Optimale Erstellung und Komprimierung von PDF-Dateien aus gescannten Dokumenten

Timo von Wysocki

20. April 2016

Version 1.1

Inhaltsverzeichnis

1 Motivation

Aufgrund der Beschränkung der Dateigröße hochzuladender Dokumente bei Onlinebewerbungen ist es nötig, die Dateigröße von eingescannten Dokumenten massiv zu verkleinern. Hierfür werden im Folgenden verschiedene Verfahren beschrieben. Diese eignen sich in erster Linie für Dokumente und weniger für etwa Fotos. Die Verfahren erheben keinen Anspruch darauf, die optimale Lösung zu sein, sie reduzieren die Dateigröße jedoch drastisch.

Verbesserungsvorschläge dürfen gerne an 77timo@gmail.com gerichtet werden.

Version 1.1: Kapitel 4.1.1 von Maximilian Naumann hinzugefügt.

2 Vorbereitung

2.1 Verwendete Software

Zur Erfassung des Dokuments wird ein Scanner benötigt, sowie eine Möglichkeit, die gescannte Datei auf den Rechner zu übertragen.

Zur komprimierung gescannter Bilddateien wird das open-source-Programm GIMP (GNU Image Manipulation Program) von `https://www.gimp.org/` verwendet.

Sollen die gescannten Dokumente zu einer mehrseitigen PDF-Datei zusammengefasst werden, wird in dieser Erklärung ausführlich der Weg über LaTeX beschrieben, es eignet sich jedoch mit Einschränkungen auch ein beliebiges PDF-Drucktool wie etwa der *PDF-Creator* von `http://de.pdfforge.org/pdfcreator` oder die Verwendung von Word.

2.2 Scannen

Eingescannte Dokumente wie Zeugnisse oder Zertifikate sollten direkt als Bilddatei vorliegen. Bei vielen Scannern ist ein Speichern als TIFF-Datei möglich, was eine verlustfreie Ausgangsdatei garantiert. Ist dies nicht möglich kann der Scan als JPG-Datei erfolgen. Dokumente in PDF-Form können jedoch ebenso wie Bilddateien in *Gimp* importiert werden. Zu beachten ist lediglich, dass dies für jede Seite separat zu erfolgen hat.

3 Verringerung der Dateigröße

3.1 Skalierung

Das Skalieren der Bildgröße des eingescannten Dokuments reduziert die Dateigröße massiv, unabhängig vom verwendeten Farbreduzierungsverfahren aus Abschnitt 3.2. Damit kann eine Skalierung immer zusätzlich erfolgen.

- Bilddatei in *Gimp* öffnen
- Bild → Bild skalieren → Breite und Höhe um den selben Wert skalieren, zum Beispiel 75%. Die Einheit Prozent kann rechts des Kettensymbols eingestellt werden.
- Die Werte für X- und Y-Auflösung sollten ebenfalls angepasst werden, um die Handhabung der Bilddateien später zu vereinfachen. Dazu wird der Wert um den selben Prozentsatz wie im Schritt zuvor reduziert. Hier kann für eine Reduzierung auf 75% direkt im Feld etwa 300*0.75 eingetragen werden.
- Mit skalieren bestätigen
- Gewünschte Farbreduzierung durchführen oder direkt als JPG-Datei speichern über Datei → exportieren als... → Dateiname.JPG → Exportieren → Qualität auf 50 oder weniger → Exportieren

3.2 Farbreduzierung

Die im Folgenden Abschnitt beschriebenen Verfahren nehmen nacheinander immer mehr Farbe des Dokuments weg und reduzieren damit die Dateigröße immer mehr. Je nach Anwendungszweck und benötigter Kompression kann ein entsprechendes Verfahren gewählt werden. Alle Verfahren können zusätzlich um die in Abschnitt 3.1 beschriebene Skalierung ergänzt werden, um die Dateigröße weiter zu verkleinern.

3.2.1 Export in Farbe, ohne sichtbare Verluste

Hier wird die Qualität der JPG-Datei soweit verringert, dass noch keine Sichtbaren Qualitätsverluste auftreten. Der Effekt ist nicht sonderlich groß, dafür liegt jedoch kein sichtbarer Verlust vor.

- Bilddatei in *Gimp* öffnen
- Datei → exportieren als... → Dateiname.JPG → Exportieren → Qualität auf 50 oder weniger → Exportieren

3.2.2 Export in Farbe, leichte sichtbare Verluste

Hier wird die Anzahl der im Bild vorhandenen verschiedenen Farben reduziert. Dabei werden sehr ähnliche Farben zu einer gleichen Farbe zusammengefasst, was die Dateigröße deutlich reduziert.

- Bilddatei in *Gimp* öffnen
- Farben → Posterisieren → Farbanzahl auf etwa 10 oder einen passenden, möglichst kleinen Wert → Ok
- Bild → Modus → Indiziert ... → Optimale Palette erzeugen mit Anzahl der Farben gleich dem Wert des vorherigen Schritts (hier 10) → Umwandeln
- Datei → exportieren als ... → Dateiname.PNG → Exportieren → Kompressionsniveau auf 9 → Exportieren

Anmerkung: JPG kann keinen indizierten Farbraum benutzen und würde so das Bild wieder in RGB-Farbraum zurückumwandeln. Daher der Export als PNG.

3.2.3 Export in Graustufen

Der Export in Graustufen wandelt das Bild in ein Bild ohne Farbe um. Man erhält das, was umgangssprachlich als Schwarz/Weiß bezeichnet wird, jedoch sind hier auch Zwischenstufen, also Grautöne möglich.

- Bilddatei in *Gimp* öffnen
- Bild → Modus → Graustufen
- Bild → Modus → Indiziert ... → Optimale Palette erzeugen mit Anzahl der Farben etwa 4 (ausprobieren, wie wenige Grautöne noch akzeptabel sind) → Umwandeln
- Datei → exportieren als ... → Dateiname.PNG → Exportieren → Kompressionsniveau auf 9 → Exportieren

3.2.4 Export in Schwarz/Weiß

Der Export in Schwarz/Weiß lässt lediglich komplett schwarze oder komplett weiße Bildpunkte zu. Da helle Farben dabei leicht als komplett weiß interpretiert werden, wird das Bild zuerst in ein Graustufenbild umgewandelt und anschließend über einen Schwellwert in Schwarz und Weiß zerlegt. Hiermit kann eine Komprimierung auf ein Zehntel der originalen Dateigröße erfolgen.

- Bilddatei in *Gimp* öffnen
- Farben → Entsättigen → Helligkeit → Ok

- Farben → Schwellwert → Guten Wert finden, bei dem die gewünschten Details noch erkennbar sind, z.B. 210 → Ok
- Bild → Modus → Indiziert ... → Schwarz/Weiß-Palette (1-Bit) verwenden → Umwandeln
- Datei → exportieren als ... → Dateiname.PNG → Exportieren → Kompressionsniveau auf 9 → Exportieren

4 Erstellung von PDF-Dateien

4.1 Latex

Am flexibelsten und speicherplatzsparendsten ist die Erstellung über LaTeX. Hierfür wird die folgende tex-Datei erstellt:

```
\documentclass[
a4paper,
]{scrartcl}
\usepackage[absolute]{textpos}
\usepackage{graphicx}
\pagestyle{empty}

\setkeys{Gin}{width=20cm}        % Bildbreite

\begin{document}

% Blockbeginn

\begin{textblock*}
{200mm}
(2.5mm,7.5mm)                    % Vertikale und horizontale
   Platzierung des Bildes, gemessen von der linken oberen
   Ecke
\includegraphics{bild.PNG}       % Muss im selben Ordner wie
   die tex-Datei liegen, ansonsten Pfad mitangeben
\end{textblock*}
\leavevmode
\newpage

% Blockende

\end{document}
```

Der Befehl `\setkeys{Gin}{width=20cm}` bestimmt die Breite der mittels des Befehls `\includegraphics{bild.PNG}` eingefügten Bilder.

Soll die so definierte Standardbreite für ein einzelnes Bild überschrieben werden, muss das Bild über den Befehl `\includegraphics[width=10cm]{bild.PNG}` eingebunden wer-

den. Die Bilddateien müssen im selben Ordner liegen, wie die tex-Datei. Andernfalls muss der Pfad explizit mitangegeben werden.

Die beiden Zahlen in runden Klammern als Option des `\begin{textblock*}`-Befehls geben die absolute Platzierung des Bildes auf der Seite an. Dabei gibt die erste Zahl den vertikalen und die zweite den horizontalen Abstand von der linken oberen Ecke der aktuellen Seite an.

Für jedes einzufügende Bild ist der zwischen `% Blockbeginn` und `% Blockende` befindliche Code zu kopieren und nach `% Blockende` einzufügen.

4.1.1 Zusammenfassung zu einem Gesamtdokument

Maximilian Naumann hat freundlicherweise einen Latex-Code zur Verfügung gestellt, mit dem sich einfach Dokumente zusammenfassen lassen, inklusive Titelseite, Inhaltsverzeichnis und Kopf- bzw. Fußzeilen.

```
\documentclass[a4paper,12pt]{article}

% Packages:
        \usepackage[]{geometry}% http://ctan.org/pkg/geometry
            %command showframe to show the frames
        \usepackage{fancyhdr}
        \usepackage[german]{babel}

        \usepackage[final]{pdfpages}
        \usepackage{verbatim}

        \usepackage{hyperref}
        % Links in blue
        \hypersetup{
            colorlinks=true,
            citecolor=black,
            filecolor=black,
            linkcolor=blue,     % change this "blue" to "
               black" to get a black table of contents
            urlcolor=black
        }

        % Disable LaTeX default twoside layout
        \setboolean{@twoside}{false}

        % Settings for pdf import
        %   DIN A 4: 210 mm wide,  297 mm high
        \geometry{
                left=1.05cm,
```

```
                right = 1.05cm,
                top=1.5cm,
                bottom = 2cm,
                headheight = 1 cm,
                headsep =0cm,
                footskip = 1cm
        }
        \savegeometry{pdf-import}

        % Settings for titel page
        \geometry{
                left=1in,
                bottom = 1in,
                right = 1in,
                top=1in,
                headheight = 1 cm,
                headsep =0cm,
                footskip = 1cm
        }


% Notes:
% use geometry package, DON'T change the default margins instead
% phantomsection must be before addcontentsline for the references in the TOC



\begin{document}

%%%%%
%% Titlepage
%%%%%

        \title{Titel \\ \vspace{1cm}
                                Titelzeile 2 \\ \vspace{1cm}
                                 Titelzeile 3}
        \author{Maximilian Mustermann}
        \date{01.01.1970}

        \maketitle
        \vspace{2cm}
```

```
69          \tableofcontents
70          \thispagestyle{empty}
71          \pagebreak
72
73  %%%%%
74  %% PDF import
75  %%%%%
76
77  % Settings for pdf import:
78          \loadgeometry{pdf-import}
79          \pagestyle{fancy}
80
81          \fancyfoot{}
82          % Header
83          \lhead{}
84          \chead{}
85          \rhead{}
86          % Footer
87          \lfoot{footer} % Textline in footer
88          \cfoot{}
89          \rfoot{\thepage} %absolute page number
90
91          % lines separating the footer and the header
92          \renewcommand{\headrulewidth}{0.4pt}
93          \renewcommand{\footrulewidth}{0.4pt}
94
95  % actual pdf import:
96          %% Sec 1
97          \phantomsection
98          \addcontentsline{toc}{section}{Section 1}
99
100         %% Subsec 1.1
101         \phantomsection
102         \addcontentsline{toc}{subsection}{Subsection 1.1}
103
104         \lhead{Header (1/2)}    % header on this page
105         \includepdf[pages=1, frame, scale=0.9, pagecommand
               ={}]{status-lua} % <--- here is where the pdf file
                is imported!!
106                 % status-lua is a sample pdf page in most TeX
                        distributions, replace it by Lebenslauf.
                       pdf or whatever other file
107
```

```
108         \lhead{Header (2/2)}
109         \includepdf[pages=1, frame, scale=0.9, pagecommand
               ={}]{status-lua}
110
111         %% Sec 2
112         \phantomsection
113         \addcontentsline{toc}{section}{Section 2}
114
115         %% Subsec 2.1
116         \phantomsection
117         \addcontentsline{toc}{subsection}{Subsection 2.1}
118
119         \lhead{Document 2 (1/3)}
120         \includepdf[pages=1, frame, scale=0.9, pagecommand
               ={}]{status-lua}
121
122         \lhead{Document 2 (2/3)}
123         \includepdf[pages=1, frame, scale=0.9, pagecommand
               ={}]{status-lua}
124
125         \lhead{Document 2 (3/3)}
126         \includepdf[pages=1, frame, scale=0.9, pagecommand
               ={}]{status-lua}
127
128 \end{document}
129 \endinput
```

4.2 PDF-Creator

Soll lediglich eine einseitige PDF-Datei aus einem Bild erstellt werden, kann das Bild über den *PDF-Creator* gedruckt werden. Dazu das Bild mit der Windows Fotoanzeige öffnen (Standard-Doppelklickfunktion) und über Drucken → Drucken... Das Druckfenster öffnen. Als Drucker den *PDF-Creator* auswählen und anschließend drucken.

4.3 Word

Mehrseitige PDF-Dokumente können sehr einfach in Word erstellt werden. Dazu die Bilder in der gewünschten Reihenfolge in eine neue Worddatei einfügen über Datei → Speichern unter → PDF als PDF-Datei abspeichern. Die Dateigröße wird hierbei jedoch deutlich größer als über den LaTeX-Weg, weshalb dieses Vorgehen einen Teil der vorherigen Verringerung der Dateigröße wieder zunichte macht.