Scantechnologien und ihre Grenzen bei der Erfassung unterschiedlicher Archivdokumente

Scanning Technologies and Limits for Archive Document Acquisition

Alexander Geschke
Preservation Academy Leipzig
Kreuzstrasse 12
04103 Leipzig
Tel.: 0341-98 388 21, Fax: 0341-98 388 20
E-mail: ageschke@pa-leipzig.com, Internet: www.preservation-academy.com

Zusammenfassung:

Um Dokumente zu scannen stehen unterschiedlichste Geräte von der Digitalkamera bis zum Dokumentenscanner bereit. Unterschiede sowie Vor- und Nachteile werden an Beispielen besprochen. Einen weiteren Schwerpunkt bilden die Leuchtquellen und der Einfallwinkel des Lichtes, der für bestimmte Dokumententypen von entscheidender Bedeutung für das Scanergebnis sein kann. Dabei kommt es darauf an, welches Ziel mit der Digitalisierung erreicht werden soll. Wenn eine anschliessende OCR erfolgen soll, ist eine möglichst schattenlose, flach wirkende Beleuchtung sinnvoll. Im Gegensatz dazu sind bei der Reproduktion historischer Dokumente die Betonung der Papierstruktur eher wünschenswert.
Die Ergebnisse des Einsatzes von Farbfiltern zur besseren Erkennung von Buchstaben werden vorgestellt und diskutiert. Ebenso wird die Rolle der Farbwiedergabe besprochen.
Alle Ergebnisse wurden im Rahmen des EU-Projekts MEMORIAL gewonnen, dessen Ziel die Erfassung personengebundener Daten aus Schreibmaschinendokumenten und Karteikarten ist.

Abstract:

For document scanning different devices are available starting from digital cameras up to feeder scanner. The differences, advantages and disadvantages are discussed. Another focal point are the position of light sources and the angle of incidence of the light sufficient for the scanning result of some document types. The decisive question is the aim of digitisation. If OCR is necessary in postprocessing it is useful to have as possible flat illumination without shadows. On the contrary for reproduction of historic documents the emphasis of paper structure is desirable.
The results of colour filter application for better character recognition are presented and discussed. Just as the role of colour reproduction is discussed.
All results are got during the EC funded project MEMORIAL. The main goal is the acquisition of data from personal files on machine typed documents or file cards.

Einleitung

Digitalisieren hat andere Charakteristika als der Mikrofilm. Die Beständigkeit ist nur gewährleistet, wenn die Daten ständig genutzt werden, d.h. eine Migration auf neue (im Vergleich zur Lupe hochentwickelte) technische Systeme und auf neue Datenspeicher realisiert wird. Dem steht der Vorteil des breitesten und unmittelbaren Zugangs gegenüber. Aber aus internationalen und nationalen Untersuchungen ist auch bekannt, dass die relativ hohen Kosten der reinen Digitalisierung nur ca. 1/3 der Gesamtkosten eines zugriffsfähigen elektronischen Verteilungssystems darstellen. In ein Gesamtprojekt gehen ja auch Transport, Logistik, Bildbearbeitung, Indexierung, Speicherung, Datenbankanpassung, Realisierung des Internetzugriff,

Internet-Site-Wartung und Datenpflege ein. Deshalb ist es im Zentrum für Bestandserhaltung wichtig, mit dem Kunden den auf ihn zugeschnittenen Gesamtrahmen zu realisieren. Dies kann in Stufen erfolgen, wie bei der bereits erfolgten Digitalisierung von Bachautographen. In der ersten Stufe wurden die Blätter in abgestimmter Reihenfolge digitalisiert, indexiert und auf CD-ROM an den Kunden übergeben. Die zweite Stufe wird der Realisierung des Internetzugriffs gewidmet sein.

Das Projekt MEMORIAL hat an einem ausgewählten Anwendungsbereich den Gesamtprozess von der Digitalisierung bis zur Verfügbarmachung von personenbezogenen Daten aus Registern u.ä. zum Ziel. Dazu werden grundsätzliche Probleme aus der gesamten technologischen Kette der Digitalisierung von Dokumenten aus Archiven untersucht und einer Lösung zugeführt.

Dokumenteneingabe

Verschiedene Eingabegeräte

Die komplexe Aufgabe erfordert ein differenziertes Herangehen. Die Möglichkeiten und Grenzen der Eingabetechnologien werden beschrieben und der derzeitige Stand wird diskutiert. Die erste Schlussfolgerung nach Beendigung des Workpackages 2 (Dokumenteneingabe) lautet, dass keine Produktbeschreibung eigene Tests unter realen Bedingungen mit Originaldokumenten und Testmiren ersetzen kann. Da es andererseits nicht möglich war alle Geräte zu erproben, wurden einige als Repräsentanten ausgewählt. Unterstützung mit Scannern für die Tests wurde durch die Fa. MikroUniverse GmbH (Berlin), das International Tracing Centre (Bad Arolsen) and the French i2s gegeben.

Im Allgemeinen ist es kein Problem digitalisierte Abbildungen von aktuellen Bürodokumenten zu erhalten, die gut genug für eine anschliessende Anwendung von OCR-Programmen sind. Kommerzielle OCR-Programme nutzen jedoch nur Binärbilder, die üblicherweise eine Auflösung von mindestens 300dpi aufweisen. Somit ist es ausreichend für aktuelle Bürodokumente, die maschinengeschrieben oder vom PC-Dokument ausgedruckt sind, binäre Scans zu verwenden. Für ein kleines Büro ist ein einfacher Flachbettscanner ausreichend. Für Büros grosser Firmen hingegen, die viele Dokumente erzeugen, sind Massenscanner erforderlich.

Dokumenten- oder Feederscanner

In der Vergangenheit wurden neue Geräte entwickelt, die fast unseren Ansprüchen genügen. Beispiele (für verschiedene Leistungsklassen) sind:

1. Canon 5080C, Desktop Document scanner, colour resolution 200 dpi
2. KODAK i250/i260 Scanner, colour resolution 300dpi (optical), 33 p/min, (7.000/8.500€)
3. Fujitsu 4990C Document scanner, colour resolution 300-400 dpi, 80p/min (50.000 €)
4. OCE Scanstar 5045, colour resolution 400 dpi, 200 images/min (65.000 €)

Elektronische Blattausrichtung und Beseitigung von Pixelfarbfehlern (dropouts) verbessern die Ergebnisse der OCR und das parallele Scannen der Vor- und Rückseite sind eingeschlossen.

Solche Scanner können für die Verarbeitung von Karteikarten genutzt werden (beispielsweise die Karteikarten von Marburg, die von Bad Arolsen oder auch die Häftlingskartei von Stutthof). Falls nicht die Notwendigkeit besteht alle Karteikarten in sehr kurzer Zeit zu digitalisieren, so ist der Einsatz eines Dokumentenscanners der kleineren Klasse (i250 simplex oder i260 duplex) sinnvoll.

Im Gegensatz zu den erwähnten Karteikarten sind solche Scanner für die Verarbeitung der Transportlisten der KZ-Gedenkstätte Stutthof nicht geeignet. Versuche mit einem dem Durchschlagpapier ähnlichen Papier zeigten, dass Dokumentenscanner solche Seiten zerstören oder zumindest zerknittern. Abbildung 1 zeigt einen Scan von Schreibmaschinen-Durchschlagpapier. Um die Beschädigungen der Papiere durch den Einzugsmechanismus zu umgehen, versuchten wir die Papiere in Folien zu betten. In diesen Fällen wurde das Papier zwar nicht zerknittert, aber es kam zu häufigen Lichtreflexen, die das Erkennen der Schrift darunter unmöglich machten. Entsprechend versagte auch an diesen Stellen die OCR. Das Reflexbild ist in Abbildung 2 dargestellt. Unter Berücksichtigung der aktuell am Markt vertretenen Gräte ist für solche empfindlichen Dokumente nur der Einsatz von Kameras, Flachbett- oder Buchscannern möglich.

Flachbettscanner

Verschiedene high end Flachbettscanner wurden getestet. Im Ergebnis kann ich Ihnen einige Resultate des Quatographic X-Finity Pro 42 vorstellen. Wir konnten keine wesentlichen Probleme erkennen, wenn man von der Scangeschwindigkeit absieht. Allerdings wurde ein allgemeines Problem aller Bildaufnahmegeräte offensichtlich. Es handelt sich um die Beleuchtung. Da die Beleuchtung und damit der Beleuchtungswinkel in Flachbett- und Dokumentenscannern fest eingestellt sind, ist es nicht möglich, die Beleuchtungseigenschaften für eine spezielle Aufgabe zu optimieren. Die meisten Flachbettscanner benutzen eine einzelne, auf einer Seite des Signalempfängers angebrachte Leuchte, die sich mit der Zeile über die Vorlage bewegt. Dadurch wird die Struktur des Papiers ebenfalls deutlich und durch Schattenbildung noch verstärkt. Das kann für die Darstellung historischer Dokumente sehr positiv sein, nicht jedoch für Erkennungsaufgaben, wie sie für Bürodokumente bei Einsatz der OCR zur Debatte stehen. Die Abbildung 3 zeigt eine den Ausschnitt einer Ormigseite a.) mit einem Flachbettscanner und b.) mit einem Buchscanner aufgenommen. Die Teile c.) und d.) zeigen jeweils die entsprechende best mögliche Binarisierung (manuell gewählt) für die Ausgangsbilder. Was das Ergebnis für eine anschliessende OCR bedeutet, kann sich jeder selbst ausmalen.

Buch- oder Kamerascanner

Am Markt werden verschiedenste Buchscanner angeboten. Vom Abbildungsprinzip kann man drei Hauptgruppen unterscheiden: Zeilenscanner in der Bildebene (z.B. BookEye BCS-2, Minolta PS7000, Zeutschel OS7000, 9000 und 10000) oder in der Objektebene (z.B. DigiBook 6002, 10000) und Matrixarrays in der Bildebene (z.B. Zeutschel OS8000, HIT-vario). Am wichtigsten scheint das kontaktlose Abtasten zu sein, das mit der Mikroverfilmung vergleichbar ist. Ein Nachteil der Zeilenscanner im Vergleich zu Matrixscannern ist die geringere Scangeschwindigkeit auf Grund der mechanischen Bewegung (die jedoch meist immer noch über der von Flachbettscannern liegt). Der momentane Nachteil von grossen Matrixkameras liegt vor allem in den Kosten. Sie werden jedoch die Aufnahmegeräte der Zukunft darstellen. Beispielsweise benutzt der OS8000 nur eine Matrix mit 3500x 2300 Pixeln (8 MPixel). Ich konnte ausgiebige Tests mit einem zum 6002 aufgerüsteten DigiBook6000 durchführen. Die Hauptprobleme mit dem alten 6000 konnte mit besserer Steuerungssoftware, grossflächigeren Sensorelementen der Zeile, modifizierter Beleuchtung und mechanischer Stabilisierung überwunden werden. In externen Tests verglichen wir den OS7000 (Zeile mit 8bit Graustufen), und den HIT vario Matrixscanner (Farbe).

Abbildung 4 zeigt das Fragment eines A4-Dokuments. Beide Scanner liefern vergleichbare Resultate, wobei das Problem des inhomogenen Hintergrunds deutlich wird und die Farbfähigkeit des Kamerascanners eine effektivere Bildverarbeitung zur Hintergrundunterdrückung gestattet.

Die erste Schlussfolgerung von allen Vergleichen ist die Notwendigkeit einer genauen Parametereinstellung einschliesslich Blende, Scangeschwindigkeit (Belichtungszeit) und Entfernungseinstellung.

Im METAe-Projekt wurden Minolta PS7000 (Graustufen) und ImageWare BookEye A2 (nur schwarz-weiss) miteinander verglichen. Der neue BookEye ist jetzt ebenfalls mit Graustufen verfügbar. Wie im METAe-Bericht ersichtlich, konnten keine signifikanten Unterschiede in der reinen Bildqualität festgestellt werden.

Da für Hybridanwendungen (Mikrofilm UND Digitalisierung) die Diskussion um die Reihenfolge eher philosophischer Natur zu sein scheint, haben wir zusätzlich einen Prototyp eines automatischen Mikrofilmscanners getestet. Die Produktivität ist ziemlich hoch und der Prozess läuft automatisch ab. Das Hauptproblem verlagert sich (wie bei allen anderen dann auch einmal) zur Indexierung.

Digitalkameras

Man kann die für unsere Zwecke am Markt befindlichen Digitalkameras in drei Klassen unterteilen: Hochauflösende Kosumerkameras (Matrix mit üblicherweise 5-8 MPixel), hochauflösende professionelle Kameras (Wechselobjektive, Matrix >5MPixel) und professionelle Mittelformat-kameras mit scan backpack (Zeile, 6000 Pixel). Ich testete die Nikon CoolPix5000. Für die nahe Zukunft ist der breite Übergang der professionellen Kameras zu 24x36mmm CMOS Sensoren mit 10-15MPixeln zu erwarten. Dies hat geringeres Rauschen zur Folge und wird im Kostensegment unter 5000 € zu finden sein.

Für A4 ergaben sich mit der CP5000 geometrische Auflösungen um 215 dpi bei einer Bildwiederholfrequenz von 12 Sekunden (5 Bilder pro Minute). Die Kamera zeigt befriedigende Ergebnisse für Formate von A4 und kleiner bei gleichzeitig im Vergleich zu Buchscannern minimalen Kosten (unter 1500 €). Die Entwicklung zu höher auflösenden Sensoren hat sich zwar verlangsamt, ist jedoch noch im Gange. Ein Beispiel ist die Sony DSC F828 mit 8 MPixel, was der Wiedergabe eines A4-Blattes mit ca. 260 dpi entspricht.

Verbesserungen und Untersuchungen

Optische Parameter der Eingabegeräte

Zur Adaption der Scanparameter an eine spezifische Dokumentenklasse kann es hilfreich sein, Parameter de Eingabegerätes zu variieren. Die Hauptparameter sind in unserem Falle die Scharfeinstellung (für leicht unscharfen und damit homogeneren Hintergrund, was zu besseren OCR-Resultaten führt) und die Belichtungsparameter (Blende und Belichtungszeit). Die Unschärfe kann effektiver und vor allem kontrollierter durch Schritte der digitalen Bildverarbeitung erzeugt werden.

Beleuchtung

Wie schon gezeigt, hat der Beleuchtungswinkel einen starken Einfluss auf die Unterstreichung oder Unterdrückung von Eigenheiten des Hintergrunds (der Papierstruktur, siehe Abb. 5). Um diesen Einfluss praktisch zu untersuchen, wurde das Beleuchtungssystem des Buchscanners von i2s verändert. In Abbildung 6 ist die veränderte Anordnung der Beleuchtung sichtbar. Im Gegensatz zur Veränderung der optischen Parameter ist die Egalisierung des Hintergrunds durch

veränderte Beleuchtung viel effektiver als in nachfolgenden Schritten der Bildbearbeitung. Der Grund liegt darin, dass die Hintergrundbearbeitung eine lokale Operation ist und beachtet werden muss, dass nur eine minimale Veränderung der eigentlichen Hauptinformation erfolgt. Das physikalische Problem für die Beleuchtung besteht in dem Kompromiss zwischen dem steilen Lichteinfall einerseits und der möglichen Reflexion andererseits. Dies ist für Matrixkameras und Zeilenkameras in der Bildebene viel schwerer zu erreichen als für Geräte mit beweglichen Zeilen in der Objektebene. Die Problematik wird in Abbildung 7 verdeutlicht. Da die Nachverarbeitungsschritte alle sehr empfindlich auf das Resultat einwirken, muss empfohlen werden alle scannerinternen Bildbearbeitungsfunktionen (wie Schärfen, Histogrammodifikationen) abzuschalten.

Optisches Filtern

Verschiedene Arten von Dokumenten in schlechter Druckqualität (Ormig, Durchschläge, Thermokopien, Faxausdrucke etc.) wurden untersucht. Für die Wellenlängenunterscheidung wurden Glasfilter mit relativ grosser Bandbreite benutzt. In jedem Falle war ein Kompromiss zwischen dem Signal/Rauschverhältnis und der passierenden Energiemenge erforderlich. Dies war der Grund, warum Interferenzfilter nicht anwendbar waren. Für das nahe Infrarot wurden Kantenfilter benutzt, die alle Wellenlängen unter 700 nm bzw. 850 nm abschnitten.

Für den NIR-Bereich trug das Papier vor allem zum Signal bei und abgesehen vom S/R-Verhältnis war die Erkennbarkeit der Buchstaben schlechter. Im sichtbaren Bereich des Spektrums ergaben einige Grün- und Orangefilter nach ausgiebiger Bildbearbeitung die interessantesten Resultate. Trotzdem muss festgestellt werden dass nur geringe bis gar keine Verbesserungen zum Originalscan mit dem gesamten Spektrum erkennbar sind. Abbildung 8 zeigt einige Resultate (von oben nach unten: ohne Filter, mit grün/orange mit orange. Von links nach rechts: unbehandelt, Kontrastbehandelt / Histogram und Farbbehandelt). Im Vergleich zum Aufwand sind die Ergebnisse zu vernachlässigen.

Schlussfolgerungen

Da die Eingabe entscheidend für die folgenden Schritte der Optimierung des Arbeitsablaufes ist, wird damit auch die Möglichkeit für eine Verbesserung der Resultate gegeben. Die Ergebnisse zur Scangeschwindigkeit, Beleuchtung und Filterung zeigen, wo eine sinnvolle Beeinflussung der Eingabe möglich ist.

Quellen

- The MEMORIAL Project description, Technical Annex, Jan. 2002 (IST-2001-33441)
- The METAe Project, EVA Berlin 2002
- Und beispielsweise der METAe newsletter in http://heds.herts.ac.uk/METAe/issue02.htm
- Zeutschel scanners: http://www.zeutschel.de
- I2s bookscanner: http://www.i2s-bookscanner.com
- MEMORIAL Deliverable 2: (see http://docmaster.eti.pg.gda.pl/)

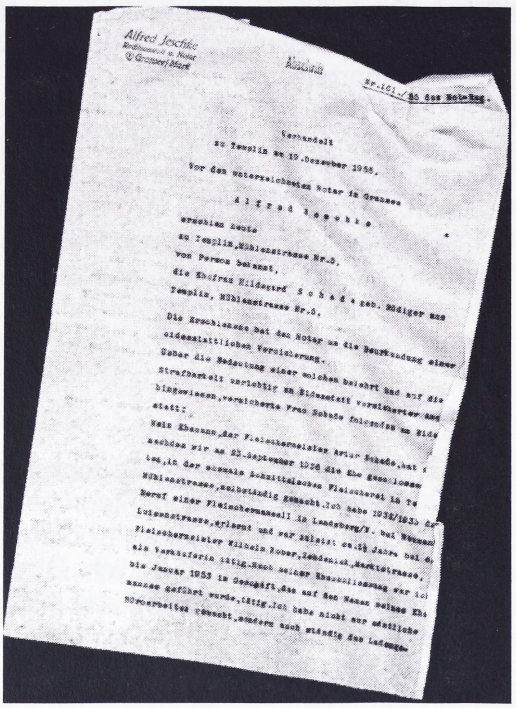

Abb. 1: Dokumentenscanner beschädigt Original

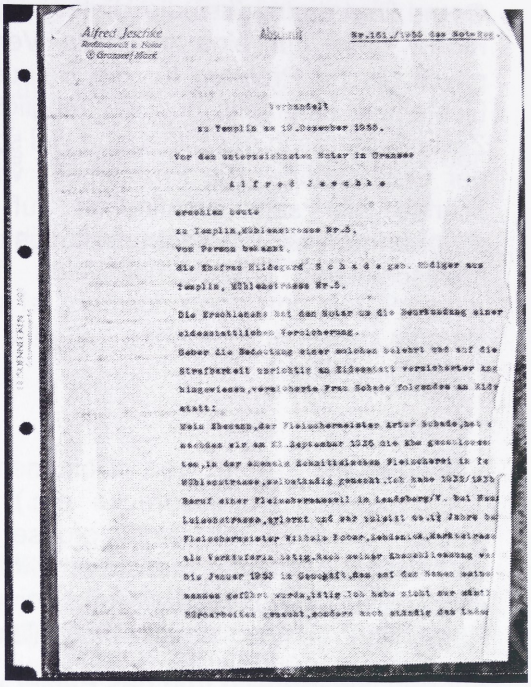

Abb. 2: Folien erzeugen Reflexe

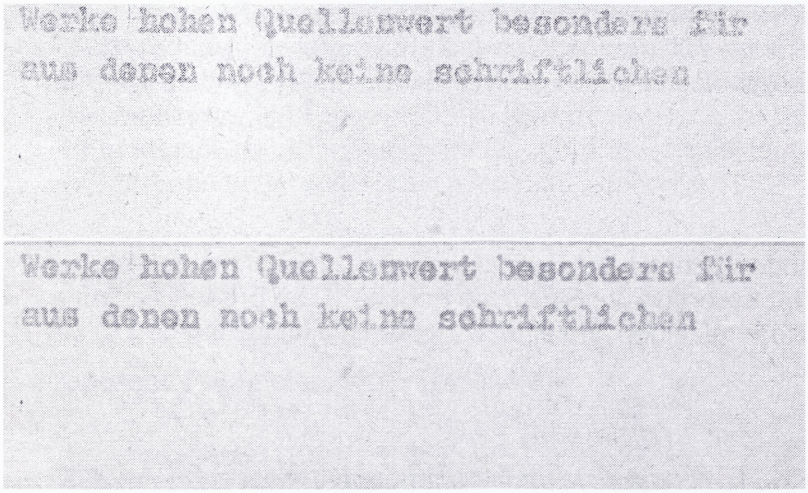

Abb. 3a: Ormigscan Fachbettscanner

Abb. 3b: Ormigscan Buchscanner

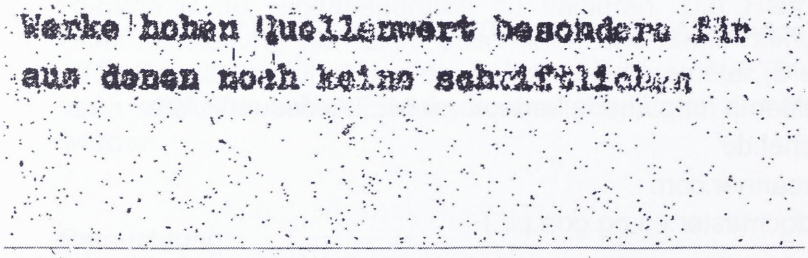

Abb. 3c: Manuelle optimale Binarisierung von 3a

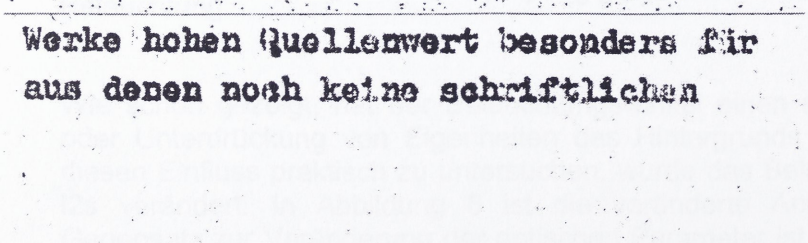

Abb. 3d: Manuelle optimale Binarisierung von 3b

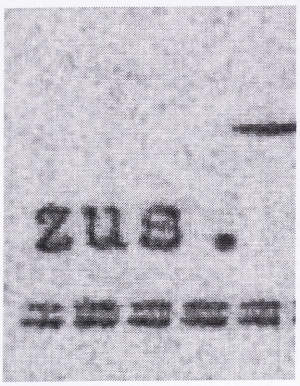

Abb. 4: Fragment eines Scans mit Buchscanner

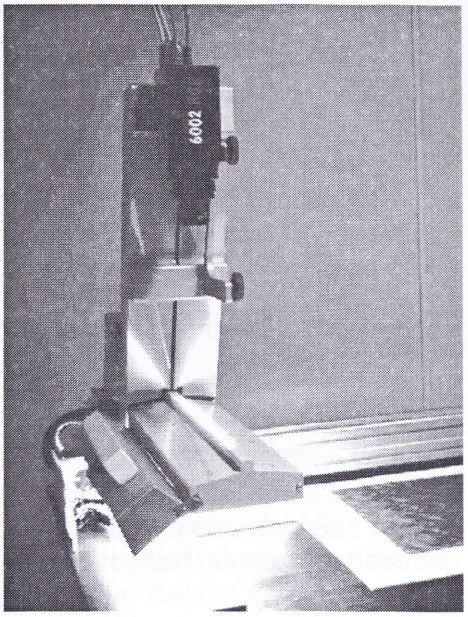

Abb. 6: Modifizierter Buchscanner mit Zusatzleuchte

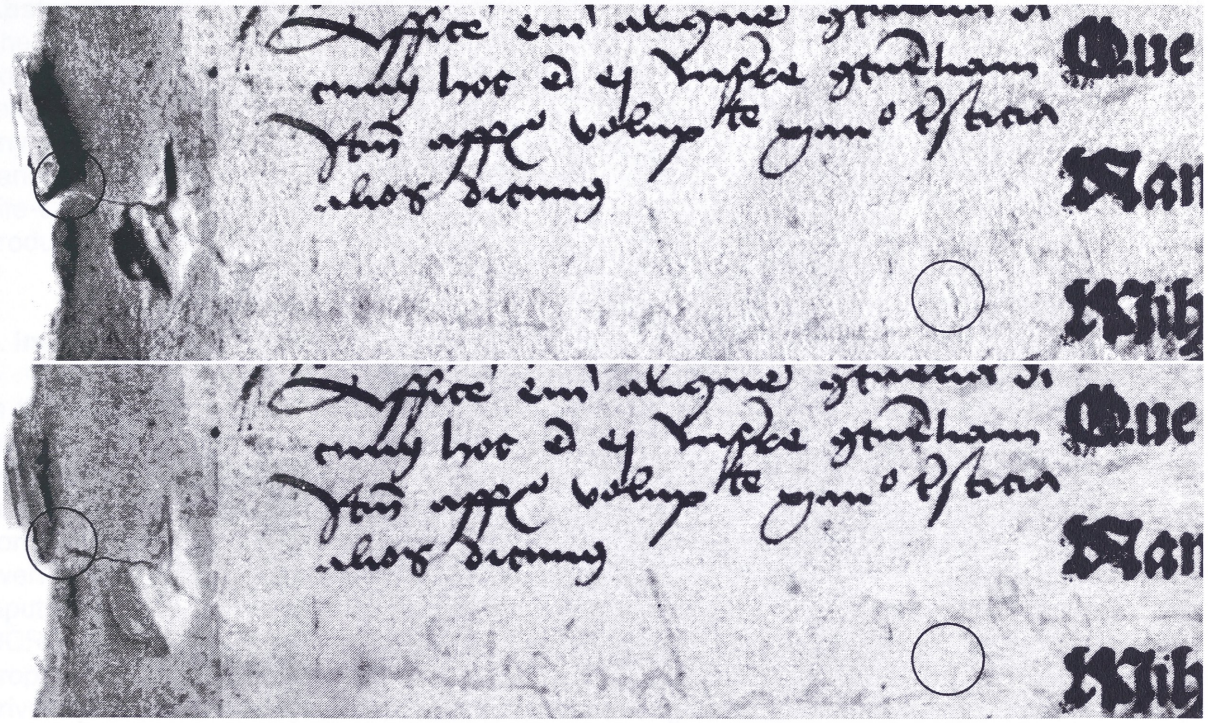

Abb. 5: Historisches Dokument mit flachem (oben) und steilem Lichteinfall gescannt

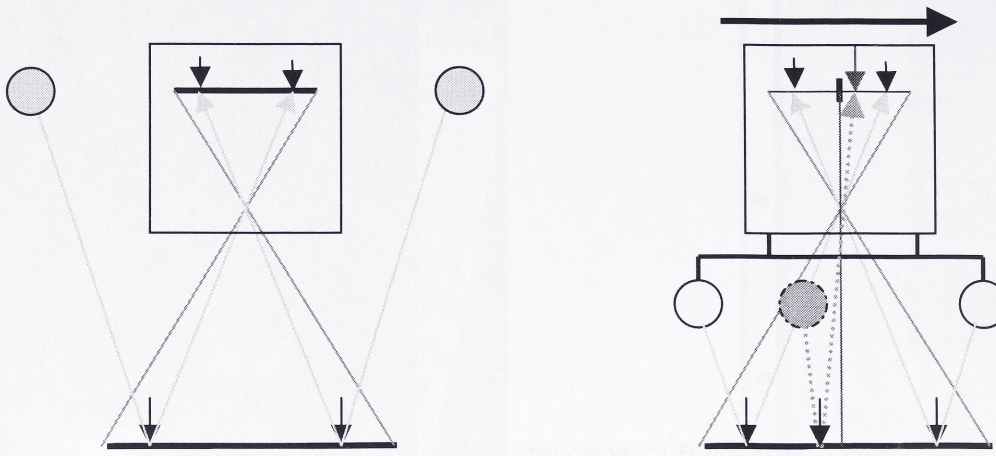

Abb. 7: Prinzip des Lichteinfalls und Verdeutlichung der Reflexionen beim Kamera- oder Bildebenenscanner (links) und einem Zeilenscanner, der sich in der Objektebene bewegt

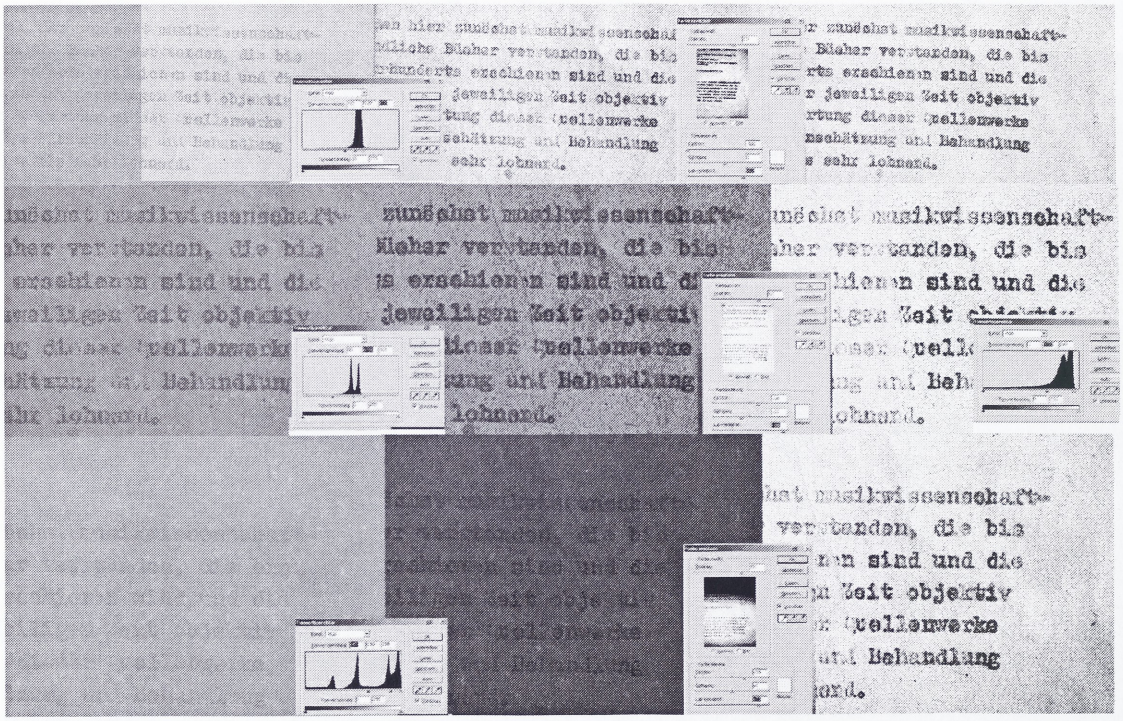

Abb. 8: Ergebnisse der Farbfilterung. Von oben nach unten: Ohne Filter, grün/ornage und orange. Von links nach rechts: Ausgangsbild, Histogrammodifikation, Farbmanipulation.