

Buchstaben-Recycling

Will man Visitenkarten in der Datenbank archivieren oder interessante Zeitungsartikel als E-Mail verschicken, erledigt Texterkennung das mühsame Abtippen.

● von Markus Zitt

Wer immer wieder Telefonbücher, Lexika oder Zeitschriften nach bestimmten Informationen durchsuchen muss, weiss um die Vorteile elektronischer Medien. Auf dem Computer sind in einer Textdatei oder Datenbank Wörter in Sekundenschnelle gefunden. Die Texte ganzer Ordner finden Platz auf nur einer Diskette. Doch erst müssen Informationen, die in gedruckter Form vorliegen, in den Computer gelangen. Mit einem Texterkennungsprogramm erledigt der PC die monotone Tipp-Arbeit und nach zwei Minuten kann der Inhalt einer A4-Seite bereits bearbeitet oder versendet werden.

Auch in diesem Moment führen Sie eine Texterkennung durch. Eigentlich sehen Sie nur schwarze Farbe auf weissem Papier. Dieses Schwarzweiss-Muster wird auf den Netzhäuten Ihrer Augen abgebildet, dort in Nervenimpulse umgewandelt und zum visuellen Cortex im Hinterkopf weitergeleitet. In jenem Teil des Hirns wird das Gesehene erstmals analysiert, mit Ihrem Wissen über das Aussehen von Buchstaben und Ihrem passiven Wortschatz verknüpft – Sie lesen.

Für die Texterkennung per Computer werden gedruckte Vorlagen mit einem Scanner digitalisiert, d.h., es entsteht ein Muster aus quadratischen Bildpunkten (Pixeln) unterschiedlicher Helligkeit. Dieses wird von einem OCR-Programm analysiert, indem Stellen mit hohem Kontrast als Objektkonturen interpretiert werden. Die so gewonnenen Objekte werden auf Gesetzmässigkeiten hin untersucht, die einen Rückschluss auf den Buchstaben erlauben. So lässt sich beispielsweise ein «C» durch seine ovale Form mit einer Öffnung rechts beschreiben. Die erkannten Buchstaben «tippt» das Programm in eine Textdatei und abschliessend erfolgt eine Rechtschreibprüfung in der vorgewählten Sprache.

Natürlich könnte man Vorlagen auch einfach nur einscannen, um sie bei Bedarf auf dem Bildschirm zu lesen – analog zu der früher üblichen Archivierung auf Mikrofilm. Solche «digitalen Fotokopien» würden aber sehr viel Platz benötigen. Um einen Buchstaben digital als Bild zu speichern, braucht es mehrere Bildpunkte. Jeder einzelne davon beansprucht etwa gleich viel Speicherplatz wie ein eingetippter Buchstabe.



FOTO MARKUS ZITT

Alles, was Sie brauchen – vom Computer abgesehen – ist ein Scanner und ein OCR-Programm. Als Scanner eignet sich jeder einfache Flachbett-, Einzugs- oder Handscanner mit einer Auflösung von mindestens 300 ppi. Höhere Auflösungen sind vor allem für Vorlagen mit kleiner Schrift von Nutzen. Für die Texterkennung reicht ein einfacher Graustufen-Scanner, ja selbst ein reines Schwarzweiss-Modell mit einem Bit Farbtiefe kann genügen. Ein reibungsloses Arbeiten ist jedoch nur mit passenden Scanner-Treibern möglich. Deshalb gilt es beim Kauf von OCR-Software zu prüfen, welche Scanner vom Programm unterstützt werden.

Die günstigen, kleinen Handscanner eignen sich speziell für den Betrieb unterwegs mit dem Notebook. Wegen Verzerrungen beim manuellen Ziehen, eher mässiger Auflösung und geringer Farbtiefe, taugen reine Handscanner kaum fürs Digitalisieren von Bildern, sondern nur für OCR. Als Handikap erweist sich oft ihre geringe Scanbreite. Dadurch können A4-Blätter nur mit zwei horizontalen Scans eingelesen werden. In der Folge werden Zeilen doppelt gescannt oder ausgelassen. Umwege über verkleinerte Fotokopien sind aufwendig und wirken sich wegen winziger Buchstaben oft negativ auf die Texterkennung aus. Auch

Die Textvorlage wird mit dem Einzugsscanner als Bild eingelesen und innert Sekunden in digitalen Text umgewandelt.

für das Scannen von Büchern oder kleinen Vorlagen braucht es etwas Geschick, denn rasch bleibt der Scanner an den Seitenrändern hängen. Abhilfe schafft eine dünne Glasplatte auf dem Buch. Visitenkarten u.ä. werden vorzugsweise unter einer klaren Plastikfolie (Sichtmäppchen, Hellraumfolie) gescannt. Für Textpassagen eignen sich besonders die «Zitatensammler». Gemeint sind Handscanner, die wie ein Markierstift bedient werden und zeilenweise einlesen, z.B. der ScanPen (Info: Arp-Datcon AG, Tel. 041/799 09 09; Preis: Fr. 269.–).

Einzugsscanner sind besonders bedienungsfreundlich. Gestartet werden sie meist mit dem Einführen einer Papiervorlage. Dank ihrer kompakten Form lassen sie sich gleich neben der Tastatur platzieren und auch einfach transportieren. Meist «schlucken» sie aber nur grössere, lose Einzelblätter. Problematisch wird es mit dicken oder beschädigten Vorlagen sowie bei kleinsten Vorlagen wie Visitenkarten und Zeitungsausschnitte. Neuste Modelle, etwa der multifunktionale Scanner FlexoScan von Logitech, können selbst visitenkartengrosse Vorlagen einziehen. Auch bei der Auflösung bieten sie mittlerweile akzeptable

Fachchinesisch

OCR

Optical Character Recognition, kurz Texterkennung, bedeutet das Identifizieren einzelner Buchstaben anhand ihres spezifischen Aussehens.

Praxis

Texterkennung

Daten von etwa 300 bis 600 ppi und 24 bis 30 Bits Farbtiefe.

Flachbettscanner haben mit unterschiedlich grossen, dicken oder auch beschädigten Vorlagen keine Mühe. Sie sind jedoch am umständlichsten zu bedienen. Besonders nützlich für die Texterkennung sind Modelle, die mit einem Stapel einzug für Einzelblätter ausgestattet werden können. Damit muss man nicht untätig dem Computer bei der Texterkennung zusehen, um den Scanner immer wieder mit neuen Vorlagen zu füttern.

OCR-Programme können einzeln erworben werden. Eine einfache Version wird häufig mit Scanner oder Grafiksoftware wie der CorelDraw Suite mitgeliefert. Vermehrt verfügen auch Faxprogramme über eine integrierte OCR-Funktion. Langjähriger Marktführer für OCR-Software ist die Caere Corporation, die sich ihre Position durch konsequente Weiterentwicklung ihrer Produkte und den Aufbau

fache Bildbearbeitungsfunktion. Der Schnelligkeit wegen sollten sich alle Arbeitsschritte auch automatisch durchführen lassen. Nützlich ist die Möglichkeit, das OCR-Programm mit Mustervorlagen zu trainieren. Werden allerdings immer wieder Vorlagen mit unterschiedlichen Schriften verwendet, lohnt sich das kaum.

Allgemein konzentrieren sich Entwickler von OCR-Programmen auf die Optimierung des In- und Outputs. Einerseits bereitet die OCR-Software die Scans für die Texterkennung auf. So weisen auch schlechte Vorlagen, wie Ausdrucke von Faxgeräten oder Nadeldruckern, graue Fotokopien oder schief eingescannte Texte, geringe OCR-Fehlerraten auf. Andererseits wird der erkannte Text möglichst ähnlich dargestellt, d.h., OCR-Programme versuchen die Seitengestaltung zu übernehmen. Dieser Fähigkeit rühmt sich das nur für Windows 95 und NT erhältliche Profi-Programm

gierende Fehler pro Textzeile macht den Zeitvorsprung zunichte.

Häufigste Ursachen für fehlerreiche Ergebnisse sind mangelhafte Vorlagen oder schlechte Scans. Oft führt der Umweg über ein Bildbearbeitungsprogramm zu besseren Erkennungsraten. Ein schiefer Scan kann dort gerade gestellt werden, Beschädigungen des Papiers retuschiert und das gescannte Textbild nötigenfalls geschärft werden. Empfehlenswert ist die gezielte Kontraststeuerung, bis die Schrift am leserlichsten erscheint. Zum Retuschieren sollte die Vorlagen als Graustufenbilder vorliegen. Da ältere und einfachere OCR-Programme nur reine Schwarzweissbilder analysieren können, müssen für diese die Graustufenbilder am Ende der Bearbeitung umgewandelt werden.

An einem Text auf farbigem Hintergrund, sei es eine einfarbig unterlegte Textbox oder ein Hintergrundfoto, scheitern heute noch alle OCR-Programme. Hier hilft folgender Profi-Tip: Mit einem Farbscanner und einem Bildbearbeitungsprogramm, das die einzelnen Farbkanaäle anzeigt, beispielsweise PhotoShop, PhotoPaint oder LivePicture, lassen sich die störenden Farben eliminieren. Der dreifarbig Scan im RGB-Farbmodell wird in ein Vierfarbenbild umgewandelt, bestehend aus den Ebenen cyan, magenta, yellow und schwarz. Dann werden die CMY-Farbkanaäle gelöscht, der Bildkontrast im übrig gebliebenen schwarzen Kanal gesteigert und das Pixelbild als Graustufenbild gesichert. Aber es geht auch ohne Bildbearbeitungsprogramme: Vorlagen können absichtlich schief in den Scanner eingelegt werden, um ein verzogenes Textbild zu kompensieren. Visitenkarten und kleine Zeitungsausschnitte oder beschädigte Vorlagen werden erst fotokopiert und dabei am besten noch vergrössert.

Wenn immer möglich sollten jedoch Originalvorlagen eingescannt werden. Werden Zeitungsartikel erfasst, kann man hinter das Blatt einen schwarzen Halbkarton legen, um das Durchscheinen des rückseitigen Drucks zu verhindern.

Möchte man Faxbriefe in editierbaren Text umwandeln, so empfiehlt sich doppelt die Anschaffung eines Faxprogrammes mit OCR. Man spart nicht nur Faxpapier und Scanzeiten; vor allem erzielt man mit einem elektronisch empfangenen Fax eine bessere Erkennung, da der qualitätsmindernde Umweg über den Scanner wegfällt. Zudem sind die integrierten OCR-Funktionen auf Faxvorlagen spezialisiert. Der Erfolg ist jedoch durch die niedrige Auflösung der Faxstandards selten befriedigend. Bei bevorstehendem Faxversand bittet man den Absender vorzugsweise eine serifenlose Schrift, wie z.B. Avantgarde, in mindestens zwölf Punkt Grösse zu verwenden und das Fax in der besseren Fine-Auflösung – aber keinesfalls in der Fotoauflösung – zu senden. Bleibt zu hoffen, dass das Fax bald endgültig von der effizienteren E-Mail verdrängt werden wird.

Einen Nachteil hat auch die beste Texterkennung: Der Lerneffekt, den viele beim Abtippen oder auch beim verkürzten Zusammenfassen erfahren, wird künftig ausbleiben. ●

Flachbettscanner eignen sich, um Textvorlagen zu erfassen. Sie sind vorteilhaft für dickere oder beschädigte Originale, lohnen sich aber nur, wenn sie auch für andere Aufgaben als OCR genutzt werden.



Typischer Zitatenfresser: Solche Handscanner, die wie ein Markierstift bedient werden, sind praktisch für kurze Texte, beispielsweise Adressen auf Visitenkarten.



FOTOS: PC WORLD

von Konkurrenzfirmen sichert. Caeres OmniPage Pro, das Ende Jahr als Version 8.0 nun auch für Windows NT 4 und – wie bis anhin – für MacOS und Win 95 erhältlich sein wird, ist die leistungstärkste, aber auch teuerste Software. Sie bietet neu automatische Erkennung von Sprachen, Integration in Microsofts Word (direktes Einlesen in Word-Dateien, gemeinsame Nutzung der Wörterbücher) sowie eine verbesserte Erkennung selbst bei schlechten Vorlagen. Immerhin lassen sich die abgespeckten Bundle-Varianten OmniPage LimitedEdition und OmniPage Lite relativ günstig auf die Vollversion aufrüsten. Info: Dative, Tel. 071/637 70 70. Preise: Vollversion Fr. 1333.–, Update auf Version 8.0 Fr. 203.–, Update Bundle-Varianten Fr. 320.– (bis Ende 97). Der stärkste Konkurrent von OmniPage ist TextBridge von Xerox, das wegen seines günstigen Preises (Info: CCS AG, Tel. 01/709 09 90, Preis Fr. 373.–) bei ausgezeichnete Leistung viele Scanner gleich als Vollversion begleitet.

Qualitativ unterscheiden sich die verbreitetsten OCR-Programme nur wenig, denn die Erkennungstechnologie stagnierte in den letzten Jahren auf einem hohen Niveau. Unterschiede finden sich deshalb eher in der Ausstattung der Programme. Ein OCR-Programm sollte sowohl vorhandene Scans bearbeiten können, als auch das direkte Einlesen vom Scanner unterstützen. Elementare Einstellmöglichkeiten sind die Vorwahl einer bestimmten Sprache und die Auswahl der zu erkennenden Bereiche auf der gescannten Vorlage. Ebenfalls zum Standard gehört eine ein-

Acrobat Capture von Adobe (Info: Adobe, Tel. 0800 55 51 54, Preis: kleine Version ca. Fr. 7000.–, grosse ca. Fr. 23'000.–). In der Klasse der erschwinglichen OCR-Programme indes ist noch einige Entwicklungsarbeit zu leisten. Zum Korrekturlesen ist die Beibehaltung des Seitenlayouts eine Hilfe. Sie erlaubt eine rasche Orientierung beim Vergleich zwischen Vorlagen und Textdatei.

Die Tücken der Texterkennung lassen sich mit Tricks umschiffen. Trotzdem macht der ganze Aufwand vom Scannen über OCR bis hin zur persönlichen Kontrolle erst ab einer gewissen Textlänge wirklich Sinn. Umgekehrt ist Texterkennung unschlagbar, wenn sich die Prozedur vom Einscannen und Erkennen dank Stapel einzug und Batchverarbeitung automatisieren lässt. In jedem Fall erweist sich eine Texterkennung als ineffizient, wenn der Text intensiv nachbearbeitet werden muss. Bereits zwei manuell zu korri-