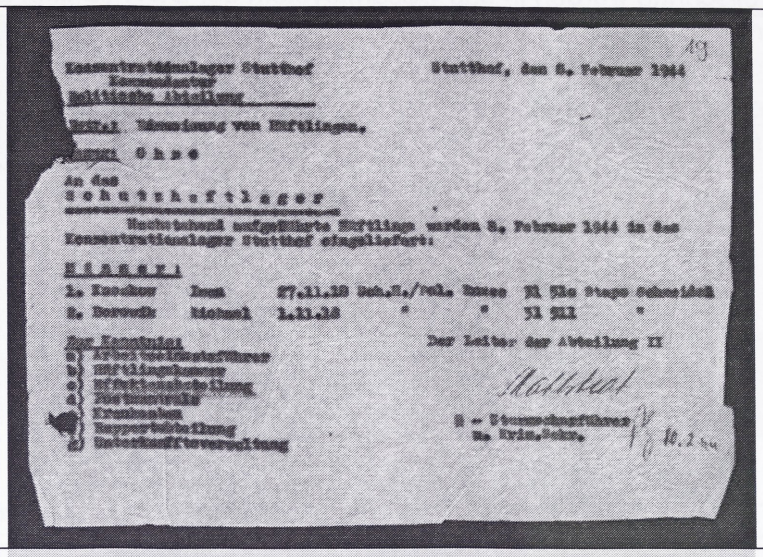
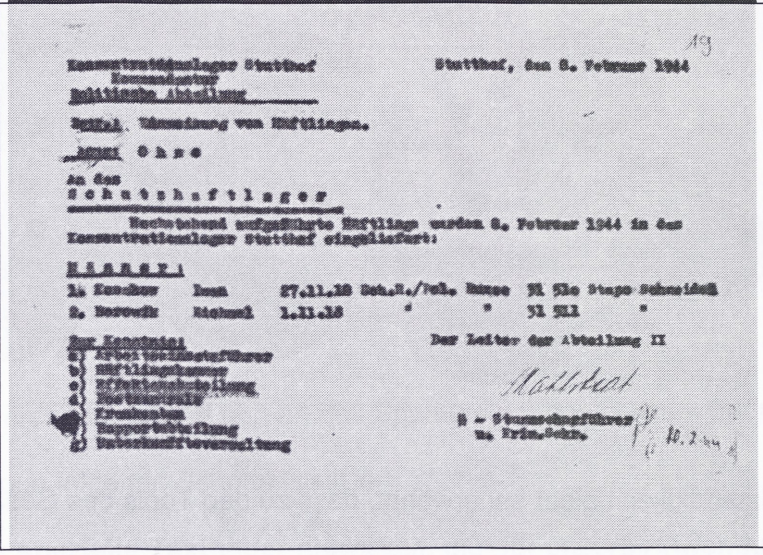
Der Vollständigkeit halber sei erwähnt, dass zu den Tools des Gesamtsystems ebenfalls

- Programme und Verfahren zur Unterstützung der Digitalisierung der Dokumente
- zum Auffinden von Dokumentenbildern in unvollständig indexierten, aber geordneten Imagefiles
- Programme zum Übertragen der gewonnenen Informationen in Archivdatenbanken
- Programme zum Austausch von Informationen über das Internet

gehören.

Ergebnisse

Erste Ergebnisse zeigen eine Verbesserung der Erkennung. Das nachfolgende Beispiel dokumentiert die schrittweise erzielte Verbesserung selbst anhand eines bereits gut konditionierten Dokuments. Die richtig erkannten Zeichen stiegen hier von 94,8 % auf 97 %.

JPG ohne Hintergrundbereinigung	
JPG mit Hintergrundbereinigung	

Ohne Hintergrundbereinigung

OCR confidence = 73%

Richtig erkannte Zeichen 94,8%

Konzentrationslager Stutthof
Kommandantur
POLITISCHE ABTEILUNG

Stutthof, den 19.2.1943

Betr.: Einweisung von Häftlingen
Bezug: O H N E

An das
Schutzhaftlager

Nachstehend aufgeführte Häftlinge wurden am 16.Februar 1943 in das Konzentrationslager Stutthof eingeliefert:

F r a u e n:							
1. Murawski	Martha	23.12.89	Erziehung	Polin	38	789	StapoBrbg.
2. Klosowski	Antonina	15.09.23	"	"	38	790	"
3. Poswiatowski	Olga	5.2.15	"	"	38	791	"
4. Zielinski	Stefanie	11.11.26	Sch.H.Pol.	"	38	792	"
5. Mazurowski	Czeslawa	15.7.21	"	"	38	793	"
6. Jurznski	Najeschda	28.9.11	"	Russin	38	794	"
7. Kletschonek	Olga	24.8.00	"	"	38	795	"
M ä n n e r							
8. Nawrocki	Daniel	7.7.02	Erziehung	Pole	38	796	StapoBrbg.
9. Lewandowski	Franciszek	29.9.23	"	"	38	797	"
10.Simon	Ernst	16.3.21	F.S.V.	R.D.	39	798	"
11.Lüdtke	Max	11.11.11	"	"	39	799	"
12.Schwartz	Gustav	29.1.20	V.H.Asozial	Deutsch	39	780	"
13.Walenko	Iwan	1912	Sch.H.Pol	Russe	39	781	"
14.F I D U R A	Marian	18.8.98	"	Pole	39	782	"
15.Wawrzyniak	Leon	8.8.12	"	"	39	783	"
16.Dzuban	Michael	28.9.21	"	"	39	784	"
17.Slawinski	Robert	29.11.01	"	"	39	785	"
18.K o t	J an	13.3.03	"	"	39	786	"

Zur Kenntnis:
a/ Arbeitseinsatzführer
b/ Häftlingskammer
c/ Effektenabteilung
d/ Krankenbau
e/Postzentrale
f/ Raportabteilung

Der Leiter der Abteilung II
Falsyfikat 1-1
SS-Sturmscharführer
u.Krim.Sekr.

Mit Hintergrundbereinigung

OCR confidence = 76%

Richtig erkannte Zeichen 95,4%

Konzentrationslager Stutthof
Kommandantur
POLITISCHE ABTEILUNG

Stutthof, den 19.2.1943

Betr.: Einweisung von Häftlingen
Bezug: O H N E

An das
Schutzhaftlager

Nachstehend aufgeführte Häftlinge wurden am 16.Februar 1943 in das Konzentrationslager Stutthof eingeliefert:

F r a u e n:							
1. Murawski	Martha	23.12.89	Erziehung	Polin	38	789	StapoBrbg.
2. Klosowski	Antonina	15.09.23	"	"	38	790	"
3. Poswiatowski	Olga	5.2.15	"	"	38	791	"
4. Zielinski	Stefanie	11.11.26	Sch.H.Pol.	"	38	792	"
5. Mazurowski	Czeslawa	15.7.21	"	"	38	793	"
6. Jurznski	Najeschda	28.9.11	"	Russin	38	794	"
7. Kletschonek	Olga	24.8.00	"	"	38	795	"
M ä n n e r							
8. Nawrocki	Daniel	7.7.02	Erziehung	Pole	38	796	StapoBrbg
9. Lewandowski	Franciszek	29.9.23	"	"	38	797	"
10.Simon	Ernst	16.3.21	F.S.V.	R.D.	39	798	"
11.Lüdtke	Max	11.11.11	"	"	39	799	"
12.Schwartz	Gustav	29.1.20	V.H.Asozial	Deutsch	39	780	"
13.Walenko	Iwan	1912	Sch.H.Pol	Russe	39	781	"
14.F I D U R A	Marian	18.8.98	"	Pole	39	782	"
15.Wawrzyniak	Leon	8.8.12	"	"	39	783	"
16.Dzuban	Michael	28.9.21	"	"	39	784	"
17.Slawinski	Robert	29.11.01	"	"	39	785	"
18.K o t	J an	13.3.03	"	"	39	786	"

Zur Kenntnis:
a/ Arbeitseinsatzführer
b/ Häftlingskammer
c/ Effektenabteilung
d/ Krankenbau
e/Postzentrale
f/ Raportabteilung

Der Leiter der Abteilung II
Falsyfikat 1-1
SS-Sturmscharführer
u.Krim.Sekr.

Mit Zeichenverbesserung

OCR confidence = 79%

Richtig erkannte Zeichen 97,0%

Konzentrationslager Stutthof
Kommandantur
POLITISCHE ABTEILUNG

Stutthof, den 19.2.1943

Betr.: Einweisung von Häftlingen
Bezug: O H N E

An das

Schutzhaftlager

Nachstehend aufgeführte Häftlinge wurden am 16.Februar 1943 in das Konzentrationslager Stutthof eingeliefert:

F r a u e n:							
1. Murawski	Martha	23.12.89	Erziehung	Polin	38 789	StapoBrbg	
2. Klosowski	Antonina	15.09.23	"	"	38 790	"	
3. Poswiatowski	Olga	5.2.15	"	"	38 791	"	
4. Zielinski	Stefanie	11.11.26	Sch.H.Pol.	"	38 792	"	
5. Mazurowski	Czeslawa	15.7.21	"	"	38 793	"	
6. Jurznski	Najeschda	28.9.11	"	Russin	38 794	"	
7. Kletschonek	Olga	24.8.00	"	"	38 795	"	
M ä n n e r							
8. Nawrocki	Daniel	7.7.02	Erziehung	Pole	38 796	StapoBrbg.	
9. Lewandowski	Franciszek	29.9.23	"	"	38 797	"	
10. Simon	Ernst	16.3.21	P.S.V.	R.D.	39 798	"	
11. Lüdtke	Max	11.11.11	"	"	39 799	"	
12. Schwartz	Gustav	29.1.20	V.H.Asozial	Deutsch	39 780	"	
13. Walenko	Mikolf Iwan	1912	Sch.H.Pol	Russe	39 781	"	
14. F I D U R A	Marian	18.8.98	"	Pole	39 782	"	
15. Wawrzyniak	Leon	9.8.12	"	"	39 783	"	
16. Dzuban	Michael	28.9.21	"	"	39 784	"	
17. Slawinski	Robert	29.11.01	"	"	39 785	"	
18. K o t	J an	13.3.03	"	"	39 786	"	

Zur Kenntnis:
a/ Arbeitseinsatzführer
b/ Häftlingskammer.
c/ Effektenabteilung
d/ Krankenbau
e/Postzentrale
f/ Raportabteilung

Der Leiter der Abteilung II

SS-Stubmscharführer
u.Krim.Sekr.