FineReader Professional

Optische Zeichenerkennung mit dem Computer

Rainer H. Bubenzer, Hamburg

NOTFALL & HAUSARZTMEDIZIN 2004; 30: 102–103

Lesen ist eine grundlegende Kulturtechnik, die nach zumeist langwierigen Lernvorgängen und aufwändigem Training schließlich von vielen Menschen beherrscht wird. Ein Teilaspekt des Lesens, nämlich die Umsetzung analog-visueller Reize (Bilder, Zeichen, Buch- oder Zeitschriften-Seiten) in digitale Informationen (entsprechend neuronaler ZNS-Signalmuster) inklusive Mustererkennung und -zuordnung (Buchstaben, Text) wird heute auch von elaborierter Software beherrscht, zumindest einigermaßen. Hauptzweck: Umsetzung gedruckter Informationen in Büchern, Zeitschriften oder Zeitungen in digital speicherbare und vor allem verarbeitungsfähige Daten.

Das technische Prinzip der optischen Zeichenerkennung (Optical Character Recognition, OCR) baut auf bereits ab dem 19. Jahrhundert entwickelten Technologien der technischen Faksimileübertragung („Telefax"; Alexander Bain/Gr. Brit., 1843), des Fotokopierens („Xerographie", Chester F. Carlson/USA, 1938) oder des Computers („Analytical Engine", Charles Babbage/Gr. Brit., 1838; „Z3", Konrad Zuse/Deutschl., 1941) auf: Ein „elektronisches Auge" (z.B. Flachbett-Scanner, Mehrseiteneinzugs-Scanner, Digital-Video- oder Fotokamera) wandelt die aufgelegten Seiten eines Buches oder einer Zeitschrift in eine digitale Bild-Repräsentation um. Dieses so genannte Computer-Image entspricht einer Kopie, wie sie von Faxgeräten oder Fotokopierern erstellt wird, ist jedoch digital und damit auf Datenträger speicherbar und anschließend von geeigneter Software weiter zu verarbeiten. Die elektronische OCR-Datenverarbeitung entspricht einer Simulation der beim Menschen von Retina und Großhirnrinde vollbrachten Mustererkennung und Informationszuordnung. Software wie das in diesem Beitrag vorgestellte Programm „Fine Reader Professional" versucht, in digitalen Bildern enthaltene Buchstaben- und Textinformation zu identifizieren und in die entsprechenden Buchstaben und Texte in der digitalen Kodierung von Computer-Standardzeichensätzen zu übersetzen. Um sich die Erkennungs-Arbeit zu erleichtern, erkennen hoch entwickelte OCR-Programme heute auch Strukturinformationen eines Bildes (z.B. Absätze, Überschriften, Bildunterschriften, Seitenzahlen usw.).

Folgende Programme entsprechen etwa der Leistungsklasse von FineReader

Omnipage Pro 14	www.scansoft.de
Readlris Pro 9	www.irislink.com
TextBridge Pro 11	www.scansoft.de

(siehe Vergleichstest in c't 2/04: www.abitz.com/ocr/ct.0204.142-147.pdf)

Contra OCR

Neben der wissenschaftlichen Herausforderung, auch Computern das Lesen beizubringen, gab und gibt es zahlreiche weitere Motive, die EDV-gestützte optische Zeichenerkennung (weiter) zu entwickeln. Ein idealistisches, aber schon fast wieder aufgegebenes Motiv seit den 60er Jahren ist die vollständige Digitalisierung der auf Papier geschriebenen und gedruckten Informationen der Menschheit. Aufgegeben deshalb, weil heute zum einen das Volumen der zu digitalisierenden Werke alle Grenzen der bestehenden Verarbeitungskapazitäten (auch die Ressourcen bei den Bibliothekaren) sprengt, zum anderen bis heute nicht bewältigte Erkennungs-Probleme bei handgesetzten (oder gar handgeschriebenen) Büchern und Zeitungen vor zirka 1870–1890 auftreten. Worüber die „OCR-Szene" jedoch nur hinter vorgehaltener Hand spricht, ist der Hauptgrund gegen eine breit angelegte Digitalisierung: Nämlich die „Vergänglichkeit der digitalen Daten". Auf Leder, Pergament, Papyrus oder Papier geschriebene Informationen können Jahrtausende überdauern, wobei der Informationsverlust bei geeigneter Lagerung (z.B. versiegelte Tonkrüge in den Höhlen von Qumran) selbst nach einigen Jahrtausenden minimal ist. Bei digitalen Speichermedien ist manchmal schon nach zehn Jahren nicht mehr zu garantieren, dass noch geeignete Lesegeräte oder Software vorhanden sind, die Speichermedien noch intakt oder die Informationen erhalten geblieben sind. Kurzum: Eine digitale Bibliothek von Alexandria, die das Wissen der Welt umfasst, wird es vermutlich niemals geben.

Pro OCR

Wozu aber dann digitale Texterkennung (in der Medizin)? Ein Pro für die Technologie sind Unterstützung der Wissenschaft und Optimierung der ärztlichen Tätigkeit. Umfangreiche Schlüsselwerke der Medizin, digital aufgearbeitet, stehen der elektronischen Datenverarbeitung zur Verfügung, zum Beispiel können Informationen weitaus schneller gefunden oder verknüpft werden als bei dem papierenen Pendant (Beispiele: Pschyrembel, Lexikon der Geschichte der Medizin, komplette Werke von Samuel Hahnemann auf CD-ROM u.a.). Die Digitalisierung ist auch ein sinnvoller Ersatz für den teuren Neudruck me-

dizinhistorisch oder für eine bestimmte Fachrichtung bedeutsamer Werke. Einige Disziplinen der Medizin haben schon vor Jahren damit begonnen, ihr in Zeitschriften und Büchern gesammeltes Wissen, beispielsweise der Homöopathie, Psychiatrie oder Phytotherapie, digital aufzuarbeiten. Beispiel: Die berühmte Arzneimittellehre des Dioskurides als Folge von knapp 600 digitalen Scans, wurde mittels Texterkennung durchsuchbar gemacht (www.heilpflanzen-welt.de/dioskurides; suchen Sie doch mal nach „Elephant“).

OCR in der Praxis

Weitere Gründe für den Einsatz von OCR sind Erfassung und Archivierung von papiergebundenen Informationen, deren Aufbewahrung zu teuer (Aktenlagerung) oder deren klassische Verteilung (Fotokopie, Versand) zu aufwändig wäre. Beispiele sind der regelmäßig zu erstellende Pressespiegel einer großen Klinik, die Digitalisierung von Vortragsmanuskripten von Kongressen oder Buchausschnitten, die als Element zum Beispiel einer Infobroschüre an Patienten abgegeben werden. Auch Einscannen und Texterkennung von standardisierten Befunden, Gutachten oder Labordaten für die digitale Patientenakte sind mit OCR-Software realisierbar.

FineReader: Eins der besten OCR-Programme

FineReader 7.0 Professional vom russischen Hersteller Abbyy Software House ist eines der ausgereiftesten Texterkennungs-Programme für Desktop-Systeme. Die Software lässt sich einfach von CD-ROM installieren, vor der Erstbenutzung ist eine Produktaktivierung erforderlich (Internetverbindung nötig). FineReader läuft mit allen TWAIN-kompatiblen Geräten, also fast allen Scannern oder vielen Digitalkameras (Liste kompatibler Geräte: www.abbyy.de). Mit der Benutzeroberfläche von FineReader sind grundlegende OCR-Aufgaben so einfach oder kompliziert, wie man es möchte. Direkt unter der Standardmenüleiste befindet sich eine Werkzeugleiste mit fünf großen Symbolen. Über eines dieser Symbole können mit einem Mausklick etliche Scan- und Lesefunktionen gestartet werden, was besonders für Anfänger ideal ist, die bei den vielen OCR-Optionen noch Unterstützung benötigen. Mit den anderen vier Schaltflächen können einzelne Schritte gesteuert werden: Scannen, Lesen, Rechtschreibung prüfen und Speichern der erkannten Seiten als Dokument (z.B. als Word- oder PDF-Dokument). Erfahrenere OCR-Anwender werden feststellen, dass die zusätzlichen Werkzeugleisten schnellen Zugriff auf erweiterte Funktionen bieten, unter anderem Anpassung des Erkennungsbereichs, Bildbearbeitung oder Drehen des Bildes. Der Zugriff auf komplexere Eigenschaften des OCR-Programms ist gegenüber Vorgängerversionen deutlich transparenter und einfacher geworden. FineReader 7.0 Professional für Texterkennungssoftware für Microsoft Windows ab Version 95 kostet 129 Euro, als Updateversion 89 Euro.

Umwandlung komplexer PDF-Dokumente zu Text

Seit Version 6 erlaubt die Software die Texterkennung bei PDF-Dokumenten. Doch erst mit der jetzt vorgelegten Version ist eine befriedigende OCR auch bei komplex aufgebauten und mit der Exportfunktion von Adobe Acrobat meistens nicht korrekt in Textdokumente umzuwandelnden PDF-Dateien möglich. Nach vielfältigen Vergleichen der OCR-Ergebnisse von FineReader und generischen PDF-Exportprogrammen (z.B. Magellan, www.bcl-technologies.com u.a.) zeigt sich: Die Text- und Struktur-Erkennungsleistung bei komplex gestalteten PDF-Dokumten (Mehrspaltensatz, spaltenübergreifende Bilder u.a.) ist bei FineReader insgesamt mit Abstand am besten, selbst wenn immer wieder unerklärliche Erkennungsfehler auftreten.

Formfiller auch für Formulare in Arztpraxis, Klinik oder Apotheke

FineReader lässt sich direkt in Microsoft-Office-2003-Anwendungen einbinden. So kann ein Dokument zum Beispiel mitten in ein Dokument in Word 2003 eingescannt und der erkannte Text dann mit dessen Rechtschreibprüfung korrigiert werden. Als Bonus erhalten registrierte Anwender „FormFiller“, ein Programm zum Dateneinlesen aus Formularen. Nach Festlegung der Datenfelder eines Formulars kann die Software aus einer Serie gleicher Formulare alle Daten extrahieren und beispielsweise im Excel-Format speichern. Praktisch wenn in Praxis, Klinik oder Apotheke immer wieder gleiche Formulardaten erfasst werden müssen (Abbyy liefert entsprechende Software-Schnittstellen – FineReader SDK).

Medizinisches Korrekturwörterbuch

Noch ein Wort zur Rechtschreibkorrektur bei mit FineReader erkannten Texten: Diese ist einfach möglich, bietet dabei ständig einen Blick aufs Originalbild und glänzt in der Version 7 jetzt mit Korrekturlexika alter oder neuer deutscher Rechtschreibung, jeweils mit oder ohne Medizinbegriffe. Diese Funktionalität übertrifft weit die Korrekturfähigkeit zum Beispiel der MS-Officekorrektur und erhöht die Erkennungsgenauigkeit bei medizinischen Texten erheblich. Noch besser: Spezielle Schriften, die FineReader nicht bekannt sind oder bei denen die Erkennungsgenauigkeit unbefriedigend ist, können bei der Erkennung trainiert werden – so wurden bei uns mehrere in Fraktur gesetzte, medizinische Werke nach diesem Training mit hoher Genauigkeit erkannt. Schade nur, dass der Hersteller die seit Jahren angekündigte integrierte Frakturschrift-Erkennung immer noch nicht realisiert hat – da nützen denn auch hunderte von eingebauten Erkennungssprachen oder exotischen Schriften nichts (für Bibliotheken gibt es eine spezielle Beta-Version zur Frakturerkennung).

Anschrift des Verfassers
Rainer H. Bubenzer
multi MED vision
Borselstraße 9
22765 Hamburg
Fax: 040/41 91 28 77
Rainer@Bubenzer.com