Abbyy FineReader Professional 6.0

Martin Schönhacker

CD ROM (593 MB); ca. Euro 129,00

Bei Abbyy FineReader handelt es sich um einen der prominenteren Vertreter in der Landschaft aktueller Programmpakete zur Texterkennung. Die Entwicklung geht entsprechend flott vor sich, und so sei gleich zu Anfang gesagt, dass die Version 6.0 bei Erscheinen dieser Besprechung schon wieder zum „alten Eisen" gehört. Aber das Prinzip ist natürlich gleich geblieben, die Erkennungsleistung der neuen Version ist mindestens so gut, und mit etwas Glück kann man die nunmehr alte Version 6.0 sogar als Schnäppchen zum Sonderpreis erwerben.

Zur Installation werden ungefähr 90 MB freier Festplattenspeicher benötigt. Die Auswahl an optionalen „Erkennungssprachen" ist dabei fast unglaublich vielfältig: 30 „Hauptsprachen" (z.B. Deutsch, Englisch, Griechisch), 86 „Erweiterungssprachen" (z.B. Latein, Rätoromanisch, Zulu), 7 „Formale Erkennungssprachen" (z.B. Basic, C/C++, Java, aber auch einfache chemische Formeln) sowie 4 „Konstituierende Erkennungssprachen" (z.B. Esperanto) stehen zur Auswahl und können im Prinzip auch alle zugleich installiert werden. Für jeden Textblock kann man dann im Erkennungsprozess getrennt auswählen, in welcher Sprache er verfasst ist.

Die Grundfunktion des Programms ist immer die gleiche: Ein Dokument liegt als Grafik vor und soll wieder in Text umgewandelt werden, der in der Folge zum Beispiel durchsucht, neu formatiert oder auf andere Weise bearbeitet werden kann. Idealerweise wünscht man sich zu diesem Zweck oft nicht bloß den unformatierten Klartext, sondern eine Datei, die möglichst dicht am ursprünglichen Original ist. Abbyy FineReader bietet hier zahlreiche Optionen an.

Zunächst kann das Eingabemedium fast beliebig gewählt werden. Natürlich gibt es die Möglichkeit, eine oder mehrere Seite(n) direkt über einen Scanner zu importieren. Aber auch mit Grafikdateien in verschiedenen gebräuchlichen Formaten kommt das Programm zurecht, wenn die Auflösung gewissen Mindestansprüchen genügt. Und schließlich kann man sogar PDF-Dateien öffnen, die in der Folge automatisch seitenweise in (übrigens ziemlich umfangreiche!) Grafiken umgewandelt und auf diese Weise der Texterkennung zugeführt werden, ohne das PDF-Format an sich decodieren zu müssen.

Dann folgt die Formaterkennung. Ein Algorithmus versucht, Blöcke mit bestimmten Eigenschaften auf der eingelesenen Seite zu identifizieren. Es gibt Textblöcke, Grafiken, aber auch Tabellen. Wenn das Programm kein gutes „Gefühl" für das Layout entwickelt haben sollte, kann man jederzeit manuell eingreifen und selbst Blöcke einfügen, ändern oder entfernen.

Der nächste und wichtigste Schritt ist die eigentliche Texterkennung. Hier wird versucht, in allen Text- bzw. Tabellenblöcken einzelne Zeichen zu identifizieren und wieder in Klartext zu verwandeln. Manchmal geht das nur mit einer gewissen Wahrscheinlichkeit, aber das wird einem auch nicht verschwiegen. Mögliche Fehler bzw. „unsichere" Stellen werden mit Farbcodes markiert und können in der Folge der Reihe nach durchkorrigiert werden. Dabei kann man Text einfügen, ändern, löschen, aber sogar auch formatieren.

Ist die Erkennung erst einmal so weit gediehen, fehlt nur noch die Umwandlung in ein Ausgabeformat. Hier werden populäre Programme wie Microsoft Word, Excel, aber auch die allgemeinen Formate RTF (für formatierten Text), HTML, CSV (für Tabellen), DBF (Datenbank), TXT (mit verschiedenen Codierungen) oder PDF unterstützt. Auf Wunsch kann ein installiertes Zielprogramm auch direkt gestartet werden, um die Qualität der Umwandlung unmittelbar überprüfen zu können.

Im Praxistest waren die Resultate teilweise erstaunlich. So wurde der eingescannte, naturgemäß mit Fremdwörtern gespickte Beipacktext eines Medikamentes in verblüffender Präzision mit nur zwei falschen Zeichen auf zwei Seiten erkannt. Auch beim Rückentext einer DVD, der weiß auf dunkelbraun gedruckt und daher scheinbar hoffnungslos war, hatte das Programm keine großen Schwierigkeiten.

Bei PDF-Dateien gab es andererseits (erstaunlicherweise, weil ja die Qualität des Ausgangsmaterials technisch makellos ist) einige Komplikationen, vor allem mit kursiv gedruckten Textpassagen. Hier wurden leider relativ oft benachbarte Buchstaben zusammengezogen und dann falsch erkannt.

Dagegen waren die Testresultate bei hochwertigen Digitalfotos von Zeitschriftenseiten (5 Megapixel für eine Doppelseite) beinahe schon erstaunlich zu nennen. Es konnten ca. 98% des Textes erkannt werden, obwohl nicht einmal alle Zeilen völlig gerade waren, weil die Zeitschrift sich in der Mitte gewölbt hatte und die Kamera nur in der Hand gehalten wurde.

Insgesamt präsentiert sich die Erfahrung ziemlich gemischt. Manche Resultate waren geradezu begeisternd, andere etwas enttäuschend. Das vorliegende Programm ist wohl eines der besten am Markt, aber daran erkennt man auch, dass der Markt noch einige wichtige Schritte vor sich hat, bis die Texterkennung zur Perfektion gelangt ist. An die Erkennung normaler Handschrift, auch wenn sie nur aus Blockbuchstaben besteht, ist mit diesem Programm übrigens noch nicht wirklich zu denken.

Für einigermaßen „normale" Texte in ordentlicher Qualität liefert Abbyy FineReader allerdings sehr gute Resultate, wenn man das Programm mit Sorgfalt und etwas Gefühl für die Erkennungsverfahren bedient. Eine manuelle Markierung an der richtigen Stelle kann wahre Wunder wirken. Unter der Bedingung, dass man bereit ist, sich mit dem Programm zu beschäftigen und selbst zum Resultat beizutragen, ist es also durchaus zu empfehlen. Schneller und abwechslungsreicher als stupides Abtippen ist es fast bei jedem Text.