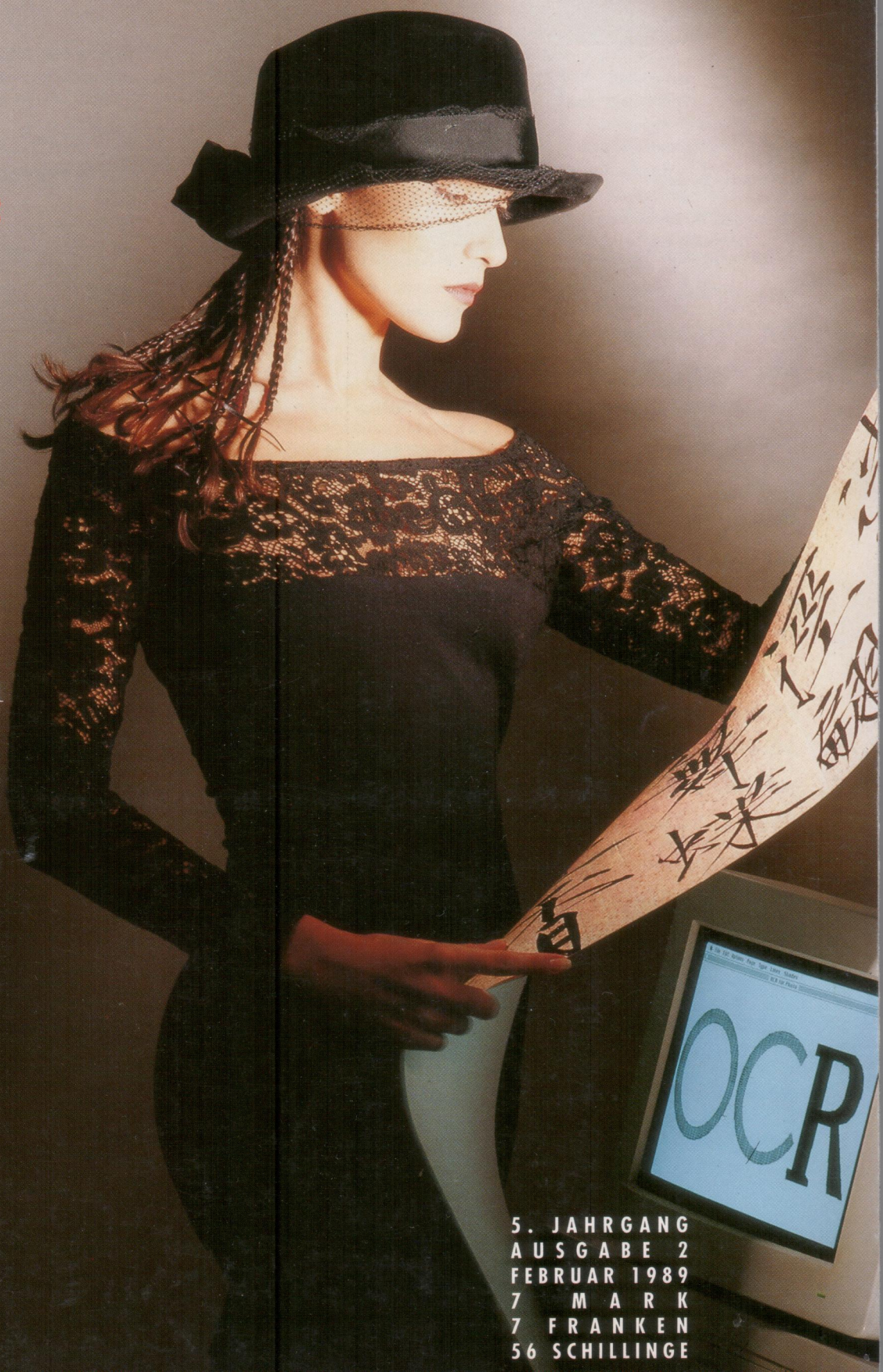
MACUP
DAS UNABHÄNGIGE
MACINTOSH-MAGAZIN

MACup

Optical Character Recognition:
SCHRIFTERKENNUNG MIT DEM MACINTOSH



Programme:
Digital Darkroom
Full Impact
Quark Style

Grappler LQ:
Preiswerter Drucken

Tutorial:
Daten retten

5. JAHRGANG
AUSGABE 2
FEBRUAR 1989
7 MARK
7 FRANKEN
56 SCHILLINGE

LESEN UND LESEN

FRIEDRUN REINHOLD

Computer können nicht lesen. Texte, die ihnen nicht eingegeben wurden, behandeln sie als sinnlose Punktsammlungen – bestenfalls als Graphik. Diesen Mangel an Verständnis soll die Technik der „Optical Character Recognition", kurz OCR, beheben. MACup-Redakteur Benjamin Heidersberger beschreibt die Verfahren und präsentiert Ergebnisse.

LASSEN

Manchmal ist es zum Verzweifeln. Da sitzt man nun vor dem teuren Equipment – PC und Scanner –, und dann soll man diesen Text hier eintippen wie damals in der ersten Klasse. Wo bleibt denn da der Fortschritt! Es muß doch möglich sein, diese Maschine dazu zu bringen, den Text abzuscannen und in eine Textverarbeitung zu übersetzen.

Wer sich diese oder ähnliche Gedanken schon mal machen mußte, ist der ideale Kandidat für eine Art Software, die dem Computer etwas gibt, was fast schon intelligent ist, nämlich OCR. OCR bedeutet „Optical Character Recognition", also das „Lesen" eines Textes, die Übersetzung der gescannten Text-Graphik in einzelne Buchstaben, um von einer für den Computer sinnlosen Sammlung von Punkten zu einem Text zu kommen.

Multimedia

Am Anfang war das Wort. Eingetippt auf einer Tastatur und dargestellt auf einem Bildschirm. Wie langweilig. Je älter der PC wird, desto mehr Sinne bekommt er, Mitteilungen aufzunehmen und sich mitzuteilen. Vorläufiges Ziel scheint der hörende und sehende, der sprechende und animierte Computer zu sein.

Wichtigstes Hindernis auf diesem Weg ist das Verstehen. Es ist nicht schwierig, wie ein Tonband etwas aufzunehmen oder wie ein Photo zu speichern. Problematisch ist die Umsetzung in Objekte, den abgetasteten und gespeicherten Bildpunkten also beispielsweise den Buchstaben „a" zuzuordnen. Hier erst fängt der Computer an, eine wichtige Leistung des Gehirns nachzuahmen.

Was macht Sinn?

Ziel des „papierlosen Büros" ist es, alle Texte elektronisch vorliegen zu haben und bearbeiten zu können. Theoretisch also ist OCR nur für eine Übergangszeit nötig, bis alles elektronisch erfaßt ist und übermittelt wird.

Praktisch ist dieser Traum schon lange ausgeträumt. Der Mensch benötigt zum Arbeiten eine schriftliche, gedruckte Vorlage. Von der Ausschaltung der sinnlich-taktilen Komponente mal abgesehen, wäre auch ein weltweit genormtes Kommunikationsformat vonnöten. Ganz zu schweigen von den interessanten Problemen, nicht fälschbare Dokumente elektronisch zu erzeugen und zu speichern – ein Aspekt, der die Computerkriminalität aufblühen ließe. Und schließlich – mal ganz banal – braucht jedes in diesem Format erzeugte Dokument Hardware, um abgespielt zu werden. Da sind die Software und Abspieler vereinenden Bücher oder Briefe doch überlegen.

Dennoch macht OCR Sinn. Denn zwar liegt weder alles Wissen dieser Welt elektronisch vor, noch wird das jemals der Fall sein; an der Schnittstelle von bedrucktem Papier zum Computer aber findet OCR ein dankbares Feld.

Beispielsweise werden bisher beim Aufbau eines Archivs von wissenschaftlichen Magazinen nur bestimmte Stichworte oder eine Kurzzusammenfassung in die Datenbank aufgenommen. Besser dagegen sind Volltextdatenbanken, die vollständige Texte erfassen, in denen gesucht werden kann. Diese Erfassung wird leichter mit OCR.

In vielen anderen Gebieten wäre OCR dagegen nur ein Umweg einer ohnehin nötigen Neustrukturierung. Bei der klassischen Zeitungsherstellung tippt der Journalist seinen Text auf der Schreibmaschine, das Korrektorat prüft die Rechtschreibung, tippt den Text gegebenfalls neu ab, und schließlich setzt ihn der Setzer nochmals auf der Setzmaschine. Heute dagegen wird der elektronisch erfaßte Text nur noch von Station zu Station wieder aufgerufen und bearbeitet – sogar die Meldungen der Nachrichtenagenturen gelangen bereits auf elektronischem Wege zur Zeitung.

Die Benutzung des Telefax führt zu weiteren interessanten Aspekten. Man tippt den Brief auf dem Computer, druckt ihn aus, geht zum Fax, das scannt ihn, komprimiert den Inhalt und übermittelt denselben per Modem. Auf der Gegenseite erfolgt der umgekehrte Vorgang bis hin zum Erfassen des Textes durch OCR. Hier wäre eine direkte Übertragung per Modem von Computer zu Computer bei weitem sinnvoller.

Ein zukünftiges Telefax wird imstande sein, eine handgeschriebene Seite per OCR zu übersetzen und somit auch optimal zu komprimieren, da die Seite nicht mehr als Graphik, sondern bereits als Text übertragen wird.

Die Technik

Auf den ersten Blick scheint das Erkennen eines Buchstabens eine recht einfache Aufgabe zu sein. Dabei vergißt man leicht, daß dem Menschen im Alltag eine Hochleistungsrechenmaschine zur Verfügung steht, nämlich das menschliche Hirn in Zusammenarbeit mit den informationsverarbeitenden Eigenschaften des Auges. Diese Kombination nachzuahmen ist für einen einfach strukturierten Automaten nicht möglich. Außerdem versteht der Mensch nur unvollkommen sein eigenes Funktionieren. So zeigen Untersuchungen, daß Wörter schneller erkannt werden als Einzelbuchstaben, eine Tatsache, die mit einer rechnertypischen linear-vertikalen Arbeitsweise, die vom Einzelnen ausgehend immer mehr begreift, nicht in Einklang gebracht werden kann.

Ein einfaches Beispiel zeigt die besonderen Fähigkeiten der Informationsverarbeitung beim Menschen, die es erlaubt, auch noch halb verstümmelte Zeilen zu lesen:

el an die vorgesehene | alles parat – auch für einen typografisch
ht. Er spart den Platz | Amoklauf. Aber bislang summierten s

Bei der maschinellen Zeichenerkennung lassen sich grundsätzlich zwei verschiedene Methoden unterscheiden. Das „Pattern matching" versucht, den gescannten Buchstaben mit einem vorgegebenen, bereits abgespeicherten Muster des idealen Buchstabens zur Deckung zu bringen und so zu erkennen. Betrachtet man jedoch vergrößerte Buchstaben genauer, so ergeben sich verschiedene Probleme.

Zum einen ist die Kontur an etlichen Stellen ausgefranst. Dem läßt sich mit Rechenmethoden beikommen, die diese Konturen begradigen und sie so dem Ideal näherbringen. Dann aber können leicht ineinanderlaufende Buchstabenfolgen zu einem unverständlichen Zeichen kombiniert oder Einschnürungen zum vollständigen Zerfallen gebracht werden.

Untersucht man in einem Text alle Exemplare eines Buchstabens, wird man eine gewisse Variationsbreite feststellen. Diese Toleranz – eine sehr wichtige und kritische Größe – muß ebenfalls der Idealbuchstabe haben. Eine zu große Toleranz erkennt nämlich auch andere Buchstaben als die gewünschten, eine zu geringe Toleranz dagegen nicht alle Buchstaben einer Sorte. Und schließlich müssen noch Drehungen, Verzerrungen und Größenunterschiede berücksichtigt werden, indem die gescannten Buchstaben in eine Normalform überführt werden.

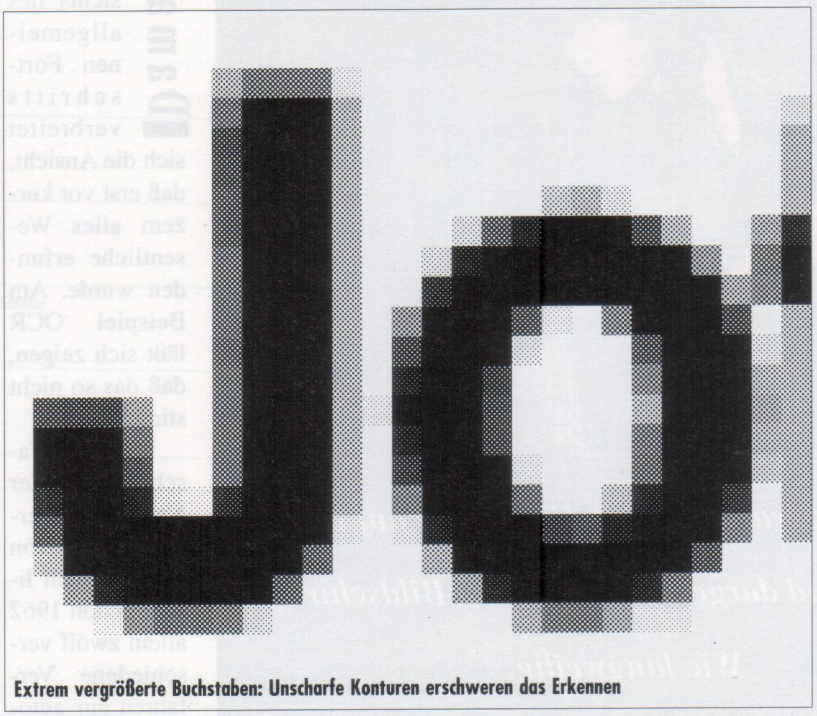
Der andere Ansatz heißt „Feature recognition" und bezeichnet die Erkennung der Eigenschaften eines Buchstabens. Dazu wird ein Algorithmus programmiert, der versucht, den Konturen des zu erkennenden Buchstabens „nachzufahren" und die Eigenschaften abzutasten. Beispielsweise ist ein „d" ein Kreis, an dem auf der rechten Seite eine Gerade tangential nach oben angesetzt ist. So lassen sich für alle Buchstaben Beschreibungsregeln finden.

Viele Probleme des Pattern matching, etwa die Überführung in eine Normalform, sind damit behoben – an ihrer Stelle tauchen andere auf. Zum einen haben bestimmte Buchstaben in einigen Fonts ein verschiedenes Erscheinungsbild, das „a" beispielsweise; zum anderen ist die Unterscheidung von kleinem „l", großem „I" und der Zahl „1" über Eigenschaften fast nicht möglich.

Vergleich

Vergleicht man Pattern matching und Feature recognition, so scheint erst einmal die letztere Methode überlegen zu sein, da die stärkere Abstraktion Varianten besser ausgleicht. Dennoch kann eine Beschreibung immer nur auf einen Zeichensatz optimiert sein. Umfaßt der zu verarbeitende Text mehrere Zeichensätze, kann das Pattern matching zu besseren Ergebnissen führen, da die Muster für verschiedene Zeichensätze parallel geladen sein können.

Ein weiterer wichtiger Vorteil des Pattern matching ist die Möglichkeit, das System zunächst zu trainieren und es mit den Buchstaben bekanntzumachen. Bei der Feature recognition dagegen müßten die zu abstrahierenden Eigenschaften eines Buchstabens aus seiner äußeren Erscheinung isoliert werden, was als rekursiver Prozeß ad absurdum führt, weil genau das OCR ist.



Extrem vergrößerte Buchstaben: Unscharfe Konturen erschweren das Erkennen

Teil des Ganzen

Nun besteht die Aufgabe von Optical Character Recognition nicht lediglich darin, einzelne Buchstaben zu erkennen. Das Ziel soll sein, normale Seiten möglichst vollständig in eine Textverarbeitung zu übersetzen. Und normale Seiten haben oft zur Verzweiflung führende Eigenschaften, nämlich verschiedene Varianten eines Schrifttyps, wie beispielsweise Fett und Kursiv, Versalien oder Unterstreichungen. Möglicherweise werden zusätzlich noch verschiedene Schrifttypen in einem Text verwendet.

Auf der nächsten Ebene, der Separation der einzelnen Buchstaben, können Buchstaben mit proportionalem oder gleichmäßigem Abstand aufeinander folgen, bestimmte Kombinationen wie „ff" oder „ft" zusammengezogen (Ligaturen) oder durch Kerning nicht mehr von einer senkrechten Linie zu trennen sein. ➤

To ft

Kerning und ...

... Ligaturen: Leseprobleme durch uneinheitliche Abstände zwischen den Buchstaben

Weiter gibt es mehrspaltiges Layout, und es wäre schade, wenn die Software versuchen würde, den Spalt zu überlesen. Schließlich können auch Bilder und Text abwechseln. Der Versuch, ein Bild als Text zu lesen, kann zu katastrophalen Ergebnissen führen.

Das Ideal

Sicher ist es schön, all die Probleme zu begreifen, mit denen sich eine OCR-Software herumzuschlagen hat. Doch diese Verständnisbereitschaft verschwindet schnell, sobald es darum geht, mit einem Produkt zu arbeiten. Dann zählt allein, wie gut es funktioniert.

Auf den ersten Blick sehr verlockend ist die Möglichkeit, einem System die Bedeutung der Buchstaben beizubringen. Das muß dann aber auch getan werden; bei kurzen Texten, die „mal eben" mit OCR bearbeitet werden sollen, lohnt es sich kaum. Ebenso führt jeder einstellbare Parameter zu einer endlosen Testerei, was denn nun am besten gehe.

Wie viele Fehlerkennungen lassen sich eigentlich tolerieren? Eine Treffsicherheit von 95 Prozent erscheint hoch, doch eine Schreibmaschinenseite mit 2000 Anschlägen hat dann immer noch 100 falsche und zu korrigierende Buchstaben – also in jeder Zeile mehr als zwei. Hier sieht man schon, daß der maschinell fehlerfrei erkannte und nicht nachzuarbeitende Text eine Illusion ist. Selbst eine Trefferquote von 99 Prozent ist ein schon selten erreichtes Traumergebnis unter idealen Voraussetzungen und würde im obigen Beispiel noch immer 20 Fehler pro Seite produzieren. Handelt es sich um einen Vertragstext mit Zahlen oder um Telephonnummern, ist kein einziger Fehler tolerierbar.

Die Überprüfung der Rechtschreibung ist dabei eigentlich nicht mehr so zeitraubend, da sie durch Spellingchecker unterstützt wird, wie sie inzwischen Teil beinahe jeder Textverarbeitung sind.

Die Alternative zu OCR ist manuelle Erfassung. Gute Kräfte schaffen mehr als 200 Anschläge pro Minute, achten von selbst auf das Wichtige und machen nicht die maschinentypischen und zum Teil haarsträubenden Fehler. Die Beispielseite wird also manuell in unter zehn Minuten erfaßt.

Die ideale OCR-Maschine hat keine Einstelloptionen, braucht keine Trainingsphase, muß nicht mit verschiedenen Zeichensätzen geladen werden und hat trotzdem eine Trefferquote von 99 Prozent und mehr. Da stöhnen sie, die Programmierer: Man drückt den Startknopf, und den Rest besorgt die Maschine.

Am Anfang war das Wort. Eingetippt auf einer Tastatur und dargestellt auf einem Bildschirm. Wie langweilig.

FRIEDRUN REINHOLD

Damals

Angesichts des allgemeinen Fortschritts verbreitet sich die Ansicht, daß erst vor kurzem alles Wesentliche erfunden wurde. Am Beispiel OCR läßt sich zeigen, daß das so nicht stimmt.

Das „Taschenbuch der Nachrichtenverarbeitung" von K. Steinbuch listete schon 1962 allein zwölf verschiedene Verfahren zur auto-

➤

Frankfurter Allgemeine Zeitung

Desktop Publishing

Kein Jota fällt aus dem Rahmen

Computer-Satz mit Calamus / Von Hans-Heinrich Pardey

Dieser Artikel, die Gestaltung dieser wie der anderen Seiten der Zeitung – das kommt alles aus dem Computer. Genauer gesagt: Rechner sind das Werkzeug für das Schreiben und Bearbeiten des Textes, aber auch für den Umbruch, die Plazierung der Artikel auf den Seiten, und schließlich für deren Satz im Laser-Belichter. Der Computer trennt Wörter, wobei man ihm genau auf die Finger gucken muß, er zieht die Spaltenlinien, läßt den Text von einer Spalte in die nächste umlaufen und stellt fest, ob der Artikel an die vorgesehene Stelle paßt oder nicht. Er spart den Platz für die Abbildungen aus, hält Abstand zwischen Text und Überschrift und Bild und Rand. Der Computer tut das alles nicht aus eigener Macht und Kenntnis. Man hat ihm mühsam genug beibringen müssen, ein vertrautes Erscheinungsbild darzustellen, das dem des Bleisatzes, der früheren Produktionsweise, entspricht. Die Maschinen, mit denen das bewerkstelligt wird, sind nicht gerade billig; ihre Bedienung verlangt Fachkräfte.

Desktop Publishing oder kurz DTP ist nun das Versprechen, auf einem einzigen Schreibtisch alle Arbeitsgänge von der Textherstellung bis zur fertigen Druckvorlage zu vereinen. Voraussetzung: Auf diesem Schreibtisch steht ein leistungsfähiger Personal Computer. Wenige Computer-Anwendungen sind so großspurig als revolutionäre Erleichterung auf den Markt

Bis zur letzten Minute läßt sich aktualisieren, Grafiken und Bilder können problemlos in den Text integriert werden, der Computer sorgt dafür, daß alles besser aussieht, als wenn geschnippelt und geklebt wird. Dazu kommt die auf Tastendruck entfesselte Kreativität bei der Gestaltung: Hier ein Kasten mit Schattenrand und da eine keck schrägstehende Schlagzeile, hier Negativdruck und da Formsatz, dazu jede Menge verschiedener Schriftarten und -größen. Der Rechner hat alles parat – auch für einen typografischen Amoklauf. Aber bislang summierten sich die Investitionen für ein professionell arbeitendes DTP-System rasch zu einem abschreckenden Einstandspreis von mehreren zehntausend Mark: ein schneller PC mit mehreren Megabyte Arbeitsspeicher und großer Festplatte, ein Großbildschirm und ein Laserdrucker, vielleicht noch ein Scanner zum Digitalisieren von Vorlagen und nicht zuletzt das umfangreiche Programm.

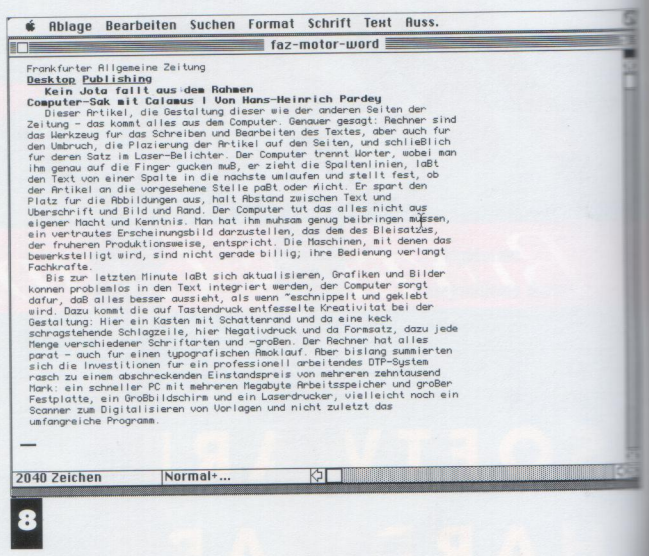
Eine preisgünstige Lösung

Beim Preis hat jetzt Atari angegriffen, mit einem kompletten System für weniger als 10 000 Mark. Es zielt auf den, der ernsthaft arbeiten, aber nicht soviel bezahlen will wie bei Apple, dem bevorzugten Ausstatter von DTP-Anwendern. Atari will die Kundschaft vom privaten power

Die Druckvorlage di

keine größeren Arbe
speicher laden lassen
zu dem Urteil, Me
und Calamus bildete
fast professionelles A

1



Ablage Bearbeiten Suchen Format Schrift Text Auss.

faz-motor-word

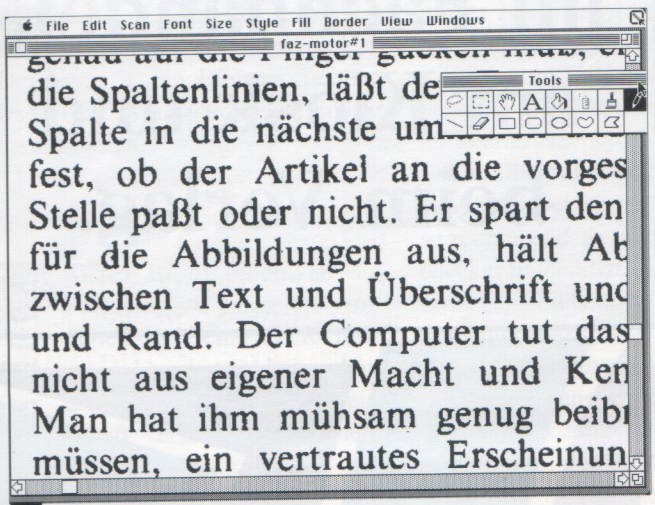
Frankfurter Allgemeine Zeitung
Desktop Publishing
Kein Jota fallt aus dem Rahmen
Computer-Sak mit Calamus ! Von Hans-Heinrich Pardey

Dieser Artikel, die Gestaltung dieser wie der anderen Seiten der Zeitung - das kommt alles aus dem Computer. Genauer gesagt: Rechner sind das Werkzeug fur das Schreiben und Bearbeiten des Textes, aber auch fur den Umbruch, die Plazierung der Artikel auf den Seiten, und schlieBlich fur deren Satz im Laser-Belichter. Der Computer trennt Worter, wobei man ihm genau auf die Finger gucken muß, er zieht die Spaltenlinien, laBt den Text von einer Spalte in die nachste umlaufen und stellt fest, ob der Artikel an die vorgesehene Stelle paBt oder nicht. Er spart den Platz fur die Abbildungen aus, halt Abstand zwischen Text und Uberschrift und Bild und Rand. Der Computer tut das alles nicht aus eigener Macht und Kenntnis. Man hat ihm muhsam genug beibringen mussen, ein vertrautes Erscheinungsbild darzustellen, das dem des Bleisatzes, der fruheren Produktionsweise, entspricht. Die Maschinen, mit denen das bewerkstelligt wird, sind nicht gerade billig; ihre Bedienung verlangt Fachkrafte.

Bis zur letzten Minute laBt sich aktualisieren, Grafiken und Bilder konnen problemlos in den Text integriert werden, der Computer sorgt dafur, daB alles besser aussieht, als wenn "eschnippelt und geklebt wird. Dazu kommt die auf Tastendruck entfesselte Kreativitat bei der Gestaltung: Hier ein Kasten mit Schattenrand und da eine keck schragstehende Schlagzeile, hier Negativdruck und da Formsatz, dazu jede Menge verschiedener Schriftarten und -groBen. Der Rechner hat alles parat - auch fur einen typografischen Amoklauf. Aber bislang summierten sich die Investitionen fur ein professionell arbeitendes DTP-System rasch zu einem abschreckenden Einstandspreis von mehreren zehntausend Mark: ein schneller PC mit mehreren Megabyte Arbeitsspeicher und großer Festplatte, ein GroBbildschirm und ein Laserdrucker, vielleicht noch ein Scanner zum Digitalisieren von Vorlagen und nicht zuletzt das umfangreiche Programm.

2040 Zeichen | Normal+...

8



File Edit Scan Font Size Style Fill Border View Windows

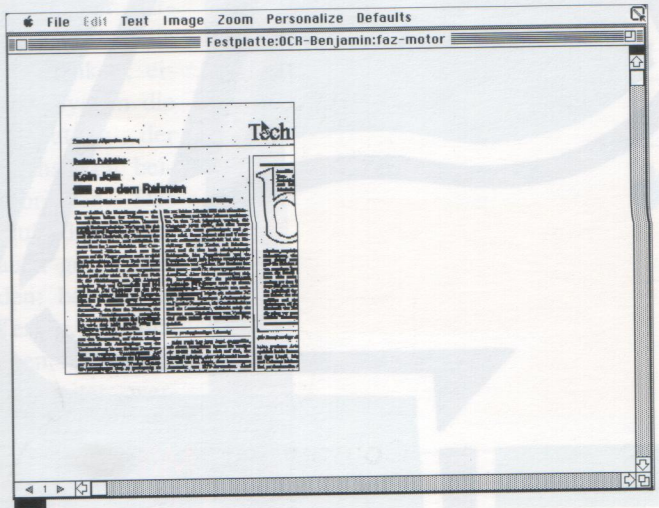
faz-motor#1

Tools

die Spaltenlinien, läßt de
Spalte in die nächste um
fest, ob der Artikel an die vorges
Stelle paßt oder nicht. Er spart den
für die Abbildungen aus, hält Ab
zwischen Text und Überschrift unc
und Rand. Der Computer tut das
nicht aus eigener Macht und Ken
Man hat ihm mühsam genug beibi
müssen, ein vertrautes Erscheinun

2

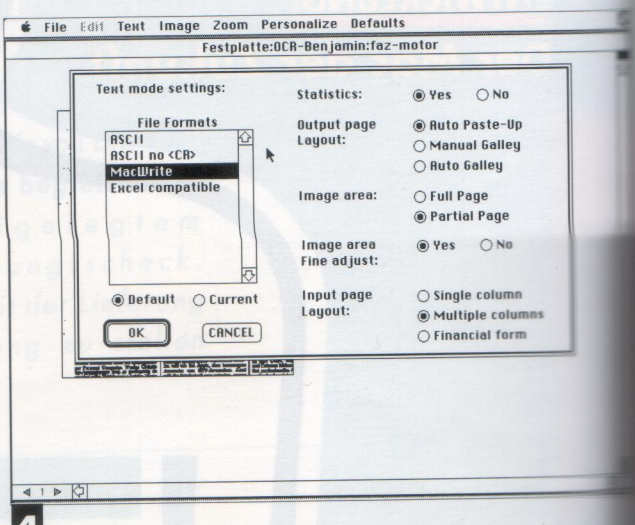
OmniPage-Lesevorgang: Die Vorlage wurde gescannt (Abbildung 1) und in guter Qualität (Abbildung 2) in OmniPage eingelesen (Abbildung 3). Dann waren zunächst die gewünschten Optionen (Abbildung 4) und der Vorlagenbereich auszuwählen (Abbildung 5), bevor sich das Programm durch schwarze Markierungen beim Lesen zusehen ließ (Abbildung 6). Der erfaßte Text erschien zunächst als Roh-Version im OmniPage-Texteditor (Abbildung 7) und schließlich – abgesehen von den Umlauten – in beeindruckender Qualität mit Fettungen und Unterstreichungen im endgültigen Textverarbeitungsformat (Abbildung 8).



File Edit Text Image Zoom Personalize Defaults

Festplatte:OCR-Benjamin:faz-motor

3



File Edit Text Image Zoom Personalize Defaults

Festplatte:OCR-Benjamin:faz-motor

Text mode settings:

File Formats
ASCII
ASCII no <CR>
MacWrite
Excel compatible

Default | Current

OK | CANCEL

Statistics: Yes | No

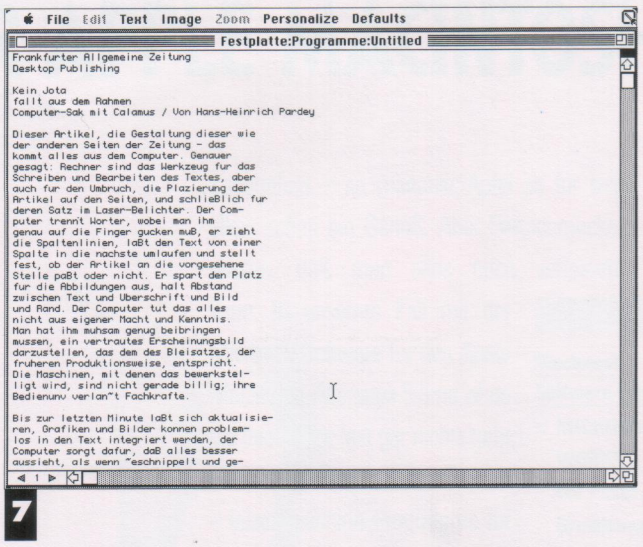
Output page Layout: Auto Paste-Up | Manual Galley | Auto Galley

Image area: Full Page | Partial Page

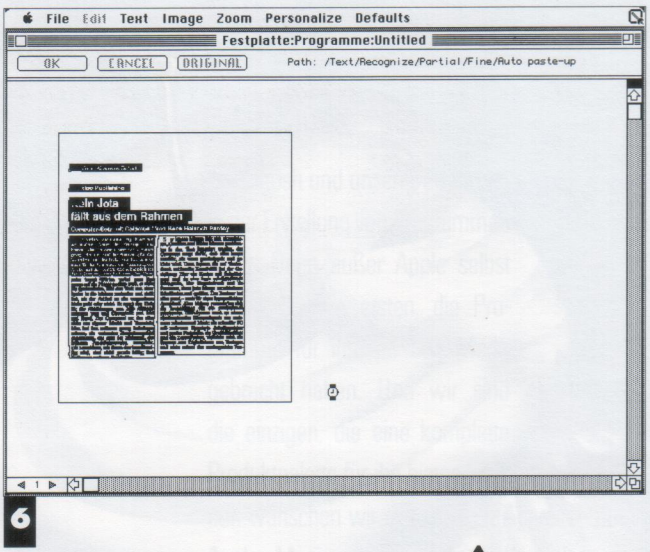
Image area Fine adjust: Yes | No

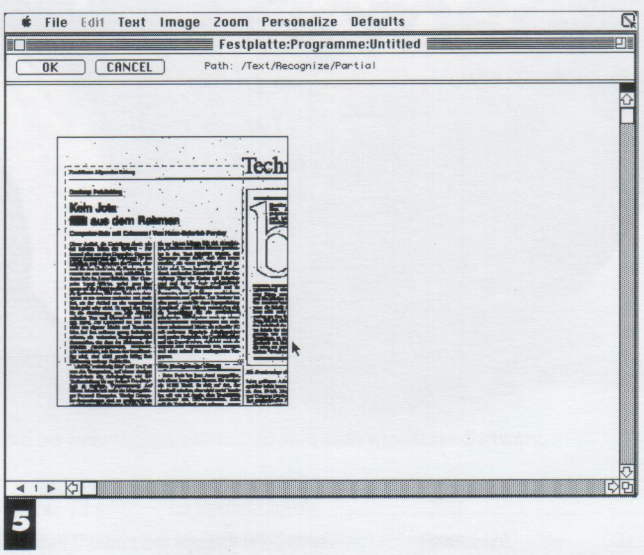
Input page Layout: Single column | Multiple columns | Financial form

4



OmniPage





matischen Zeichenerkennung auf. Beschrieben werden nicht nur Pattern matching und Feature recognition, die Firma Standard Elektrik Lorenz hatte zu dieser Zeit mit der ZL 57 auch schon eine Maschine auf dem Markt, die 400 Ziffern pro Sekunde erkennen konnte.

Doch die Anfänge liegen noch weiter zurück. Mit dem Deutschen Reichspatent 66 247 vom 7. 5. 1929 („Vorrichtung zur Steuerung von Maschinen durch strahlende Energie") hat G. Tauschek die Zifferner-kennung mit Hilfe des Pattern matching beschrieben (siehe Abbildung Seite 24). Da das Verfahren sehr instruktiv aus der Zeichnung hervorgeht, sei es hier beschrieben. Die Lampe beleuchtet eine Ziffer, welche durch die Linse auf die rotierende Trommel mit dem Negativ der zehn Ziffern abgebildet wird. Die sich in der Trommel befindende Photodiode empfängt die Differenz der Ziffer auf dem Papier zu dem vorgegebenen Pattern. Stimmen beide überein, geht der Photostrom auf ein Minimum, die Ziffer wird erkannt, hier als „1".

Heute

Neben Blatt- und Beleglesern verschiedener Hersteller ist die amerikanische Firma Kurzweil Dreh- und Angelpunkt der OCR-Industrie. Aus einem Projekt am MIT mit dem Ziel, eine Lesemaschine für Blinde zu entwickeln, wurde 1974 eine Firma, die diese Lesemaschine perfektionieren und vermarkten sollte. Die Kurzweil Recognition Machine konnte nicht nur Texte lesen – und zwar sowohl Schreibmaschine als auch Buchdruck –, sondern diese auch in gesprochene Sprache umwandeln, und stieß in öffentlichen Einrichtungen auf großes Interesse.

Die mit der KRM gemachten Erfahrungen führten zur Entwicklung der Kurzweil Data Entry Machine, die 1978 das Licht des Marktes erblickte. Sie ist der Maßstab, an dem noch immer OCR-Systeme gemessen werden. Es handelt sich um eine auf OCR spezialisierte Maschine mit dediziertem Prozessor und Scanner. Das System ist trainierbar und bietet das obere Leistungsspektrum des Erreichbaren – allerdings auch im

Preis, der mit 130 000 bis 250 000 Mark je nach Konfiguration einem Einsatz für jedermann bislang im Wege steht.

Mit zunehmender Verbreitung leistungsfähiger Scanner im PC-Markt und dem großen Interesse an Optical Character Recognition lag es nahe, durch die Entwicklung geeigneter Software in den Bereich bisher dedizierter Systeme einzudringen.

Auf Messen und in den Entwicklungsabteilungen tauchten ab 1987 erste Programme auf, die aber entweder im Preis teilweise weit über 10 000 Mark lagen oder aber von den Scanner-Herstellern als weiteres Argument zum Kauf eines Geräts kostenlos beigefügt werden sollten. Während erstere leidlich funktionierten, waren die Prototypen letzterer Spezies völlig unakzeptabel.

Inzwischen hat sich die Situation verändert. Im Jahr 1988 kamen zur MacWorld-Expo mehrere Hersteller mit neuen Produkten auf den Markt.

OmniPage

Das amerikanische Unternehmen Caere (gesprochen „kähr"), gegründet 1973, hat sich bisher mit optischen Lesegeräten beschäftigt. Einer der Köpfe von Caere ist Dr. Robert Noyce, Mitgründer von Intel und Erfinder eines zur Herstellung integrierter Schaltungen wichtigen Prozesses. Als OEM-Lieferant (Firmen, die Produkte herstellen, auf die andere Firmen ihr Label kleben) für IBM, Digital, Memorex und ITT hat Caere Bar-Code-Lesegeräte entwickelt (bekannt zum Beispiel vom Ablesen der Lebensmittel-Etiketten an den Kassen von Supermärkten). Doch dieser Markt ist inzwischen hart umworben, auch von den Japanern.

Um ein zweites Standbein aufzubauen, hat man sich der Entwicklung von OCR-Software zugewandt und nach einem Aufwand von 25 Mannjahren „OmniPage" für den IBM PC und den Mac herausgebracht. Das IBM-Produkt funktioniert auf der Basis einer 68020-Koprozessor-Karte mit zwei Megabyte Speicher. Für den Mac ist keine Zusatzkarte nötig, dafür beansprucht die Software aber vier Megabyte Hauptspeicher. ➤

OmniPage ist ein erstaunliches Produkt. In seiner Arbeitsweise ist es eher dem Feature-recognition-Verfahren zuzuordnen, da es nicht trainierbar ist und auch sonst kaum Optionen enthält. Dennoch hat diese Software in unseren Tests bei weitem am besten abgeschnitten und generell über 99 Prozent der Buchstaben richtig erkannt. Auskünfte über die Arbeitsweise von OmniPage waren recht spärlich, das System scheint während des Leseprozesses selbständig zu lernen und wird dementsprechend immer schneller.

Die Handhabung entspricht fast schon der idealen OCR-Maschine. Obwohl die Software absolut stabil auch unter dem MultiFinder läuft, gab es allerdings ein paar kleinere Probleme:

• Bisher wird kein internationaler Zeichensatz unterstützt. Das bedeutet, daß die Umlaute nicht erkannt werden, ein „ä" demzufolge als „a" gelesen wird. Im Februar 1989 soll OmniPage jedoch kostenlos für den internationalen Zeichensatz upgedatet werden.

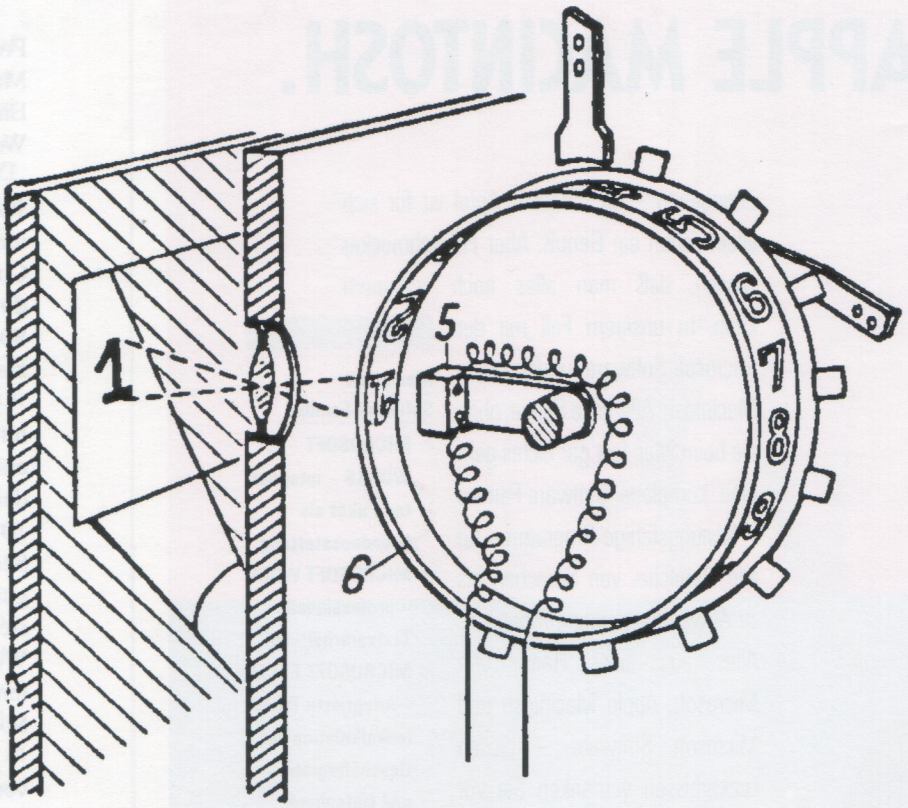
• In MacWrite erstellte Texte wurden mitunter nicht als solche erkannt.

• Der zum Test benutzte Microtek MS 300 QS wird nicht unterstützt, was aber ein Handhabungsproblem ist, da TIFF-Files eingelesen werden. – An dieser Stelle sei darauf hingewiesen, daß eine DIN-A4-Seite im TIF-Format ein Megabyte Plattenspeicher braucht. – Die Zahl unterstützter Scanner soll jedoch noch vergrößert werden.

OmniPage liest einwandfrei mehrspaltige Texte und setzt die Spalten im übersetzten Text hintereinander. Graphiken werden eindeutig erkannt und ausgespart, nur manchmal fügt die Software an der Trennstelle von Text und Graphik ein falsches Textzeichen ein. Soweit möglich, erhält OmniPage Fettdruck und Unterstreichungen im endgültigen Format.

Die wirklich benötigte Zeit zum Einlesen eines Textes: Ein Artikel von 16 Seiten und etwa 15 000 Anschlägen, der uns zugefaxt worden war, sollte mit OmniPage erfaßt werden. Zur Verfügung standen ein Mac II mit fünf Megabyte Speicher, ein Microtek-Scanner und genügend Platz auf der Platte (ein seltener Fall). Dennoch dauerte das Scannen und Übersetzen des Textes aus den Einzelseiten über zwei Stunden, ohne Rechtschreibkorrektur. Manuell wäre der Text in anderthalb Stunden erfaßt worden, mit einer zu akzeptierenden Rechtschreibung.

OmniPage ist bisher nur in Englisch mit entsprechendem Manual verfügbar, kostet 2268,60 Mark und ist nicht kopiergeschützt. Der deutsche Distributor ist die Firma Prisma. Ende Februar 1989 wird zum Preis von 2838,60 Mark die deutsche Version mit deutschem Zeichensatz und Unterstützung des Siemens-Scanners erscheinen. Schade nur, daß der zum Arbeiten nötige Speicher so viel wie die Software selbst kostet.



Pattern-matching-Verfahren zur Zeichenerkennung: Reichspatent 66 247 von G. Tauschek aus dem Jahre 1927

Read-it

Das Produkt „Read-it" der Firma Olduvai repräsentiert die klassische OCR-Software nach dem Pattern-matching-Verfahren. Neben den 41 mitgelieferten Fonts bietet sie außerdem die Möglichkeit, interaktiv neue Fonts zu trainieren.

Read-it ist ein einfaches, konsequentes und durchschaubares Programm, mit dem man gut arbeiten kann. Es läuft einwandfrei unter dem MultiFinder, zwei Megabyte Speicher sind zwar sehr zu empfehlen, aber nicht unbedingt nötig. Unterstützt werden wiederum viele gängige Scanner, das TIF- und das MacPaint-Format – vorzugsweise sollte ersteres benutzt werden.

Gäbe es OmniPage nicht, wäre dieses Programm erste Wahl. Doch vor all dem Luxus, den OmniPage bei sehr hoher Erkennungsrate bietet, erscheint Read-it eher wie aus einer vergangenen Generation. Wie gesagt erweisen sich alle Optionen als unnötig und hinderlich. Das fängt damit an, wie man einen Schrifttyp erkennt, um die vorgegebenen Sätze benutzen zu können. Wer kein Graphiker oder Setzer ist, erkennt eben keine Times. Das nächste Problem ist, daß der zum Teil recht lange dauernde Erkennungsvorgang nicht abgebrochen werden kann. So muß man immer warten, bevor ein neuer Zeichensatz ausprobiert werden kann.

Das oftmals notwendige Erstellen eines neuen Zeichensatzes in der Trainingsphase dauert laut Handbuch zwischen einer und drei Stunden, wenn eine 99prozentige Erkennungsrate angestrebt wird – die beim Te- ➤

sten jedoch nie erreicht wurde. Einen wirklich guten neuen Zeichensatz zu erstellen ist nicht ganz leicht, da die letzten Optimierungen eines aufmerksamen Studiums der Lesestatistiken bedürfen. Eine gewisse Einarbeitungsphase sollte man daher schon einplanen.

Read-it ist bislang nur in englischer Version mit englischem Handbuch erhältlich, das Programm ist nicht kopiergeschützt und bei der Firma Pandasoft zu beziehen. Das Preis-Leistungs-Verhältnis ist mit 987 Mark vergleichsweise sehr gut.

Da stöhnen sie, die Programmierer: Man drückt den Startknopf, und den Rest besorgt die Maschine.

FRIEDRUN REINHOLD

Readstar II+

Die französische Firma Inovatic wurde 1985 gegründet und war eine der ersten, die OCR-Software nach dem Pattern-matching-Verfahren mit Trainingsmöglichkeiten anbot. Nach einer IBM-Version existiert jetzt auch eine für den Macintosh mit dem Namen „Readstar II+".

Die uns zur Verfügung gestellte Demo-Version war – bis auf fehlende Speichermöglichkeit – ein vollständiges Produkt. Sie braucht zwei Megabyte Speicher. Grundsätzlich war die Erkennungsquote sehr gut, wenn eine entsprechende Trainingsphase eingelegt wurde. Für das Verfahren gilt das bereits über das Pattern-matching Gesagte. Leider unterstützt Readstar nicht das TIF-Format, bei der Darstellung des VersaScan-Formates gab es Schwierigkeiten, so daß nur ein schwarz gerasterter Querbalken zu sehen war und danach nichts mehr passierte. Beim Erstellen eines neues Schrifttyps stürzte die Software unter dem MultiFinder ab. Das fragmentarische englische Handbuch soll ergänzt werden, die englische Software kostet nach Preissenkung 13 680 Mark, ist kopiergeschützt und bei der Firma IPT zu haben.

TextScan

Ebenfalls eine OCR-Software nach dem Pattern-matching-Verfahren bietet die spanische Firma CTA an, die hier unter dem Namen „TextScan" in einer deutschen Version (2.03) angeboten wird. CTA war vor der Umstrukturierung von Apple in Europa Apple Spanien. Das Programm wurde zusammen mit der Universität Barcelona entwickelt.

TextScan hat das am weitesten entwickelte Interface bei der Erstellung neuer Zeichensätze, es lassen sich falsch identifizierte Zeichen zurücknehmen und oft verwechselte Buchstaben wie „i" und „l" besonders berücksichtigen. Das Programm sollte zum Arbeiten ein Megabyte Speicher haben, läuft aber auch auf dem Mac 512 und unterscheidet sich damit angenehm von anderen Speicherfressern. Es unterstützt alle gängigen Scanner.

Leider gab es auch hier unerklärliche Abstürze unter dem MultiFinder, deren Ursachen aber inzwischen beseitigt sein sollen. Die Mac-Benutzerschnittstelle bedarf der Überarbeitung, und außerdem fällt eine Unterscheidung von Klein- und Großbuchstaben wie „u" und „U" in der Trainingsphase schwer. Auch dieses Manko hat man inzwischen erkannt und verbessert. Mit den vorgegebenen Schriftsätzen waren nur schlechte Erkennungsraten möglich, allerdings lassen sich mit selbst erstellten Schriftsätzen sehr hohe Erkennungsraten erzielen. In zukünftigen Versionen will sich CTA weiter in Richtung Feature recognition entwickeln, um auf die Trainingsphase verzichten zu können.

Die neuste Version 2.1 von TextScan ist für 4902 Mark über die Firma SEP, Köln, als Distributor zu beziehen und nicht kopiergeschützt.

Die Entscheidung

Mit OCR-Software für PCs und Scanner war es bisher so wie mit dem Hund, der Schach spielt. Man freut sich, daß es überhaupt geht. Doch mit OmniPage wurde eine neue Generation von OCR-Software kreiert. Es ist das herausragende Produkt, das in seiner Unkompliziertheit und Erkennungssicherheit Erstaunliches leistet. Die Konkurrenz wird natürlich versuchen, so rasch wie möglich aufzuholen.

Auch bei OCR sollte man sich nicht vom Glauben an die Technik hinwegtragen lassen. In vielen Fällen ist die manuelle Erfassung eines Textes oder aber ein bißchen mehr Planung bei der Arbeit sinnvoller, etwa die Übertragung eines Textes per Modem.