Abbildtreue und Kompression gescannter Dokumente bei PDF/A

Empfehlungen geeigneter Einstellungen

Datum: 29.11.2013

Autor: Axel Rehse
LuraTech Imaging GmbH

Thomas Zellmann
LuraTech Europe GmbH

Inhalt

Einleitung

Die Archivierung realer Dokumente wird in vielen Bereichen durch die digitale Archivierung abgelöst. Hieraus ergibt sich eine deutliche Verringerung hinsichtlich des Platzbedarfs aber auch des Aufwandes beim Zugriff auf die Informationen. Die Kostenersparnis durch die digitale Archivierung hängt von deren Effizienz ab. Die nötigen Informationen sollen mit einer möglichst geringen Datenmenge erhalten werden.

Bei gescannten Dokumenten handelt es sich um Bilddaten, die neben den eigentlichen Informationen auch redundante oder nicht relevante Daten enthalten können. Verschiedene Kompressionsverfahren bieten die Möglichkeit, die Datenmenge so zu verringern, dass vor allem die eigentlichen Informationen erhalten bleiben.

Die Kompressionsverfahren werden in verlustfreie und verlustbehaftete Verfahren unterschieden. Bei den verlustfreien Verfahren wird lediglich die Redundanz entfernt, so dass die Bilddaten zu 100% rekonstruierbar sind. Die Abbildung der komprimierten Bilddaten unterscheidet sich in diesem Fall nicht von der Abbildung der Bilddaten vor der Kompression. Bei den verlustbehafteten Kompressionsverfahren wird darüber hinaus versucht, die nicht relevanten Daten zu verringern.

In den folgenden Kapiteln werden Datenformate im Scan-Bereich sowie einzelne Kompressionsverfahren und deren Eigenschaften, insbesondere hinsichtlich Abbildungstreue, erläutert.

Formate gescannter Bilddaten

Nach dem Scanprozess können Bilddaten in verschiedenen Formaten vorliegen. Die Formate unterscheiden sich in folgenden Merkmalen: Auflösung, Farbtiefe, Kompression und Dateiformat.

Die Auflösung legt fest, wie viele Bildwerte in einem bestimmten räumlichen Bereich erfasst werden. Dadurch ergibt sich ein maßgeblicher Einfluss auf die Datenmenge und die Qualität. Beim Scannen von Papier sind Werte zwischen 150 und 600 dpi üblich. Ein geringerer Wert bedeutet eine reduzierte Datenmenge aber auch schlechtere Qualität. In der Praxis wird für Papier-Akten häufig eine Auflösung von 300 dpi als guter Kompromiss zwischen Qualität und Datenmenge verwendet. Beim Scan von Mikrofilm muss beachtet werden, dass die Dokumente bereits verkleinert vorliegen. Die technische Auflösung beim Abtasten des Films muss daher um den Verkleinerungsfaktor größer sein. Damit entspricht die Abtast-Auflösung nicht der Auflösung im Verhältnis zum abgelichteten Original, wie sie in der Ausgabedatei als Auflösung angegeben wird.

Die Farbtiefe bestimmt den Grad der Unterscheidung zwischen einzelnen Bildwerten. Bei 1 Bit Farbtiefe ist nur die Unterscheidung zwischen zwei Werten möglich, Schwarz und Weiß. Für die Darstellung von 256 Graustufen sind 8 Bit Farbtiefe nötig. Farbe wird bei gescannten Bilddaten in der Regel über drei Farbkanäle mit je 8 Bit realisiert (zusammen 24 Bit). Die Farbkanäle Rot, Grün und Blau ergeben gemischt die eigentlichen Farben. Die Wahl der Farbtiefe hängt vor allem davon ab, welche Rolle die Farbe im Dokument spielt und ob Farb- oder Helligkeitsabstufungen für den Inhalt des Dokuments wichtig sind. Fax-Dokumente sind zum Beispiel bereits eine Schwarz/Weiß-Vorlage, die sicher keinen Farb-Scan erfordert. Farbige Anmerkungen auf solchen Dokumenten können aber wiederum ein Scannen in Farbe erfordern.

Bereits während des Scanvorgangs können Bilddaten komprimiert werden. Oft wird von dieser Möglichkeit auch Gebrauch gemacht, insbesondere dann, wenn die gescannten Daten über ein Netzwerk zur Verfügung gestellt werden. Die Datenmenge ohne Kompression ist unter Umständen zu groß für die zur Verfügung stehenden Übertragungs- und Speicherkapazitäten. Außerdem kann so die Übertragungs- und Verarbeitungszeit verringert werden. Beim Einsatz verlustfreier Kompressionsverfahren, wie LZW, Deflate oder Fax, ergeben sich auch keine Nachteile hinsichtlich der Qualität. Farbige Bilddaten können aber nur begrenzt verlustfrei verringert werden. Außerdem stehen entsprechende Verfahren nicht bei allen Scannern oder jeder Scansoftware zur Verfügung. Die verlustbehaftete JPEG-Kompression ist dagegen weit verbreitet und wird bei allen bekannten Herstellern angeboten. Die Qualität kann hier im Allgemeinen durch eine entsprechende Einstellung variiert werden. Der bei höchster Qualitätseinstellung auftretende Qualitätsverlust ist in der Regel nicht sichtbar. Trotzdem kann dieser Qualitätsverlust in Einzelfällen zu Nachteilen in der späteren Verarbeitung führen. Die Qualität der OCR kann ebenso verringert

werden, wie die Effizienz und Qualität folgender weiterer Kompression. Wir empfehlen daher, wenn möglich, verlustfreie Kompression einzusetzen oder zunächst auf eine Kompression zu verzichten. Letztendlich ist eine Abwägung im Einzelfall zu treffen.

Für gescannte Bilder werden im Wesentlichen folgende Dateiformate eingesetzt: TIFF, JPEG und PDF. Bei TIFF und PDF handelt es sich um sogenannte Containerformate, die Bilddaten in verschiedene Bittiefen und Kompressionsverfahren enthalten können. Außerdem können in diesen Formaten mehrere Seiten abgelegt werden. Das JPEG-Dateiformat erlaubt dagegen nur ein Bild pro Datei und für die Kompression wird ausschließlich das gleichnamige JPEG-Kompressionsverfahren verwendet. Das Format ist für Graustufen- und Farbbilder ausgelegt und legt die Bittiefe auf 8 Bit (je Farbkanal) fest. In TIFF und PDF-Dateien können Daten auch unkomprimiert abgelegt werden. Das TIFF-Format unterstützt unter anderem folgende Kompressionsverfahren: LZW, Deflate, Fax G4 und JPEG. Die gleichen Kompressionsverfahren werden auch im PDF-Format unterstützt. Zusätzlich ist hier aber auch die Kompression mit JBIG2 und JPEG2000 möglich. Im Scanbereich haben die zuletzt genannten Bildkompressionsverfahren allerdings immer noch eine geringere Verbreitung als die weniger effizienten älteren Verfahren. Beim PDF-Format ist anzumerken, dass es sich hier eigentlich nicht um ein Dateiformat explizit für Bilddaten beziehungsweise Rasterbilder handelt. Vielmehr werden Dokumentseiten beschrieben, auf denen verschiedene Objekte platziert sein können. Neben Bilddaten können das auch Text, Vektorgrafiken und andere Inhalte sein.

Wir empfehlen für den Scan zunächst eine möglichst verlustfreie Kompression (z.B. TIFF LZW), da sich eine bestehende verlustbehaftete Kompression negativ auf die weitere Verarbeitung auswirken kann. JPEG-komprimierte Bilddaten sind daher nur bedingt geeignet und sollten zumindest nur mit maximaler Qualitätseinstellung verwendet werden.

Kompressionsverfahren

Die genannten Raten beziehen sich auf den Scan eines typischen Dokuments (hoher Textanteil) im A4 Format. Die Datenmenge für einen solchen Scan ergäbe ohne Kompression bei einer Auflösung von 300 dpi in Farbe ca. 24,5 MB Daten und in Schwarz/Weiß ca. 1 MB. Bei 200 dpi in Farbe ca.10,6 MB und in S/W ca. 450 KB.

Fax G4

Dieses Verfahren kann nur für Bilddaten mit einer Farbtiefe von 1 Bit eingesetzt werden, also Schwarzweiß-Bilder. Die Kompression ist verlustfrei und ändert die Qualität nicht. Die Rate der Datenverringerung hängt stark vom Inhalt ab und liegt für sauber gescannte A4 Seiten im Mittel bei 1/15.

JBIG2

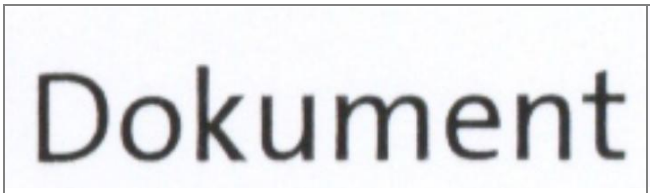
Diese Kompression ist ebenfalls nur für Schwarzweiß-Bilder (1 Bit Farbtiefe) anwendbar. Im Gegensatz zu Fax G4 bietet JBIG2 allerdings eine Vielzahl an Optionen, die unter anderem auch eine verlustbehaftete Kompression erlauben. Darüber hinaus ist eine deutlich bessere Datenverringerung möglich. Die mögliche Rate der Datenverringerung liegt für verlustfreie Kompression im Mittel bei ca. 1/30 und ist ebenfalls vom Inhalt abhängig. Bei Anwendung verlustbehafteter Kompression ist für Textdokumente im Mittel sogar eine Reduzierung auf 1/50 möglich.

Bei verlustbehafteter Kompression werden einzelne Bildwerte verändert. Die Veränderungen sind bei kleineren Symbolen sichtbar. Unter ungünstigen Umständen kann dadurch die Bedeutung kleiner Zeichen nicht mehr klar erkennbar sein oder gar verfälscht werden. Zum Beispiel kann sich anstelle des zuvor noch erkennbaren Großbuchstabens „B“ nach der verlustbehafteten Kompression ein Symbol befinden, das eher als Zahl „8“ interpretiert wird. Der verlustbehaftete Kompressions-Modus wird daher nur empfohlen, wenn eine Auflösung von mindestens 300 dpi verwendet wird und keine relevanten Informationen unter einer Schriftgröße von 10 Punkt vorhanden sind.

JPEG

Das JPEG-Verfahren dürfte die am weitesten verbreitete Kompression für Farb- und Graustufenbilder sein. Nur in Marktnischen, in denen eine höhere Kompression und bessere Qualität erforderlich ist, konnten sich andere Verfahren, wie zum Beispiel JPEG2000, durchsetzen. Die Rate der Datenverringerung ist vor allem von der Qualitätseinstellung aber auch vom Bildinhalt abhängig. Bei maximaler Qualität wird für eine typische Dokumentseite eine Rate von etwa 1/20 erreicht. Maximale Qualität bedeutet bei JPEG jedoch keine verlustfreie Kompression. Die Veränderung der Bildwerte ist bei maximaler Qualität allerdings nicht mehr wahrnehmbar. In der Regel

wird jedoch eine stärkere Verringerung der Datenmenge bei geringerer Qualität angewendet. Die typische Rate liegt dann bei ca. 1/100 In diesem Fall sind bei genauerer Betrachtung Störungen sichtbar.

	
original (vergrößert)	JPEG komprimiert mit Qualität 6 (1/100)

JPEG2000

Ursprünglich als Ablösung für JPEG gedacht, zeichnet sich die etwas komplexere JPEG2000-Kompression vor allem durch bessere Kompression und Qualität aus. Außerdem ist es möglich verlustfrei zu komprimieren. Bei verlustfreier Kompression wird die Datenmenge gescannter Bilder durchschnittlich auf etwa 1/3 bis 1/4 reduziert. Der Bildinhalt bestimmt hier maßgeblich die Rate. Die verlustbehaftete Variante erlaubt eine deutlich stärkere Verringerung. Bei maximaler Qualität wird in der Regel eine Rate von 1/50 erreicht. Bei nur etwas geringerer Einstellung der Qualität kann kaum eine Verschlechterung wahrgenommen werden. Trotzdem ist eine deutliche Verringerung möglich. Bei Qualität 8 beträgt die Rate etwa 1/70. Bei Qualität 6 mit ca. 1/100 zeichnet sich die Verringerung der Qualität eher durch eine leichte Unschärfe in bestimmten Bereichen aus.

JPEG2000 kann in PDF-Dokumenten ab Spezifikation 1.5 (Kompatibilität für Reader 6) beziehungsweise PDF/A-2 verwendet werden.

Deflate

Dieses verlustfreie Verfahren wird unter anderem auch im ZIP-Format zur Kompression allgemeiner Daten und speziell für die Bildkompression im PNG-Format verwendet. Im PDF Format kann es zur Kompression von Datenströmen ebenso eingesetzt werden, wie für Bilddaten. Da es sich um ein verlustfreies Verfahren handelt, werden die Bilddaten nicht verändert und in voller Qualität erhalten. Die Rate der Datenverringerung hängt hier besonders stark vom Inhalt ab. In Extremfällen kann gar keine Reduzierung der Datenmenge erreicht werden. Gescannte Dokumente können in der Regel auf 1/3 bis 1/4 reduziert werden. Über eine Vorfilterung der Daten kann die Datenmenge insbesondere bei Bildern noch weiter verringert werden. In diesem Fall ist das Ergebnis meist aber noch etwas größer als bei verlustfreier Kompression mit JPEG2000.

MRC

Im engeren Sinne handelt es sich bei MRC (Mixed Raster Content) um eine Technik, bei der ein Algorithmus zur Bildsegmentierung eingesetzt wird, um bestehende Algorithmen zur Bildkompression selektiv anzuwenden. Die Technik kombiniert die Vorteile verschiedener Bildkompressionsverfahren. Aus diesem Grund bleiben

Farbinformationen sowie eine gute Lesbarkeit bei deutlicher Verringerung der Datenmenge erhalten.
Durch die Segmentierung wird das Bild in drei Bildanteile oder Ebenen (Layer) zerlegt: ein Schwarzweiß Textbild, eine Vordergrundfarbebene und ein farbiges Hintergrundbild.
Die Vordergrundfarbebene wird auf den Text abgebildet und färbt diesen somit ein. Das so erhaltene Farb-Text-Bild überlagert den Hintergrund, so dass sich daraus das Gesamtbild ergibt. Die einzelnen Ebenen werden entsprechend ihres Inhalts komprimiert. Die Schwarzweiß Text-Ebene wird standardmäßig mit verlustfreiem JBIG2 komprimiert. Für die Farbebenen wird dagegen JPEG2000 eingesetzt. Alternativ können aber auch FaxG4 und JPEG verwendet werden. Die Einstellungen zur Auswahl der jeweiligen Kompression sind im entsprechenden vorherigen Abschnitt genannt. Je nach Qualitätseinstellung wird die Auflösung der Farbinformationen außerdem verringert. Die Textanteile behalten jedoch immer die volle Auflösung.

Die MRC Kompression erlaubt je nach Qualitätseinstellung und Dokumentinhalt eine Verringerung auf 1/300 bis 1/700 der ursprünglichen Datenmenge. Das Ergebnis ist damit trotz zusätzlicher Farbinformation oft nicht größer als ein Fax-G4-komprimiertes Schwarzweiß-Bild.

Zusammenfassung

Bei der Archivierung gescannter Dokumente ist eine Entscheidung zu treffen, wie stark und in welcher Weise sich die Abbildungen vom Original unterscheiden dürfen. Beim Scannen wird bereits festgelegt, ob zum Beispiel Farbinformationen übernommen werden und wie groß die Auflösung ist.

Die Entscheidung für ein Kompressionsverfahren hängt einerseits davon ab, welche Informationen in den Dokumenten wichtig sind. Andererseits ist entscheidend, wie groß, beziehungsweise teuer, der zur Verfügung stehende Speicherplatz ist.

Schwarzweiß-Dokumente belegen bereits wenig Speicher. Für diesen Dokumenttyp ist daher eine völlig verlustfreie Archivierung noch vergleichsweise günstig. Als optimales Kompressionsverfahren bietet sich hier JBIG2 im verlustfreien Modus an.

Für Farb- oder Graustufen-Dokumente ist JPEG weit verbreitet. Andere Verfahren wie JPEG2000 und MRC erlauben jedoch eine viel höhere Kompression bei guter Dokumenten-Qualität.

Referenzen / Standards

Deflate	RFC 1951
Fax G4	ITU-T T.6
JBIG2	ITU-T T.88, ISO/IEC 14492
JPEG	ITU-T T.81, ISO/IEC IS 10918-1
JPEG2000	ITU-T T.800, ISO/IEC 15444-1 und ITU-T T.801, ISO/IEC 15444-2
MRC	ITU-T T.44, ISO/IEC 16485
PDF	ISO 32000
PDF/A	ISO 19005-1, ISO 19005-2 und ISO 19005-3
TIFF	Revision 6.0 (1992), http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf