OCR2

Rubrik: Lexikon

Der Begriff OCR kommt aus dem Englischen von „Optical Character Recognition“ und bedeutet so viel wie optische Zeichenerkennung oder auch einfach Texterkennung. OCR ist eine spezielle Technik, mit der zum Beispiel innerhalb von Bildern Texte automatisch erkannt werden können. Bilddateien sind dabei Rastergrafiken, die entweder eingescannt oder digital fotografiert wurden. Die Texterkennung mit OCR funktioniert dabei dreistufig. Zunächst erfolgt die Seiten- und Gliederungserkennung, bei der eine Bilddatei in verschiedene Bereiche unterteilt wird. Dazu gehören zum Beispiel Überschriften, Fließtexte, Bildunterschriften, die von den unwichtigen Teilen wie Weißräumen, Linien und Grafiken getrennt werden. Als Nächstes erfolgt die Mustererkennung, deren Qualität stark vom Kontrast der Vorlage abhängig ist. Nach einer Fehlerkorrektur auf Pixelebene und dem Mustervergleich sowie der Fehlerkorrektur auf Zeichen- und Wortebene können in vielen OCR-Programmen noch manuelle Korrekturen vorgenommen werden. Anschließend erfolgt die Codierung in das Ausgabeformat, zum Beispiel in eine Textdatei oder in eine Datenbank. Auch als HTML oder PDF können die Daten ausgegeben werden.

Die Qualität des Ergebnisses beim OCR-Verfahren ist von verschiedenen Faktoren abhängig, zum Beispiel von Umfang und Qualität der Muster-Datenbank, von den Wörterbüchern sowie von der Qualität der Layouterkennung und der Algorithmen zur Fehlerkorrektur. Auch Farben, Kontrast und Schriftart des Originals spielen eine Rolle.

Zum Einsatz kommt das OCR-Verfahren zum Beispiel, um Textinformationen aus Bildern weiter verarbeiten oder elektronisch durchsuchen zu können. Damit können Schriftstücke einsortiert, Bilder und PDFS durchsucht oder auch Gegenstände wie beispielsweise Kfz-Kennzeichnen erkannt und registriert werden. Auch als Hilfsmittel für blinde Menschen kommt das Verfahren zum Einsatz, zum Beispiel, um Texte einzuscannen und am Computer per Sprachausgabe vorlesen zu können.