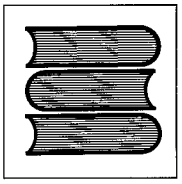
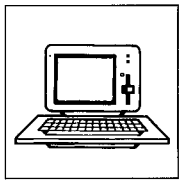



"Make sure that what comes out closely resembles what goes in. Otherwise you will be headed for frustration and failure"

(BYTE, A Simple Equation for Choosing OCR, April 1991, S. 236).

Optische Zeichenerkennung

Maximilian Herberger

Noch vor ca. 10 Jahren bedurfte es mühevoller Überzeugungsarbeit, um Juristen die Bedeutung der OCR-Technologie speziell für juristische Projekte deutlich zu machen. Inzwischen hat sich die Situation geändert. Wo Juristen EDV-Projekte diskutieren, ist auch von Scannern und deren notwendiger Anschaffung die Rede. Oft geht mit derartigen Vorschlägen aber ein unkritischer Optimismus einher, der für die Auswahl des richtigen OCR-Systems und die Planung von OCR-Projekten nicht förderlich ist. Deswegen geht es jetzt darum, die Beurteilungskompetenz an den entscheidenden Stellen so zu stärken, daß die optische Zeichenerkennungstechnologie nicht wegen fehlerhafter Systemwahl in Mißkredit gerät. Denn eins steht fest: Die OCR-Technologie hat einen Stand erreicht, der sie in einem genau definierbaren Anwendungsbereich als Produktionsinstrument tauglich und der Erfassung von Hand überlegen macht. Die zahlreichen OCR-gestützt produzierten juristischen CD-ROM's beweisen es.

Ein wenig Geschichte

Mit automatischer Zeichenerkennung (abgekürzt: *OCR* für *Optical Character Recognition)* beschäftigt man sich in der Computerwelt seit ca. 30 Jahren. Schon 1961 publizierte das "National Bureau of Standards" der Vereinigten Staaten einen Bericht über den Stand der erreichten Technik.[1]

Der Belegleser IBM 1275

Das Interesse richtete sich zunächst auf das maschinelle Erfassen von Belegen. Von daher ist der Terminus "Belegleser" immer noch verbreitet. Eines der ersten Systeme dieser Art stammte von IBM und trug die Bezeichnung "IBM 1275 Recognition System".[2]

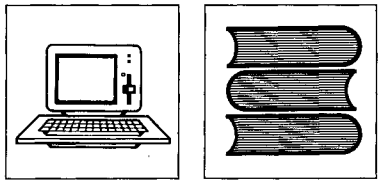
Allerdings handelt es sich bei dem "Beleglesen" um ein Spezialproblem der Zeichenerkennungstechnologie. Denn einerseits ist die Aufgabe hier schwerer als bei der Interpretation maschinell erzeugter Vorlagen: Handschriftenvorlagen weisen selbst bei Verwendung von Blockschrift eine wesentlich größere Variationsbreite auf als Druckvorlagen. Andererseits ist die Aufgabe aber auf Grund der besonderen Konstellation auch wieder leichter als bei beliebigen Druckvorlagen: Beim Beleg gibt es bestimmte Zonen, in denen man nur mit bestimmten Zeichen (etwa Zahlen) oder einer eingeschränkten Menge von Zeichenkombinationen zu rechnen hat. Deswegen sind vielfältige Syntax- und Plausibilitätskontrollen denkbar.

Recht bald wandte sich das wissenschaftliche Interesse allgemeinen Zeichenerkennungsproblemen zu. Ende der sechziger Jahre arbeitete die Cognitive Information Processing Group des "M.I.T- Research Laboratory of Electronics" an einem Projekt mit dem Ziel, eine "Lesemaschine" mit Sprachausgabe für Blinde zu entwickeln. 1968 existierte ein experimentelles System, das weitere Forschungen als aussichtsreich erscheinen ließ.[3]

1 Vgl. National Bureau of Standards (Hrsg.), Automatic Character Recognition – A State-of-the-Art Report. Technical Note Nr. 112, Mai 1961.

2 Vgl. zur Entwicklung dieses Systems H. van Steenis, The IBM 1275 Recognition System and Its Development, in: Otto-Joachim Grüsser/Rainer Klinke (Hrsg.), Zeichenerkennung durch biologische und technische Systeme, Berlin 1971, S 253–261.
Auch heute noch bildet die Frage der maschinellen Interpretation von Belegen beispielsweise im Bankbereich ein wichtiges technologisches Problem.

3 Vgl. Samuel J. Mason/Jon K. Clemens, Character Recognition in an Experimental Reading Machine for the Blind, in: Paul A. Kolers/Murray Eden (Hrsg.), Recognizing Patterns Studies in Living and Automatic Systems, Cambridge 1968, S. 156–167.



OCR-Pionier: Raymond Kurzweil

Wie bei zahlreichen amerikanischen Forschungsprojekten mündete auch dieses schließlich in eine wirtschaftliche Umsetzung. Im Jahre 1974 gründete Raymond Kurzweil in den USA eine Firma, um eine Lesemaschine für Blinde zur Produktionsreife zu entwickeln. 1976 war ein Prototyp fertiggestellt, den man mittlerweile im Bostoner Computermuseum besichtigen kann. Mit dem "KRM" (Kurzweil Reading Machine) genannten Gerät konnten maschinenschriftliche und gedruckte Texte gelesen werden. Gleichzeitig verfügte die "KRM" über eine Sprachausgabeeinrichtung, mit der sich die gelesenen Texte in gesprochene Sprache umsetzen ließen. Die Maschine fand großes Interesse. Bis 1978 wurden in amerikanischen Blindenanstalten und Bibliotheken über 50 Exemplare installiert. Nach den ersten Erfahrungen mit der "KRM" lag der Gedanke nicht fern, ein Gerät dieser Art in industriellen Anwendungen dazu zu benutzen, um maschinengeschrieben oder gedruckt vorliegende Texte in computerlesbare Form zu bringen. Ergebnis dieser Überlegungen war der Bau der "KDEM" (Kurzweil Data Entry Machine), die 1978 auf den Markt kam. Die Situation auf dem Gebiet der Zeichenerkennungsgeräte ist viele Jahre lang durch den Maßstab geprägt worden, den die "KDEM" gesetzt hat. Allerdings unterstützte der Preis eines "KDEM"-Systems (je nach Ausstattung und verfügbaren Rabatten ab 130.000 bis über 200.000,– DM) seinerzeit nachhaltig die Annahme, es könne Zeichenerkennung nur auf einem für dezentrale Lösungen kaum erschwinglichen Kostenniveau geben.[4]

Eine neue Generation von Scannern

Inzwischen sind die Scanner, die notwendigen Endgeräte für jeden EDV-gestützten Zeichenerkennungsprozeß, preiswerter geworden. Man kann bereits für unter 5.000,– DM leistungsfähige Geräte dieser Art erwerben. Es gibt sogar zu Preisen von deutlich unter 1.000,– DM kleine Scan-Geräte mit beachtlicher Auflösung, die man über Vorlagen hinwegbewegt, um diese so graphisch zu erfassen.

Wegen dieser Entwicklung bei den Scannern hat das Interesse an Zeichenerkennungssystemen zugenommen. Denn es sieht jetzt so aus, als könne Zeichenerkennung an den einzelnen Arbeitsplatz gebracht werden.

Optimistische Erwartungen und ein weit verbreiteter Irrtum

Die sich angesichts immer preiswerterer Scanner einstellenden optimistischen Erwartungen beruhen in der breiteren Öffentlichkeit oft auf einem (häufig von der Werbung planvoll unterstützten) Mißverständnis: Es wird nicht erkannt, daß zwischen zwei Arbeitsvorgängen genau unterschieden werden muß, nämlich

- erstens dem Scannen der Vorlage mit dem Übertragen von Schwarz-Weiß-Werten in den Rechner

und

- zweitens der Interpretation dieser Schwarz-Weiß-Werte als Buchstaben durch ein Zeichenerkennungsprogramm.

Der erste Vorgang ist in unproblematischer Weise "Stand der Technik". Solange man die erfaßte Textvorlage nur als Bild weiterverarbeiten will (d.h. ohne Interpretation der in diesem Bild enthaltenen Buchstaben), wird meist sogar schon mit dem Scanner eine entsprechende Software geliefert.

Grundsätzlich anders sieht es aber aus, wenn die eigentliche Zeichenerkennung zur Debatte steht, d.h. die Interpretation der übertragenen Schwarz-Weiß-Werte als Buchstaben. Dabei geht es um komplizierte Fragen der Mustererkennung, die bezogen auf unterschiedliche Schriftvorlagen bisher unterschiedlich gut gelöst sind. Um sich hier zu orientieren, muß man einige Differenzierungen beachten, die im Rahmen dieses Beitrags dargestellt werden.

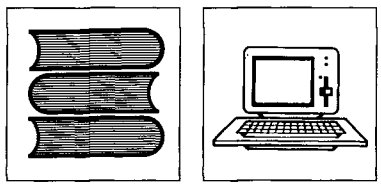
Etwas Zeichenerkennungstheorie

Menschliche und "maschinelle" Kompetenz beim Zeichenerkennen

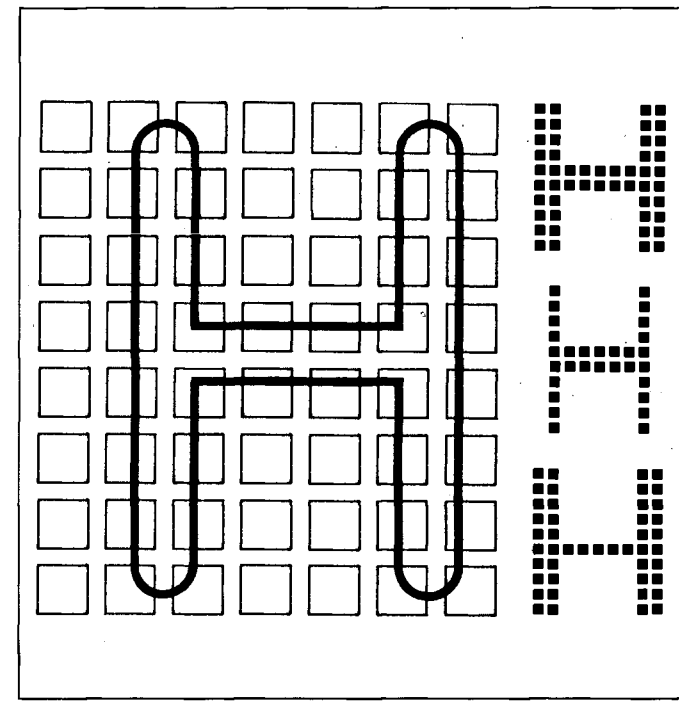
Auf den ersten Blick will nicht einleuchten, warum maschinelle Zeichenerkennung besonders schwierig sein sollte. Man ist geneigt, diese Aufgabe als leicht einzustufen, weil man sie selbst ohne größere Probleme bewältigen kann. Bei genauerer Betrachtung zeigt sich aber, daß für Mensch und Maschine ganz unterschiedliche Ausgangssituationen bestehen. Von Bedeutung sind hauptsächlich drei Punkte.

Die Entstehung der Abbildungsungenauigkeiten beim Scannen kann durch das Überlagern eines diskreten Rasters (üblicherweise 300 bzw. 400, zunehmend auch 600 Dot pro Inch) erklärt werden. Die Abbildung zeigt schematisch das Zustandekommen unterschiedlicher Pixelbilder durch Rasterverschiebung.

4 Vgl. als Erfahrungsbericht zum Einsatz der KDEM bei rechtshistorischen Projekten Maximilian Herberger, Die Maschine, die liest, Rechtshistorisches Journal 2 (1983), S 227–233.



Normalisierung der Konturen



Die Entstehung der Abbildungsungenauigkeiten beim Scannen kann durch das Überlagern eines diskreten Rasters (meistens 300 Dot pro Inch) erklärt werden. Die Abbildung zeigt schematisch das Zustandekommen unterschiedlicher Pixelbilder durch Rasterverschiebung.

Erstens erscheinen dem Menschen die Buchstaben von vornherein in einer "normalisierten" Form. Man hat etwa bei normalen Buchschriften den Eindruck, daß alle Zeichen eine "scharfe" Kontur haben. Für die Maschine "sieht" das anders aus. Sie erhält ein Muster, bei dem die Konturen auch des regelmäßigsten Buchstabens an der Oberfläche als leicht "ausgefranst" erscheinen. Daraus ergibt sich beim maschinellen Zeichenerkennen die Aufgabe der "Normalisierung" schon dort, wo es um die (vom menschlichen Auge unbewußt geleistete) Festlegung der äußeren Buchstabenlinie geht.

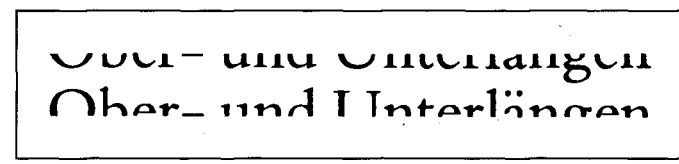
Ganzheitliche Wahrnehmung und Ergänzung von Bruchstükken

Zweitens ist der Mensch, der eine Schrift kennt, in der Lage, unvollständige oder schlecht erkennbare Buchstaben zur vollen Buchstabengestalt zu ergänzen. Das funktioniert – Kenntnis des betreffenden Alphabets vorausgesetzt – sogar in einer Sprache, deren Wortbedeutungen man nicht beherrscht. Wir sehen trotz verschiedener Deformationen, daß ein A ein A ist, weil wir wissen, wie "ein A" aussieht. Ein Zeichenerkennungsprogramm kann diese Ergänzung zur Vollgestalt nicht in der gleichen flexiblen Weise vornehmen, weil es nicht in allen Deformationen "ein A" als dahinterstehendes Muster erkennen kann.

Kontextanalyse

Drittens verwendet der Mensch beim Lesen (in einer ihm vertrauten Sprache) den Bedeutungskontext der Worte immer mit. Auch das hilft ihm dabei, beschädigte oder nur teilweise vorhandene Buchstaben korrekt zu behandeln. Ein berühmtes Experiment dazu, das man leicht selbst nachvollziehen kann, läuft folgendermaßen ab: Man deckt (etwa in diesem Text) mit einem Lineal die untere Hälfte einer Zeile ab und versucht dann, den Text zu lesen. Das gelingt ohne weiteres. Man kann sogar das Lineal noch über die Hälfte der Zeile nach oben schieben, ohne daß die Erkennungsleistung beeinträchtigt wird.

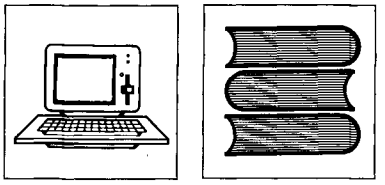
Bei der Erkennung von Zeichen durch den Menschen spielt das vollständige Zeichenbild nur eine untergeordnete Rolle. Das Weglassen der Unterlängen erschwert das Lesen eines Textes nur unwesentlich.



Methoden der maschinellen Zeichen-"Erkennung"

Musterüberlagerung ("Pattern matching")

Der einfachste (und "maschinennächste") Gedanke bei der "maschinellen" Zeichenerkennung läuft darauf hinaus, das vom Scanner übertragene Grauwertmuster mit Mustern zu vergleichen, die man vorher als "Normalmuster" der Buchstaben abgespeichert hat. Man bezeichnet diese Methode als "pattern matching". Eine genaue Betrachtung der Buchstaben in dem vom Scanner gelieferten Ausgangsmaterial zeigt aber, daß unter Normalbedingungen nahezu kein Exemplar eines Buchstabens vollständig mit dem anderen übereinstimmt. Es gibt



also keine perfekte Übereinstimmung ("match"). Deshalb ist man schon beim Ansatz des "pattern matching" gezwungen, das Ausgangsmaterial an Buchstabenexemplaren in irgendeiner Form zu "normalisieren". Dazu wurden Techniken der Konturenbegradigung ("smoothing") entwickelt. Außerdem muß man einen gewissen Toleranzbereich hinsichtlich dessen zulassen, was man noch als Übereinstimmung gelten läßt. Derartige Toleranzen, die wegen der Heterogenität des Ausgangsmaterials unumgänglich sind, bilden allerdings gleichzeitig eine mögliche Fehlerquelle, wenn der Toleranzbereich zu großzügig gewählt wird.

Identifikation über Eigenschaften ("feature recognition")

Ein anderer Ansatz der Zeichenerkennungstechnologie arbeitet auf der Basis von Buchstabeneigenschaften. Diese Methode wird als "feature recognition" bezeichnet. Ein "O" beispielsweise ist eine runde, geschlossene Form. Wenn man jetzt einen Konturenverfolgungsalgorithmus programmiert, der sich um die Buchstabengestalt herumbewegt, so würde dieser beim "O" an den Ausgangspunkt zurückkehren, ohne eine Linie gekreuzt zu haben. Das reicht allerdings noch nicht aus, um das "O" von anderen Buchstaben zu unterscheiden. Vielmehr muß man noch die Information hinzunehmen, daß es einen "Innenraum" gibt, aus dem man sich nicht "herausbewegen" kann, ohne eine Linie zu kreuzen. Auf diese Weise läßt sich der Buchstabe O innerhalb der Menge der Buchstaben eindeutig mit Hilfe zweier Eigenschaften beschreiben, die von den Unregelmäßigkeiten der Oberflächenkontur unabhängig sind, solange man es nicht mit Unterbrechungen des Linienzugs zu tun hat.

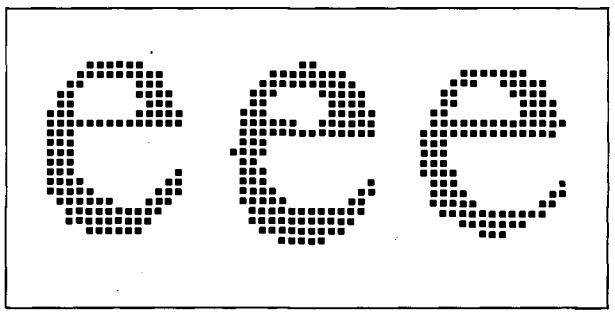
"Pattern matching" vs. "feature recognition"

Das Verfahren der Identifikation über Eigenschaften scheint auf den ersten Blick dem Prinzip der Musterüberlagerung eindeutig überlegen zu sein. Denn durch die stärkere Abstraktion von Unregelmäßigkeiten der Vorlage wird allem Anschein nach ein größerer Wirkungsgrad erzielt. Diese Einschätzung ist aber nicht uneingeschränkt richtig. Vielmehr muß man für die durch die Beachtung von Buchstabeneigenschaften erzielbaren Vorteile andere Nachteile in Kauf nehmen.

Mehrdeutigkeiten von Eigenschaftsbeschreibungen

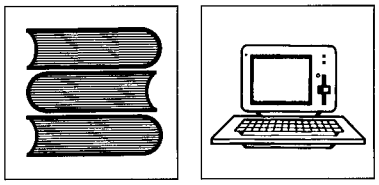
Schon bei dem eben gewählten Beispiel des Buchstabens O etwa wäre die Eigenschaftsbeschreibung bereits dann nicht mehr eindeutig, wenn man zu den Buchstaben die Zahlen hinzunimmt. Denn die geschlossene Form ist in gleicher Weise für das O und die Null charakteristisch. Weitere differenzierende Strukturmerkmale dürften nur sehr schwer zu finden sein. Es scheint also auf der Ebene der Eigenschaftsbeschreibung kaum oder auch überhaupt nicht behebbare Mehrdeutigkeiten zu geben. Aus dieser Beobachtung läßt sich ein guter Test ableiten: Man wähle eine Vorlage mit Zahlen und Buchstaben und prüfe, ob eine ausreichende Unterscheidung möglich ist. (Ein weiterer Test dieser Art betrifft das kleine "l" und die Zahl "1".) Da man bei vielen Anwendungen in der anschließenden Verarbeitung Buchstaben und Zahlen unterscheiden muß, hat dieser Test wesentliche Bedeutung für die Qualität des auszuwählenden Systems.

Alphabet-Abhängigkeit von Eigenschaftsbeschreibungen



Mehrere gleiche Buchstaben aus derselben Vorlage. Man erkennt deutlich die Variabilität, die beim Scannen mit diskretem Raster auftritt.

Ein weiteres Einschätzungskriterium beim Vergleich von "pattern matching" und "feature recognition" ergibt sich aus der Alphabet-Abhängigkeit dieses Ansatzes. Eigenschaftsbeschreibungen können nur für bestimmte Zeichensätze getroffen werden. Der Algorithmus muß sich deshalb auf nationale Alphabete beschränken und ist nicht mehr uneingeschränkt international flexibel. Das erklärt die manchmal erstaunlichen Ausfälle der für das lateinische Alphabet optimierten Erkennungsalgorithmen im Bereich der nicht-lateinischen Schriften (z.B. griechisch oder kyrillisch). Der denkbare Ausweg, die Eigenschaftsbeschreibungen auf eine Vielzahl von Alphabeten zu erstrecken, wirft (abgesehen von dem schwierigen Problem der Trennschärfe in großen Mengen von Eigenschaftsbeschreibungen) zeitkritische Probleme der Verarbeitungskapazität auf.



"Omnifont" – oft nur "Multifont"

Auf den eben angesprochenen Punkt muß man übrigens auch dann achten, wenn die Vorlage zwar Buchstaben des lateinischen Alphabets enthält, diese aber mit vielfältigen Akzenten versehen sind (die z.T. nicht einmal im ASCII-Zeichensatz vorkommen). Hier versagen zahlreiche OCR-Systeme und scheiden damit auf Grund eines einfachen Kriteriums als unbrauchbar aus. Wenn sie sich dann in der Werbung noch mit dem schmückenden Beiwort "omnifont" präsentieren, ist eine kritische Grenze erreicht.

Guter Test: Fraktur-Vorlagen

Ein guter Test in Richtung auf "Musterflexibilität" ist aus den genannten Gründen eine Vorlage in einem nichtlateinischen Alphabet. Als besonders brauchbares Testobjekt haben sich aber wegen der von der lateinischen Schrift stark abweichenden Buchstabenform auch gedruckte Fraktur-Vorlagen erwiesen. Man erkennt bei nicht-lateinischen Schriften relativ schnell, ob der betreffende Algorithmus "lateinische" Normalformen erwartet oder nicht. (Drei Algorithmen zur Zeichenerkennung, die anschaulich den Unterschied von "pattern matching"- und "feature recognition"-Techniken klarmachen, sind dargestellt bei: Elaine Rich, Artificial Intelligence, New York 1983, S. 11–16. Dort werden auch Vor- und Nachteile beider Methoden gegeneinander abgewogen.)

Einige wichtige Beurteilungskriterien

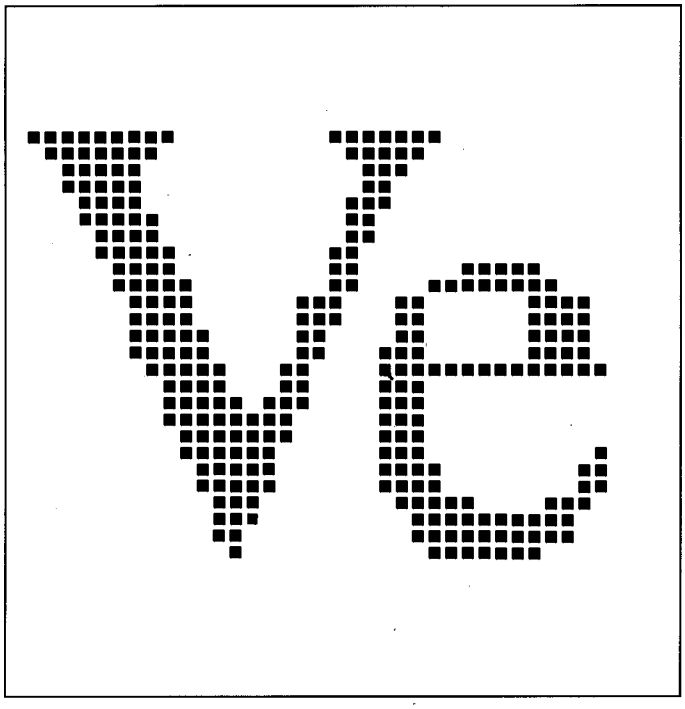
"Trainierbare" und "Nicht-trainierbare" Systeme

Eine erste, für die Beurteilung von OCR-Systemen wichtige Unterscheidung ist die zwischen "trainierbaren" und "nicht- trainierbaren" Systemen.

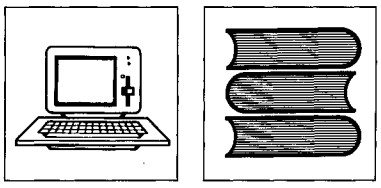
"Nicht-trainierbare" Systeme sind auf feste Schriftvorlagen hin konzipiert. Sie sind in der Lage, diese Schriftvorlagen genau in der vorgegebenen Größe und Gestalt zu "erkennen", nicht aber beliebige andere. Ein bekanntes Produkt dieser Orientierung war der AEG-Blattleser. Inzwischen werden einige "nicht-trainierbare" Systeme auch mit unterschiedlichen Größen der Schrift fertig, für die sie ausgelegt sind. Üblicherweise gibt die Produktbeschreibung (meist in Punkt als Schriftgrößenmaß) an, wie weit sich dieser "Toleranzbereich" erstreckt.

Die Trainingsphase

"Trainierbare" Systeme werden in einer "Trainingsphase" mit unbekannten Schriftmustern konfrontiert, die sie dann dieser Einarbeitungszeit (deren Ergebnisse gespeichert werden können) als "bekannt" verarbeiten. Verschiedentlich gibt es inzwischen auch die Vertriebspolitik, mit trainierbaren Systemen feste "Trainings" für bestimmte Zeichensätze vorzubereiten und diese dann dem Anwender zur Verfügung zu stellen. Das ändert aber nichts an der prinzipiellen Unterscheidung zwischen "trainierbaren" und "nicht-trainierbaren" Systemen. Prototyp der "trainierbaren" Systeme ist die eingangs erwähnte "KDEM".



Eine Unterschneidung am Beispiel der Buchstabenkombination Ve. Die rechte Linie des Buchstabens "V" reicht über den linken Teil des Buchstabens "e". Es existiert keine senkrechte Trennlinie zwischen den beiden Buchstaben.



Verarbeitungsgeeignete Schriftarten

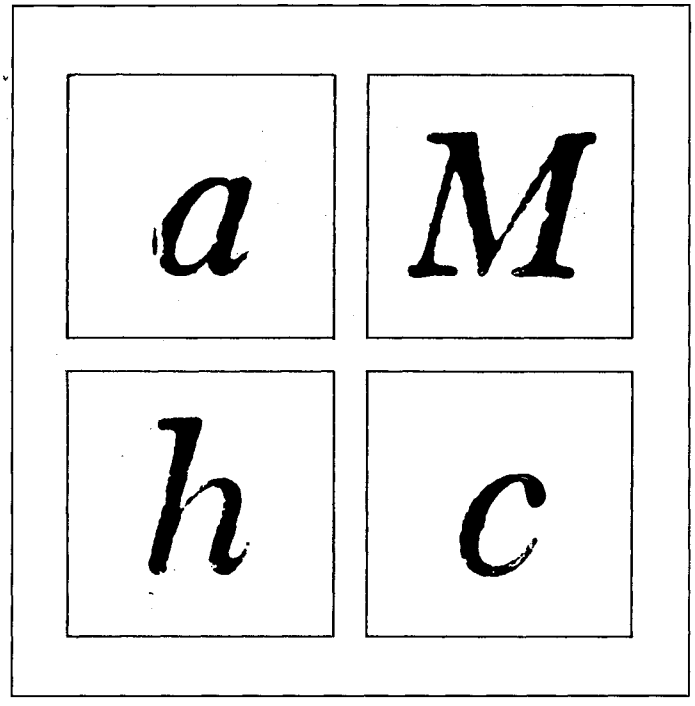
Die zweite bei der Beurteilung eines Systems zu beachtende Unterscheidung betrifft die Schriftart der Vorlage. Alle bisherigen Systeme zeigen (sofern sie überhaupt in der Lage sind, Schreibmaschinenschriften und Buchschriften zu verarbeiten) deutlich unterschiedliche Ergebnisse bei diesen beiden Vorlagenarten. Das hängt in erster Linie damit zusammen, daß der Satz viel mehr Feineinstellungen zu Buchstabengröße und Buchstabenabstand erlaubt, als dies bei Schreibmaschinenvorlagen der Fall ist. (Als ein besonders charakteristisches Beispiel seien nur die sog. Unterschneidungen angeführt).
Zwar gibt es als Sonderfall auch die in Proportionalschrift geschriebene Schreibmaschinenvorlage. In diesem Falle wird aber nur der Buchstabenabstand verändert. Deshalb sind selbst im Vergleich damit die nuancierten Gestaltungsmöglichkeiten in Buchvorlagen noch als weitaus individueller einzustufen.

Verarbeitungsgeschwindigkeit

Ein weiteres wichtiges Beurteilungskriterium, auf das viele Anwender zu recht großen Wert legen, ist die Verarbeitungsgeschwindigkeit. So einig man sich über die Bedeutung dieser Eigenschaft ist, so sehr herrscht doch oft auch Unklarheit über die Einschätzungsmethoden, die hier am Platze sind.

Die Parameter

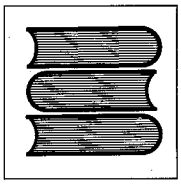
Die Geschwindigkeit wird bestimmt durch die mechanischen Grenzen des Scanners und die Verarbeitungskapazität der Zeichenerkennungssoftware. Dabei ist es (was oft nicht erwartet wird) bei schwierigen Vorlagen so, daß die "mechanische" Scan-Leistung nicht voll ausgeschöpft werden kann, weil die Zeichenerkennungssoftware mit der Abarbeitung des Stroms der ankommenden Informationen nicht im gleichen Tempo fertig wird.
Zwar ist es (die Fehlerrate bei der Erkennung vorläufig einmal ausgeklammert) immer noch so, daß das Scannen geschwindigkeitsmäßig dem Eingeben von Hand überlegen ist. Trotzdem summieren sich die reinen Scan-Zeiten bei einigen Systemen im Falle größerer Projekte beachtlich, was genaue Tests zu diesem Punkt als geboten erscheinen läßt. Bei diesen Tests ist es übrigens wichtig darauf zu achten, wie viele Buchstaben sich auf einer Seite befinden. Denn die "Lese"-Geschwindigkeit (zu unterscheiden von der reinen Grauwerterfassung der Vorlage) richtet sich aus den eben genannten Gründen nach dieser Zeichenanzahl und ist deshalb nicht etwa pro Blatt als gleich anzusetzen. Deshalb verdienen alle die Werbeaussagen eine kritische Prüfung, die ohne genaue Angabe zur verarbeiteten Zeichenanzahl pro Zeiteinheit mit Angaben wie "x Blatt in y Minuten" werben. Ein weiteres kommt hinzu: Die Verarbeitungszeit hängt auch von der Qualität der Vorlage ab. Je unregelmäßiger die Schriftkonturen, desto mehr Vergleiche werden der Zeichenerkennungssoftware (unabhängig vom gewählten Algorithmus) abverlangt. Auch hierfür sei ein Test empfohlen:

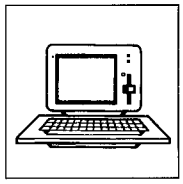


Beim Bleisatz entstehende Unregelmäßigkeiten in starker Vergrößerung.

Die Qualität der Vorlage

Eine mit Bleisatz gedruckte Vorlage. Die optisch ansprechenden Qualitäten des Bleisatzes hängen (u.a.) mit den individuellen Unregelmäßigkeiten zusammen, die mit dieser Drucktechnik verbunden sind. Für den hier empfohlenen Test relevant sind die Unregelmäßigkeiten





der Buchstabenkonturen, die dadurch entstehen, daß der Buchstabe durch Druck einer metallischen Kontur auf das Papier erzeugt wird (vgl. dazu die vergrößerten Abbildungen einiger mit Bleisatz gedruckter Buchstaben).

Für die Beurteilung eines Systems muß man auch auf die Zeichensatzkapazität achten. Um sich die Bedeutung dieses Kriteriums klarzumachen, kann man annäherungsweise folgende Berechnung anstellen: Das deutsche Alphabet umfaßt (ohne die Sonderzeichen) 30 Buchstaben, die als Groß- oder Kleinbuchstaben auftreten können. Pro Buchstaben speichern alle gängigen Zeichenerkennungssysteme mehrere Muster ab. Bei der KDEM heißen diese Muster MPD's (für "Multiple Property Description"). Für Schreibmaschinenschriften kommen effektive Algorithmen schon mit zwei Mustern pro Buchstaben aus. Bei Druckschriften kann es nötig sein, deutlich mehr Muster pro Buchstabe abzuspeichern. Damit ist man schon für den einfachsten Fall im Deutschen bei mindestens 120 Buchstaben-Mustern angekommen, die als Mindestausstattung unverzichtbar sind. Nimmt die Vorlagenqualität ab, so verschlechtert sich die Muster-/Buchstaben-Relation meist drastisch.

Zeichensatzkapazität

Berücksichtigt man zusätzlich, daß mehrere Schriftarten gemischt sein können (bei Spezialanwendungen etwa Kyrillisch, Griechisch und Lateinisch), so sieht man unschwer, daß die Zeichensatzkapazität hohe Reserven aufweisen muß, wenn man nicht auf einen engen Bereich optimaler Normalfallanwendungen festgelegt sein will. Beim Test eines Systems sollte man die Grenze der Zeichensatzkapazität dadurch auszuloten versuchen, daß man eine Vorlage mit sehr reichhaltiger Schriftmischung verwendet. Häufig stößt man schon bei ca. drei Druckschriften an systembedingte Grenzen. Dabei zeigt sich dann auch, daß es ein wichtiges Beurteilungsmerkmal ist, ob Systeme zu erkennen geben, mit wievielen (und welchen) Mustern sie intern arbeiten. Manchmal ist dies nämlich für den Normalbenutzer von außen nicht feststellbar. Die Systeme, die mit Meldungen wie "Kapazitätsgrenze erreicht" o.ä. die Begrenzung dokumentieren, sind unter dem hier relevanten Beurteilungsaspekt als offen dokumentiert anzusehen, selbst wenn es natürlich wenig erfreulich ist, erst nach längerer Arbeit auf diese Begrenzung zu stoßen. Es gibt aber auch den Fall, daß diese Grenze nicht nach außen in Erscheinung tritt, weil einfach (stillschweigend) keine neuen Muster mehr akzeptiert werden. Dann ist der Benutzer auf indirekte Einschätzungskriterien angewiesen. Ein Kriterium dieser Art ist etwa darin zu sehen, daß ab einem bestimmten Punkt hinzukommende weitere Schriften drastisch schlechter erkannt werden als die vorhergehenden.

Multifont-Vorlagen

Das letzten Endes entscheidende Beurteilungskriterium ist die Erkennungsquote. Damit war und ist es (entgegen mancher wohlklingender Werbeaussage) bei vielen jetzt günstig zu preiswerten Scannern angebotenen Systemen nicht weit her. Zu einem der Niedrigpreis-Systeme schrieb Steve Ciarcia in BYTE (August 1987, S. 70) als Antwort auf eine Leseranfrage, die ein unberechtigtes Verschweigen dieser preiswerten Systeme durch die Presse vermutet hatte:

"I don't think there is a conspiracy to suppress character readers. The technology is still in its infancy and has been quite limited until recently. Test reports on the (...) unit complained that very few type (!) fonts are readable at all, and the error rate is high in any case."

Was 1987 galt, ist auch heute noch im Niedrigpreis-Bereich eine reale Möglichkeit:

"... simply does not do a very good job. All too often, it either fails to read characters or, even worse, reads them incorrectly. It had particular trouble differentiating between lower- and uppercase letters" (Stanford Diehl/Howard Eglowstein, Tame the Paper Tiger, BYTE, April 1991, S. 220ff., 232).

Erkennungsquote

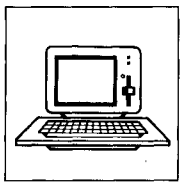
Im übrigen muß man sich bei der Beurteilung der Erkennungsquote gegen einen Fehleindruck wappnen, den wohlklingende Prozentangaben leicht hervorrufen. Mancher ist schon positiv für ein System eingenommen,

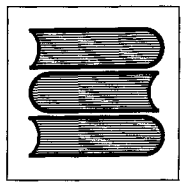
wenn als Erkennungsrate "98%" angegeben werden. Bei näherer Betrachtung erweist sich aber selbst dieser hohe Wert als problematisch, wie eine einfache Rechnung zeigt. Angenommen, eine Zeile bestehe aus 50 Buchstaben. Dann entfällt bei einer Erkennungsquote von 98% auf jeweils eine Zeile ein Fehler. Kalkuliert man die dann anfallende Nacharbeitszeit (man muß die Fehler ja auch noch ausfindig machen!), so zeigt sich oft, daß das manuelle Erfassen dem maschinellen vorzuziehen ist.

Erkennungsrate 98%: 1 Fehler/Zeile bei 50 Zeichen/Zeile

Als Beurteilungskriterium gerade für juristische Projekte mit entscheidend (aber oft übersehen) ist die Frage, ob das OCR- Programm in der Lage ist, Fontwechsel zu markieren. Das ist deswegen entscheidend, weil bei einer typographisch gut gestalteten Druckvorlage die Fontwechsel zumeist auch eine Informationsbedeutung haben. So werden beispielsweise Eigennamen häufig in Kapitälchen gesetzt, Zitate kursiv, Urteilsüberschriften in einem anderen Font als der Urteilstext, die im Urteilskopf stehende Normenkette erscheint halbfett

Fontwechsel-Markierung

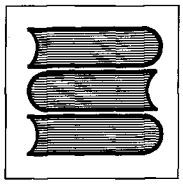


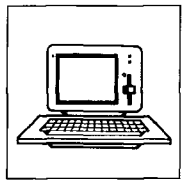


usw. Wenn nun ein OCR-Programm nur in der Lage ist, die Buchstaben zu erkennen, nicht jedoch festzuhalten, aus welchem Font bzw. welcher Auszeichnungsart sie stammen, geht die dadurch zum Ausdruck gebrachte Information verloren. Man kann dann, um bei den angeführten Beispielen zu bleiben, nicht mehr automatisch ein Namens- oder Zitatenregister erstellen, die Überschriften für eine Liste extrahieren oder ein Paragraphenverzeichnis anfertigen.
Man fragt sich angesichts der Bedeutung der Fontwechselmarkierung unwillkürlich, warum zahlreiche Programme diese Eigenschaft nicht aufweisen. Die Ursache ist bei den oben diskutierten Erkennungsmethoden zu suchen. Wenn ein Programm eigenschaftsorientiert arbeitet, enthält es sehr stark verallgemeinerte Beschreibungen von Buchstabeneigenschaften. Nun weist aber beispielsweise ein "A" in der einen Schriftart auf dieser Abstraktionsebene meist dieselben Eigenschaften auf wie ein "A" in einer anderen Schriftart, so daß beide Buchstabenvarianten zwar erkannt, aber nicht unterschieden werden können. Im Unterschied dazu bieten trainierbare Systeme, deren OCR-Methode mindestens eine "pattern matching"-Komponente aufweist, die Möglichkeit, Fontwechsel zu "erkennen" und zu markieren.

Die Situation bei Preis und Leistung: Ein praktisches Resume

Vor dem Hintergrund der eben diskutierten Beurteilungskriterien kann man die gegenwärtige Situation auf dem Gebiet der Zeichenerkennungssysteme mit Blick auf das Preis- Leistungsverhältnis folgendermaßen zusammenfassen.
Wenn die Vorlagen guter Qualität sind und es sich um Schreibmaschinenschriften handelt, kann man mit recht preiswerten Systemen (knapp unter/knapp über 1.000$) zu recht akzeptablen Ergebnissen kommen. Der zitierte BYTE-Test resumiert das entsprechende Anforderungsprofil als "if your needs are less demanding" (a.a.O., S. 238) und favorisiert in dieser Preisklasse das Programm "WordScan Plus" von Calera. Die eigenen Erfahrungen gehen, was das nötige Budget für optimale Zeichenerkennung angeht, in eine ähnliche Richtung und stützen auch die Schlußfolgerung, die Diehl und Egloffstein bezogen auf den amerikanischen Markt für das gehobenste OCR- Anforderungsprofil ziehen: Sie empfehlen eine Investition von 20.000$ (das System ist die Kurzweil K5200) und fügen hinzu: "It's a hard call" (a.a.O., S. 238).
Aus der Sicht des europäischen Marktes kann man in diesem Leistungsbereich noch ein von BYTE nicht getestetes System (OPTOPUS) mit auf die Evaluationsliste setzen. So resümiert der Rundbrief der Arbeitsgemeinschaft philosophischer Editionen: "Einhellig werden das Kurzweil-System sowie OPTOPUS als Systeme der Spitzenklasser herausgestellt, deren Preis (inkl. Hardware) allerdings eine Größenklasse über den meisten anderen Systemen liegt, ...". (Prof. Dr. N. Henrichs, Newsletter 91 der Technik-Kommision der Arbeitsgemeinschaft philosophischer Editionen) OPTOPUS hat unter anderem bei der Produktion juristischer CD-ROM's (etwa zum Einigungsvertrag und zum weitergeltenden Recht der DDR – auf der Grundlage der DDR-Gesetzblätter ein wahrer OCR-Härtetest) seine Leistungsfähigkeit unter Beweis gestellt. Es erfordert zwar kein Investitionsvolumen von 20.000$, sondern (ohne Scanner) "nur" eines von knapp 20.000.-DM. Aber "a hard call" bleibt das immer noch. Man sollte sich aber bei der Meinungsbildung nicht allein an der für das OCR-System aufzuwendenden Summe orientieren, wenn es um eine Investitionsentscheidung geht. Wichtiger sind die Kosten, die insgesamt bei einem OCR-Projekt entstehen. Auch diesbezüglich muß man Diehl und Eglowstein recht geben: "Although it costs a bundle, you'll recover a good bit of that expense by avoiding many worker-hours of proofing, rescanning and trouble-shooting" (a.a.O., S. 238).
Für die Auswahl des "richtigen" OCR-Systems gibt es – das sollte nach allem Dargelegten deutlich geworden sein – kein Patentrezept. Selten wird der Weg vor einer Entscheidung an einem gründlichen Test vorbeiführen. Ein solcher Test muß das Erfassen eines gut gewählten Text-Samples samt gründlichem Korrekturlesen einschließen. Ist das Text-Sample zweckmäßig ausgesucht, verfügt man nach dem Test auf jeden Fall bereits über eine brauchbare Volltextdatenbank. Selbst wenn der Aufwand dafür dann höher war als das anderweitige Erfassen der Texte, rechtfertigt der doppelte Zweck den Aufwand.





Menschliches und maschinelles Lesen: Eine wahre Geschichte zum Schluß

An einer OCR-Station wird Text erfasst. Der Bearbeiter legt eine Buchseite auf den Scanner, der Scanner tastet die Seite ab, übermittelt das Bild an den Rechner, die OCR-Software setzt das Bild in Buchstaben um, die auf dem Bildschirm erscheinen und dort vom Bearbeiter zu Kontrollzwecken gleich mit betrachtet werden. Ein Jurist betritt den Raum und fragt, warum man den diesen ganzen Umweg mache und das Buch nicht gleich selbst lese.
Meist stellt sich beim Erzählen dieser Anekdote Heiterkeit ein. Aber in einer Hinsicht macht der "iurista in fabula" doch auf etwas Wesentliches aufmerksam: Die Maschine liest nicht - das wird der Mensch immer selbst tun müssen.