



Entwirrter Zeichensalat

Lettern, Linien, Fliegendreck – im PC landet alles, was der Scanner digitalisiert hat. Programme zur optischen Zeichenerkennung machen aus dem Buchstaben-salat wieder lesbare Texte. CHIP zeigt, worauf es bei OCR ankommt.

Kleine Kinder, die gerade lesen und schreiben lernen, haben Computern viel voraus. Sie haben bereits gelernt, sich mit Sprache zu verständigen, durch Sätze ihre Wünsche mitzuteilen und können Gegenständen Bezeichnungen zuordnen. Der Blechkasten hingegen weiß nichts von Wörtern, grammatikalischen Regeln oder gar von der Bedeutung einzelner Sätze.

Genau dies macht es ihm so schwer, die Zeichen, die er lesen und umwandeln soll, richtig zu deuten. Wie soll er wissen, ob in den Zusammenhang eher eine Hose oder ein Hase paßt? Um das zu deuten, müßte er wissen, was eine Hose oder ein Hase ist.

Mit solchen Schwierigkeiten haben Hersteller von Software für optische Zeichenerkennung (OCR) zu kämpfen. Die

Erkennungsrate haben sie zwar schon auf über 99 Prozent getrieben, doch bedeutet das, daß auf einer Schreibmaschinenseite noch etwa 40 Buchstaben falsch erkannt werden.

Um die Rate weiter nach oben zu treiben, arbeiten die Programme mehrstufig. Nach der Analyse des Aufbaus einer Seite läuft die eigentliche Zeichenerkennung: Die Software liest die einzelnen Buchstaben. Neben der Überdeckungsanalyse – die nach dem intern gespeicherten Zeichen mit der größten Überdeckung mit dem Bild sucht – kommt die Topologieerkennung zum Einsatz: Dabei interpretiert die Software die Zahl der Linien eines Buchstabens und deren Verlauf. Vier Linien, wobei drei an der größeren vierten hängen, stellen mit ziemlicher Sicherheit ein „E“ dar.



Lehning

Abschließend untersucht die Software, ob die Wörter, die sie erkannt zu haben meint, in ihrem Wörterbuch zu finden sind. Hier kommt die Sprache ins Spiel: Die Software muß auf die Sprache des Dokuments eingestellt sein. Programme wie *Recognita Plus* verstehen sich auf eine ganze Litanei an Sprachen.

Omnipage oder *Recognita* gehen aber noch weiter: Sie rekonstruieren den Aufbau eines Dokuments. Als Ergebnis erhält man eine Rich-Text-Datei, die alle Bilder und Textkästen enthält und das gleiche Layout aufweist wie das Original.

Um die Erkennungsrate noch mehr nach oben zu drücken, wäre ein weiterer Schritt wünschenswert: die Analyse nach Inhalt und Grammatik und eine kleine Portion Allgemeinbildung. Das würde die Hose vom Hasen trennen, zur richtigen Einordnung eines Binde- und eines Trennstriches führen oder in einer Telefonnummer Buchstaben verbieten.

Ebenfalls ein Punkt der Wunschliste wäre eine Option, die dem Programm für lange Zeit die komplette Rechenleistung des PC zur Verfügung stellt. Auf diese Weise könnte die Software wichtige Do-

kumente wie Zahlentabellen in vielen Stunden, dafür aber mit maximaler Sicherheit in Text umwandeln.

Zwei Regeln helfen, optimale Erkennung zu gewährleisten. Eine höhere Auflösung bringt nicht unbedingt ein besseres Ergebnis. Das Scannen mit 300 dpi hat sich als guter Wert herausgestellt.

Beim Scannen sollte man darauf achten, daß die Seite wenig verschmutzt ist. Schlechte Kopien weisen meist schwarze Sprenkel auf, die der OCR-Software das Leben ungeheuer schwer machen.

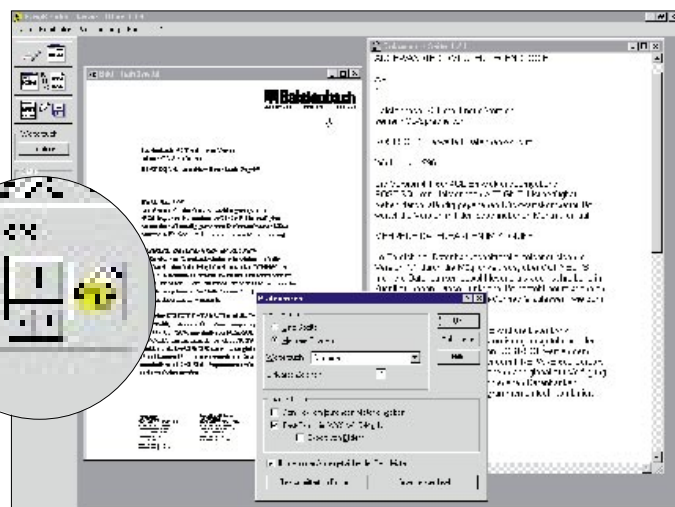
○ Easyreader Classic

Freunde von kleinen Programmen sind mit *Easyreader Classic* gut beraten. Lediglich zwei Disketten umfaßt die Software. Auch bei der Bedienung zeigt sich die Windows-95-Software von der schlanken Seite. Die wenigen Knöpfe sorgen für Übersichtlichkeit. Im Prinzip wird die Software über drei Schaltflächen bedient: Man liest ein Bild über den Scanner ein, startet den OCR-Vorgang oder speichert den erkannten Text. Ein Knopf, der alles auf einmal ausführt, fehlt allerdings.

Als besonderes Merkmal weist die Software eine Kopierfunktion auf. Mit Hilfe dieser Funktion setzt man das Programm zusammen mit Scanner und Drucker wie ein Kopiergerät ein: Original in den Scanner legen und den Startknopf drücken. Die Software holt dann das Bild in den PC und schickt es ohne Texterkennung an den Drucker. Über einen weiteren Knopf startet der Anwender die Faxfunktion. Ähnlich wie die Kopierfunktion holt die Software das Bild vom Scanner und schickt es an den Faxtreiber eines Modems. Gerade für kleine Büros sind das sicherlich zwei hilfreiche Funktionen.

Im Modus Bruttotext erkennt Easyreader Classic den reinen Text. In diesem Modus gehen alle Formatierungen verloren. Im Modus Wysiwyg (What you see is what you get) erhält die Software auch die Auszeichnungen wie Fett- oder Kursivdruck. Dieser Modus funktioniert nur über Drag and Drop in Zusammenarbeit mit einer Textverarbeitung, die Rich Text darstellen kann. Die Möglichkeit, den Text als Rich-Text-Datei oder als Word-Dokument zu speichern, bietet Easyreader nicht.

Bei der Bedienung fällt negativ auf, daß der Anwender nicht in der Lage ist, die Vergrößerung des gerade bearbeiteten Textbildes zu wählen. Lediglich eine Lupenfunktion gestattet einen detaillierteren Blick auf das gescannte Bild: In



Als Bonus eine Kopierfunktion: Ein Knopfdruck schickt die Daten vom Scanner an den Drucker. Der interne Editor von *Easyreader* zeigt nur Text an, das Layout beachtet er nicht.

einem Extrafenster zeigt die Software einen stark vergrößerten Ausschnitt des Bildes, den man über die Position des Mauszeigers verschieben kann.

Bei der Erkennungsrate ist Easyreader nicht eindeutig einzuordnen. So ging das Programm bei der einfach aufgebauten Testseite mit nur fünf Fehlern als zweitbestes durchs Ziel. Dafür hatte es aber bei komplexeren Seiten mit Bildern und mehreren Spalten seine Probleme. Es trennte beispielsweise sehr enge Spalten, wie sie im vierspaltigen Satz vorkommen, nicht voneinander.

Abhilfe schafft das manuelle Trennen der Spalten: Mit der Maus legt man einen Rahmen um eine Spalte und zieht den Rahmen per Drag and Drop in ein Textverarbeitungsprogramm wie beispielsweise Winword oder Wordpad. Die Software analysiert dann nur den Inhalt des Rahmens und schickt den Text an die Textverarbeitung. Auf diese etwas aufwendige Art und Weise stückelt man aber auch komplexe Seiten zusammen.

welcher Qualität das Dokument ist, ob das Dokument aus mehreren Spalten besteht und wie Textbridge mit dem Layout verfahren soll. Wählt man die Option »Alles neu zusammenstellen«, versucht die Software, das Layout der Seite zu rekonstruieren, zumindest was die Breite der Spalten betrifft. Aber auch etwaige Bilder legt die Software in der Ausgabedatei im Rich-Text-Format ab.

Um die Erkennungsrate zu erhöhen, bietet die Software einen interaktiven Trainingsmodus an. Das heißt, daß Textbridge das erste Wort, bei dem es sich nicht sicher ist, anzeigt und dem Anwender die Chance gibt, den Ausdruck zu korrigieren.

Unangenehm fällt auf, daß man nach dem Laden eines Dokumentenbildes keine Möglichkeit mehr hat, die Optionen, mit denen Textbridge das Dokument verarbeitet, zu verändern. Die Software macht schon während des Einlesens eine Grobanalyse, die von den gewählten Parametern abhängt. Will der Anwender

Schnell ein Auge drauf werfen

Die Erkennungszeiten hängen von der Beschaffenheit des zu lesenden Dokuments ab. So brauchte Recognita etwa für einen einspaltigen Text um die zehn, für ein vierspaltiges Dokument mit Bildern und weniger Text über 14 Sekunden (Pentium-PC, 166 Megahertz, 32 Megabyte RAM).

Auch die Erkennungsrate ist stark an die Beschaffenheit des Dokuments gekoppelt. Unsere einfache Testseite bestand aus einspaltigem Text (1964 Buchstaben) in Schriften wie Times New Roman, Courier New und Arial in Größen von 10 bis 18 Punkt und Schnitten wie normal, fett und kursiv.

Klickt der Anwender anschließend auf den Knopf »Go«, verarbeitet Textbridge nur die umschlossenen Bereiche.

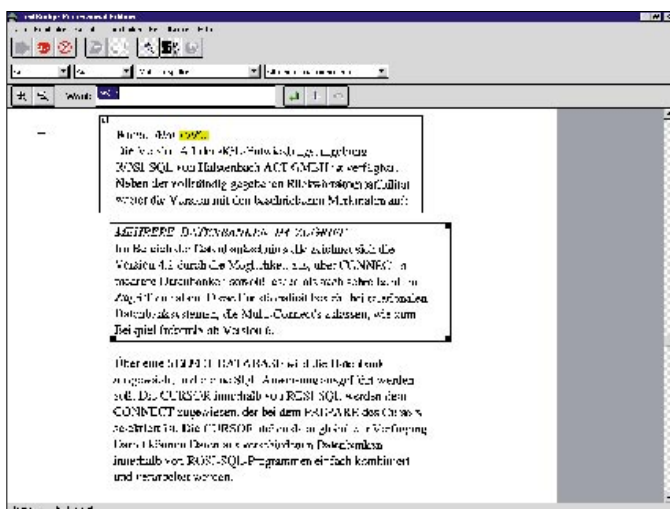
Die neue Version für Windows 95, die uns zum Test in einer englischen Version vorlag, unterscheidet sich von der Oberfläche und der Bedienung unwesentlich von der Windows-Version. Vor allem ist sie nicht schneller.

○ Omnipage

Caere hat mit *Omnipage* die Versionsnummer 7 erreicht. Diese Windows-95-Software lag zum Test noch in einer Beta-version vor. Aus diesem Grund bezieht sich der Test auf den Vorgänger, die Version 6.0 für Windows. Die Software gehört wie Textbridge und Recognita zu den dickeren Paketen. Nachdem man sechs Disketten in den Laufwerksschacht geschoben und das Programm gestartet hat, präsentiert sich eine recht aufgeräumte Oberfläche. Wie bei Easyreader dominieren die wesentlichen Schaltflächen: eine Schaltfläche, um eine Datei zu laden, eine zweite für den OCR-Vorgang und eine, um den Text auf der Festplatte abzulegen. Die größte dient der ökonomischen Einknopfbedienung: Sie startet den gesamten Ablauf.

Nach der Texterkennung zeigt Omnipage sowohl das Textbild als auch den erkannten Text in zwei Fenstern an. Dadurch ist ein einfacher, unmittelbarer Vergleich zwischen Original und Abbild möglich. Ausdrücke, bei denen sich das Programm nicht sicher war, sind in anderer Farbe gekennzeichnet.

Omnipage versucht das Layout des Dokuments originalgetreu nachzubilden.



Kastenweise: Einzelne Rahmen sagen Textbridge, welche Bereiche es bearbeiten soll. Im Trainingsmodus kennzeichnet es unsichere Wörter farbig.

○ Textbridge Professional 3.02

Zum Portfolio der Firma Xerox gehört die OCR-Software *Textbridge*. Mit sieben Disketten ist das Programm eines der dickeren Produkte. Doch trotz des Mehr an Bits und Bytes erscheint die Oberfläche angenehm aufgeräumt. Eine Menü- und zwei Symbolleisten sind alles, was der Anwender nach dem Start zu sehen bekommt. Über den großen Knopf »Go« startet man den ganzen Ablauf: Scannen oder Datei öffnen, Erkennen und Text speichern.

Unterhalb dieser Leiste geben mehrere Listboxen die Möglichkeit, die Einstellungen für das zu erkennende Dokument zu verändern. Hier teilt der Anwender beispielsweise dem Programm mit, von

nach dem Einlesen eine Option verändern, bleibt ihm nur übrig, den Vorgang zu stoppen, die Einstellung zu verändern und das Einlesen erneut zu starten.

Bei der Erkennungsrate lag Textbridge bei der einfachen Testseite mit zehn Fehlern an 3. Stelle. Textbridge war das einzige Programm, das die Seite bei höherer Auflösung besser erkannte. Bei komplexen Seiten dauern das Einlesen und Erkennen recht lange. Schade ist, daß man den Ergebnistext nur speichern oder in die Zwischenablage kopieren kann, ihn also innerhalb des Programms nicht zu Gesicht bekommt. Dafür hat Xerox eine Vorschaufunktion auf das eingelesene Bild eingebaut. In diesem Modus lassen sich auch einzelne Rahmen aufziehen, die den zu erkennenden Text kennzeichnen.

Allerdings verkettet die Software die einzelnen Textkästen nicht miteinander. Zwar durchwandert der Cursor den Text in der logischen Schlangenbewegung der Spalten, wenn man ihn über die Nach-unten-Taste durch den Text bewegt; doch rutscht der Text aus dem nächsten Kasten nicht nach, wenn man eine Zeile löscht.

In der Version 7.0 für Windows 95 hat Caere auch dieses Merkmal in die Software eingebaut. Dann hängen die Texte in den einzelnen Kästen aneinander, und der Textfluß gehorcht dem Lesefluß.

Sehr hilfreich ist die Funktion »Automatische Bereiche«. Wählt man diesen Menüpunkt an, teilt die Software das Dokument in die groben Elemente auf. Will der Anwender nun bestimmte Bereiche des Dokuments von der Erkennung ausschließen, genügt ein Klick auf den jeweiligen Rahmen. Die Software ändert daraufhin die Farbe des Rahmens und überspringt ihn bei der Texterkennung.

Die fertigen Texte legt Omnipage in verschiedenen Formaten wie Winword oder Ami Pro ab. Auf Wunsch speichert es aber auch alle Daten, wie erkannten Text, das ursprüngliche Textbild und Rahmen, in einer eigenen Datei.

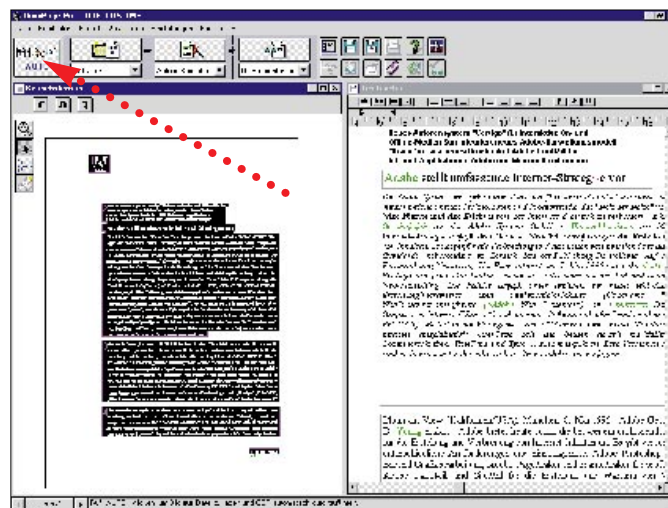
Wie Textbridge bietet die Omnipage-Software auch einen Trainingsmodus an. Nach der Eingabe der richtigen Zeichen kann die Datei für künftige Dokumente auf der Festplatte abgelegt werden.

Omnipage hatte mit der einfachen Testseite die meisten Probleme: Elf Fehler fanden sich im Text. Allerdings kam es mit komplexeren Dokumenten gut zurecht.

o Recognita Plus 3.0

Das Programm der ungarischen Firma Recognita gehört zu den Klassikern des europäischen OCR-Marktes. Auf sieben Disketten findet das Programm Platz. Die Oberfläche beherbergt nur wenige Schaltflächen. Wie bei Omnipage gibt es bei Recognita einen zentralen Knopf, der den gesamten Erkennungsvorgang startet. Nach dem Laden des Textbildes teilt die Software das Dokument grob in einzelne Bereiche ein. Dabei zeigt es auf Wunsch durch Pfeile an, wie seiner Meinung nach der Text fließt.

Bei problematischen Seiten kann der Anwender das Programm veranlassen, eine Zwangspause einzulegen. Damit erhält er die Gelegenheit, selbst Hand an die Rahmenzahl und -größe zu legen. Bei sehr engem Spaltenzwischenraum zwingt er Recognita auf diese Weise etwa, die Spalten dennoch zu trennen und sie nicht horizontal zu verkettet oder sie als



Direkter Vergleich: Omnipage zeigt in zwei Fenstern das Original und den erkannten Text an. Die große Schaltfläche startet den gesamten OCR-Ablauf (Pfeil).

Pfeilgerade: Den erkannten Textfluß zeigt Recognita auf Wunsch durch Pfeile an (rechte Seite). Der horizontale Rollbalken bedient beide Fenster (Pfeil).

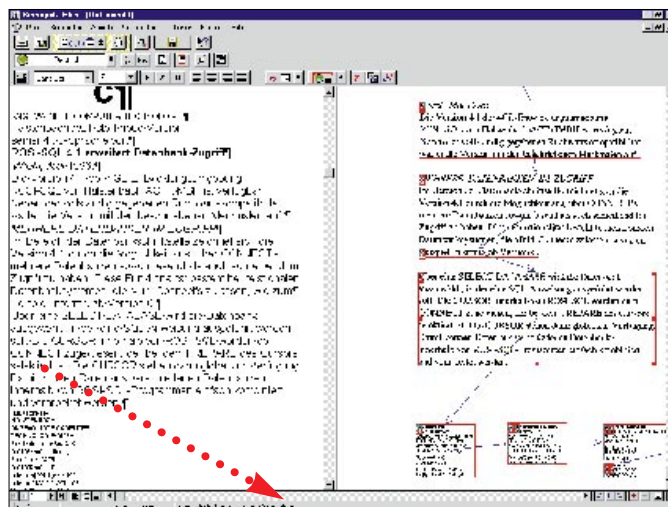


Tabelle anzusehen. In letzterem Fall trennt das Programm die Spalten mit einem Tabulator.

Den Inhalt der einzelnen Rahmen kann der Anwender mittels Schaltflächen auszeichnen. So bringt man Recognita sehr schnell bei, daß ein Rahmen keineswegs eine Tabelle, sondern Fließtext enthält. Hier stehen neben Fließtext, Grafik und Tabelle auch Balkencode, binäre Ankreuzfelder und handschriftliche Zeichen zur Verfügung.

Hat die Software die Erkennung beendet, stellt sie sowohl das Ausgangsbild als auch das Ergebnis in einem horizontal oder vertikal geteilten Fenster dar. Die Bedienung dieses Fensters ist etwas gewöhnungsbedürftig, da jeweils ein Rollbalken für beide Fensterhälften zuständig ist. Je nachdem, welche Hälfte aktiv ist, verschiebt der Rollbalken die eine oder andere Hälfte.

Bei der Darstellung hat der Anwender die Wahl zwischen Modi von reinem Text bis zum komplett nachgebauten Dokument. Recognita verpackt die einzelnen

Rahmeninhalte in Positionsrahmen von Winword. Dabei sind die Rahmeninhalte voneinander getrennt, ein Nachrutschen der Zeilen aus anderen Kästen funktioniert also nicht.

Bei der Erkennungsrate gehört Recognita zu den Besten. Bei der Testseite machte es nur zwei Fehler, ging somit als Sieger durch das Ziel. Auch die Rekonstruktion des Layouts im Textdokument funktioniert recht gut. **Tilman Börner**



Anbieteradressen: Easyreader Classic: Jolo Data, Goschenstr. 37, 31134 Hildesheim, Tel.: 05121/37555, Fax: 05121/37556, Preis: ca. 250 Mark, ca. 900 Mark (Elite)

Textbridge Pro: CCS Vertrieb GmbH, Hauptstr. 25, 73087 Bad Boll, Tel.: 07164/9419-0, Fax: 07164/9419-11, Preis: ca. 1000 Mark
Omnipage: Caere, Innere Wiener Str. 5, 81667 München, Tel.: 089/458735-0, Fax: 089/458735-20, Preis: ca. 1200 Mark
Recognita Plus: Recognita Büroautomatisierung, Lützowstr. 11a, 04155 Leipzig, Tel.: 0341/589060, Fax: 0341/5890641, Preis: ca. 1200 Mark