

A-33

Zeichenerkennung (OCR)

„Pattern Matching“ vs. „Feature Matching“ – Ansatz im Vergleich

1. Einleitung: Warum OCR?

Optical Character Recognition (OCR) ist der Begriff für die Technik, die es einem Computer ermöglicht, Schriftzeichen zu erkennen.

Scanner gibt es bereits seit der 2. Computergeneration, als die englische Firma EMI (Electric and Musical Industries Ltd.) 1955 erstmals solche „elektronischen Bildabtaster“ herstellte.

Ein Scanner ist von Grund auf „dumm“, da er das eingescannte Datenmaterial nicht strukturieren kann, sondern lediglich als Bitmap wiedergibt. Die meisten Heimanwender benutzen einen Scanner lediglich dazu, Ferienfotos und sonstige Bilder einzuscannen. Dabei spielt lediglich die Auflösung und Farbwiedergabe des Scans eine Rolle. Wie steht es nun aber, wenn wir ein mehrseitiges Textdokument einscannen und danach auf dem PC weiterverarbeiten wollen?

In diesem Fall nützt uns ein Bitmap ziemlich wenig. Auch eine Auflösung von 600dpi tut hier nichts zur Sache. Was wir brauchen ist ein Programm, das die Schriftzeichen analysieren und erkennen kann. Der Computer benötigt also eine gewisse Spur von Intelligenz, um dies zu bewerkstelligen.

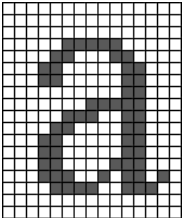
Heute ist die Ziffern-, Text- und Handschrifterkennung schon ziemlich weit fortgeschritten. Währenddem die Texterkennung eine Erkennungsrate von ca. 97-99% hat, werden einzelne Ziffern der marktüblichen OCR-A und OCR-B Schriften bereits mit 100% Fehlerfreiheit erkannt. Bei der Handschrifterkennung haben wir mittlerweile eine Trefferquote von 90% im günstigsten Fall erreicht.

Gründe für die Texterkennung sind vielseitig:

- ✓ Weiterverarbeitung von Text
- ✓ Geringe Datenmengen von Text im Vergleich zu Bilddaten
- ✓ Interaktion Computer-Mensch:
Schriftzeichen sind vom Menschen lesbar im Gegensatz zu Strichcodes u.ä.

Im Folgenden sollen die zwei Techniken der automatischen Zeichen- und Mustererkennung erläutert und miteinander verglichen werden: Pattern-Matching vs. Feature-Matching.

2. Pattern-Matching vs. Feature-Matching



Grundsätzlich versuchen die OCR-Programme zunächst, die einzelnen Zeichen der nach dem Scannen als Bitmap vorliegenden Seite als einzelne Zellen zu erfassen. Dabei suchen sie nach den Zwischenräumen. Mittlerweile sind die Erkennungsverfahren so ausgereift, dass auch aneinanderklebende Zeichenpaare oder -tripel zuverlässig getrennt werden.

← Beispiel eines eingescannten Zeichens als Bitmap [1] [2]

Hinter dem automatischen Umsetzen gescannter Texte stehen technisch aufwendige Verfahren. Zur Erkennung einzelner Zeichen gibt es im Wesentlichen zwei unterschiedliche Methoden. Die erste und bei den anfänglichen Entwicklungen überwiegend eingesetzte ist die „Musterübereinstimmung“ (Pattern Matching), die zweite nennt sich „Merkmalsübereinstimmung“ (Feature Matching).

Pattern Matching

Die Musterüberlagerung führt einen Vergleich mit vorher erfassten Referenzzeichen durch und versucht, die bestmögliche Übereinstimmung zu ermitteln. Meist wird diese durch die Anzahl übereinstimmender Bildelemente (PeI) bestimmt. Da es jedoch sehr unterschiedliche Schriften, Grössen und Ausprägungen sowie Attribute wie kursiv und fett gibt, erweist sich dies als recht problematisch. So ist es erforderlich, unterschiedliche Schriften zu trainieren und zur Erkennung den passenden Schrifttyp zu laden. Durch Training lässt sich die Erkennungsrate verbessern. Aufgrund des hohen Rechenaufwandes und der mangelhaften Flexibilität kommt Pattern Matching heute in Reinform jedoch kaum noch zum Einsatz.

Feature Matching

Grundlage der Texterkennung ist heute meist das Verfahren der Merkmalsübereinstimmung (Feature Matching), das jedes Zeichen auf seine Charakteristika hin untersucht. Die meisten Programme setzen dieses Verfahren ein. Konturen wie senkrechte und waagrechte Striche, Kreuzungspunkte und Bögen werden ermittelt und mit Musterformen verglichen. Hierbei spielen weder Schriftart oder Größe noch deren Attribute eine Rolle.

Somit erweist sich Feature Matching als weitaus zuverlässiger und universeller einsetzbar als das Pattern Matching.

Hier ein Beispiel von Feature Matching: [1] →

Der Buchstabe „a“ wird beschrieben durch einen Kreis, einen Balken rechts davon und einen Bogen oberhalb. Der Bogen könnte auch weggelassen werden, je nach Schriftart. Kann nun das OCR-Programm den eingescannten Buchstaben in diese Bestandteile aufteilen, wird das „a“ korrekt erkannt.



Feature Matching kann neben der Schriftzeichenerkennung natürlich auch in vielen anderen Bereichen angewandt werden. Als Beispiel kann hier der geografische Einsatzzweck genannt werden: [3]

Das Suchen nach Features stellt sich oft als nicht ganz so banales Problem heraus. Folgende Verfahren können dabei angewandt werden: Momente, Charakteristische Schnitte oder ‚Loci‘, Projektionen, Flächen, Schwerpunkte, Mittelwerte und mittlere quadratische Abweichungen. [4]

Wir wollen nun also die zwei Verfahren einander gegenüberstellen:

	Pattern Matching	Feature Matching
PRO	<ul style="list-style-type: none"> ✓ bei grosser Anzahl von Referenzzeichen und richtigem „Training“ bessere Zeichenerkennung als Feature Matching 	<ul style="list-style-type: none"> ✓ geringer Rechenaufwand, dank abstrahierter Sicht auf wenige Merkmale ✓ tolerant, verzerrte Schriftarten werden erkannt ✓ robust
CONTRA	<ul style="list-style-type: none"> • hoher Rechenaufwand („brute-force approach“) • nur anwendbar auf gespeicherte Schriften/Referenzzeichen, intolerant bzgl. Schriftarten, Formatierungen • Training notwendig 	<ul style="list-style-type: none"> • Herausforderung, geeignete Merkmale zu finden • unterbrochene Linien: Gefahr vor <i>void's</i>

3. OCR-Produkte und ihre Technologien

Auf dem Markt gibt es einige professionelle Texterkennungssystemen, die auf unterschiedliche Einsatzzwecke ausgerichtet sind. Eine gute Übersicht bietet folgende Website: [5]

Folgende Produkte haben sich durchgesetzt: Recognita Plus / Omnipage Pro, TextBridge (www.scansoft.com) und FineReader (www.finereader.com). OmniPage und auch die meisten anderen Produkte wenden hauptsächlich Feature Matching an und verzichten z.T. komplett auf Pattern Matching. Dies ermöglicht eine Zeichenerkennung ohne mühsames „Training“. Oft werden die zwei Ansätze aber miteinander kombiniert. So ist es in OmniPage auch möglich, einzelne Zeichen einzutrainieren, was natürlich besonders nützlich ist bei unbekanntem Zeichen.

4. Dokumente und Links:

- [1] Understanding OCR, Mustek
<http://www.mustek.com/Class/ocrinfo.html>
- [2] PCTechGuide
<http://www.pctechguide.com/18scan2.htm#OCR>
- [3] Feature Matching in Geographic Information Systems (GIS)
<http://www.gisdevelopment.net/aars/acrs/1999/ps1/ps1068pf.htm>
- [4] Skript Multimedia SS2002, Prof. Dr. P. Stucki
 Kap. 8. Einführung in die Bild- und Klangverarbeitung; S. 5-13
- [5] Uni Tuebingen, Scan-Systeme zur Texterfassung
<http://www.uni-tuebingen.de/zdv/zrlinfo/scanner.html>