OCR – Was macht's? Was bringt's?

OCR ist die Abkürzung für Optical Character Recognition und bedeutet optische Zeichenerkennung. Software, die so etwas kann, wird dazu benutzt, dem Computer das Lesen beizubringen. Sicherlich fragt man sich nun, wie wird das gemacht, denn schließlich hat ja der Computer keine Augen, und wozu braucht man so etwas.

Der Hintergrundgedanke ist, gedruckte Dokumente so in den Computer einzugeben, dass man danach den digitalisierten Text mit normalen Textverarbeitungsprogrammen – wie z. B. Microsoft Word – bearbeiten kann. Natürlich möchte man das Dokument nicht abtippen, und außerdem soll auch noch das Layout erhalten bleiben.

Vom Dokument zur Textdatei

Für die Zeichenerkennung sind drei wesentliche Arbeitsschritte nötig:

1. Scannen des Dokuments
2. Konvertieren des gescannten Rasterbildes in Textzeichen
3. Abspeichern des Dokumentes im ursprünglichen Layout in ein gewünschtes Textformat zur Weiterbearbeitung

Für das Testen von Zeichenerkennungssoftware standen zwei Programme zur Verfügung, zum einen Recognita 5.0 der Firma Recognita Corp. Hungaria und zum anderen OmniPage Pro 10 der Firma Caere Corp. (1996 erwarb Caere die Firma Recognita und verband die Technologien zur Zeichenerkennung beider Firmen im Produkt OmniPage 10).

Beide Programme bieten eine gute Unterstützung beim Scannen von Dokumenten, z. B. ist es möglich, schief eingescannte Dokumente zu drehen. Man kann aber auch mit normalen Scannprogrammen die Dokumente scannen und sie dann als Bilddatei in einem gängigen Pixelformat (jpg, tif, pcx, img, bmp) abspeichern. Diese Dateien können dann in den OCR-Programmen geöffnet werden. Beim Scannen empfiehlt es sich, in Graustufen oder Farbe zu scannen und eine Auflösung von 300 dpi zu wählen. Eine höhere Auflösung verbessert die Ergebnisse nicht.

Automatische Zeichenerkennung

Beide Programme können automatisch arbeiten. Sie teilen das eingelesene Dokument in Zonen bzw. Bereiche (Bezeichnung ist softwareabhängig) ein. Diesen Bereichen werden Typen zugeordnet, wie z. B. Text, auch mehrspaltiger, Bild oder Tabelle. Die OCR-Software versucht dann, den Text als Text zu erkennen, die Bilder als Grafiken beizubehalten und die Tabellen als Tabellen aufzubereiten. Es wird sogar von den Herstellern behauptet, dass Tabellen von den meisten Datenbank- und Tabellenkalkulationsprogrammen erkannt werden.

Hinter diesen Erkennungsalgorithmen verbergen sich Konturanalysen für Zeichen, viele Expertensysteme, komplizierte mathematische und statistische Berechnungen und Wörterbücher, um zum einen die richtigen Zeichen und zum anderen ein richtiges Wort zu bestimmen. Und so ganz nebenbei soll natürlich der Erkennungsprozess so schnell wie möglich vonstatten gehen.

Beeinflusst wird die Genauigkeit der Erkennung durch die Einstellung der Optionen, die man sich vor der Benutzung der OCR-Software genau ansehen sollte. Hier findet man Scannereinstellungen, Spracheinstellungen, Schriftarteinstellungen und anderes.

Beide Programme sind nur dafür ausgelegt, gedruckte Dokumente mit sachlichen Schriftarten zu verarbeiten. Schreibschriften oder Schriften mit vielen Verzierungen werden sehr schlecht oder gar nicht erkannt. Eine kleine Ausnahme bei der Erkennung bietet Recognita. Hier kann man als Zonentyp auch *handgeschriebene Zahlen*, *Ankreuzungen* (auf Formularen), *Strichcode* und *Nadeldruck* zusätzlich zur besseren Beschreibung und Erkennung der Zonen angeben.

Dann setzt der Korrekturprozess ein. Falsche Worte und das Layout können korrigiert werden. Ist die Bearbeitung abgeschlossen, wird die Datei in einem der vielen verschiedenen Textformate abgespeichert.

Manuelle Zeichenerkennung

Mitunter ist es notwendig, die OCR-Software ohne Automatik zu betreiben. Das ist ganz wichtig, wenn man die Erkennung von Bereichen verbessern möchte und der OCR-Software die größtmöglichste Unterstützung bei der Erkennung geben will. Außerdem wird diese Methode auch benutzt, wenn man nur Teile eines Dokuments im Computer verfügbar machen will. Dazu legt man mittels Maus die Bereiche fest und weist diesen die entsprechenden Eigenschaften (Text, Tabelle oder Grafik) zu.

Wie sollten diese Bereiche optimal ausgewählt werden? Beim Programm Recognita erweist es sich als günstig, das Dokument in möglichst kleine Bereiche einzuteilen. Das unterstützt neben der besseren Schrifterkennung auch genauer die Layoutnachbildung. Die Erfahrungen mit Recognita zeigen, dass fette Schriften selten wiedergegeben werden, eingerückte Gliederungspunkte nicht im richtigen Layout erscheinen und die Schriftart in einem etwas kleinerem Schriftgrad dargestellt wird.

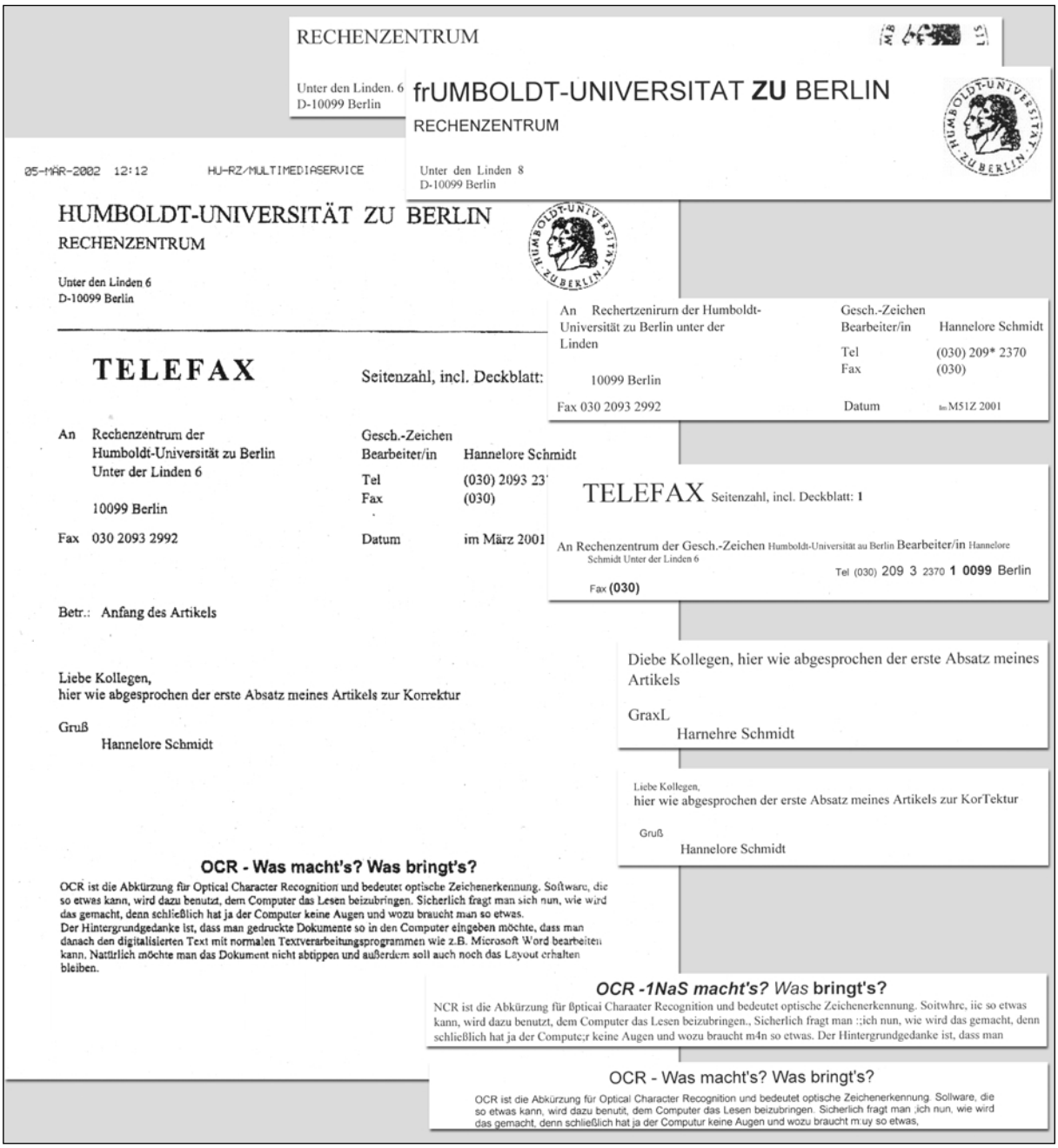

Abb. 1: Fax mit automatischer Zeichenerkennung durch OmniPage Pro (obere Auszüge) und Recognita (untere Auszüge) konvertiert

Bei der Anwendung des Programms OmniPage Pro sollte das Dokument in möglichst große Bereiche eingeteilt werden, da das Programm dann neben einer guten Texterkennung auch ein besseres Layout erstellen kann.

In beiden Programmen sind Einstellungen zur Schriftart möglich. Während man OmniPage Pro mit der Festlegung der Schriftarten, die im Dokument vorhanden sind, bei der Texterkennung helfen kann, lässt Recognita nur bestimmte Schriftarten zum Abspeichern der erkannten Dokumente zu.

Eine weitere Verbesserung der Erkennung kann erreicht werden, wenn man die Lernfähigkeit der Programme ausnutzt. Recognita und OmniPage Pro bieten die Möglichkeit, Buchstabenerkennung zu trainieren. Bei schlecht gedruckten oder schwer erkennbaren Zeichen sollte man diese bei der Korrektur nicht einfach ersetzen, sondern das Training starten. Das ermöglicht dann die automatische Korrektur dieser Zeichen im gesamten Dokument anhand der in einer Trainingstabelle enthaltenen trainierten Zeichen.

Zur Erkennungsverbesserung können auch Benutzerwörterbücher herangezogen werden. Die mitgelieferten

Wörterbüchern enthalten meist nur den normalen Wortschatz. Handelt es sich bei den Dokumentvorlagen jedoch um Fachtexte mit Fachbegriffen, sollten diese in ein nutzereigenes Wörterbuch aufgenommen werden, damit sie automatisch angewandt werden können.

Lohnt sich ein Einsatz von Zeichenerkennungsprogrammen?

Wie so oft ist der Einsatz einer bestimmten Software von der Art der Aufgabe abhängig. Wenn man sehr viele gedruckte Dokumente mit einheitlicher Schriftart auf dem Computer verfügbar machen will und beim Test mit der OCR-Software nur wenige Fehler aufgetreten sind, dann lohnt sich der Einsatz schon. Beide Programme sind, wie oben beschrieben, lernfähig. Man kann mit ihnen solange trainieren, bis eine richtige Erkennung erfolgt. Dieser Aufwand lohnt sich aber nicht, wenn man viele gedruckte Dokumente im Computer erfassen soll, die von unterschiedlichen Druckern stammen und unterschiedliche Schriftarten enthalten. Denn hier müsste der Lernprozess für jedes andersartige Dokument erneut durchgeführt werden.

Angepriesen wird auch das Lesen von gefaxten Dokumenten. In der Abb. 1 sind das Originalfax und die vom OmniPage Pro und Recognita erkannten Textauszüge zum Vergleich dargestellt. Leider lässt sich schlecht abschätzen, ob man hier nicht das Fax schneller in den Computer eingetippt als mit OmniPage Pro bzw. Recognita korrigiert hat.

Ein wichtiger Hinweis noch zuletzt. Zwar hat sich im Laufe der Jahre die OCR-Software gewaltig entwickelt, aber eine hundertprozentige Zeichenerkennung und ein Layout wie im Originaldokument kann sie noch nicht erzeugen. Daraus ergibt sich, dass man jede Seite einzeln betrachten, Korrekturen durchführen und das Layout verbessern muss. Wenn man so etwas sehr gewissenhaft macht, nimmt es doch viel Zeit in Anspruch.

Hannelore Schmidt
hschmidt@rz.hu-berlin.de