SEALSYSTEMS
THE DIGITAL PAPER FACTORY

FACTSHEET
Texterkennung - OCR

Optische Zeichenerkennung von Texten in eingescannten Dokumenten

Was ist OCR Texterkennung?

Mit der optischen Zeichenerkennung können Texte in eingescannten Dokumenten erkannt werden. Das OCR-Verfahren von SEAL Systems funktioniert für Raster- und Vektordaten und kann in automatisierte Verfahren integriert werden.

Was leistet OCR Texterkennung?

OCR-Techniken können solche Texte maschinenlesbar machen. Sie sind dann automatisch durchsuchbar. Große Dateimengen werden zusätzlich durch Suchmaschinen voruntersucht, sodass das Finden über den Gesamtdateibestand sehr schnell durchgeführt werden kann.

Wer braucht OCR Texterkennung?

Anwender, die Dokumente verarbeiten wollen, die keinen eindeutigen Zeichendoce enthalten.

Ihre Vorteile

- Informationen können in Dateien schneller gefunden werden, wenn die Suche nicht nur über Beschlagwortungen im DMS erfolgt, sondern auch direkt in den Dateien nach relevanten Begriffen gesucht werden kann. Dazu muss der sichtbare Text aber recherchierbar sein.
- Der Datenaustausch in Lieferantenketten bedingt, dass Dokumente nicht immer nur über ein DMS verwaltet werden können. Die Nutzbarkeit von Dateien wird deutlich erhöht, wenn man relevante Schlagworte zum Einordnen der Dateien direkt der Datei einnehmen kann.
- PDF/A löst zunehmend das Rasterformat TIFF als Archivformat ab. Bestandsdateien in TIFF und gescannte Vorlagen lassen sich besonders einfach in das PDF-Format umwandeln. Ohne zusätzliche OCR-Behandlung bringt aber diese Konvertierung keinen Mehrwert. Das Ergebnis-PDF besitzt außer einem Rasterbild keine weiteren Nutzdaten. Erst die Anreicherung mit Textelementen bringt einen zusätzlichen Nutzen.

OCR - Texterkennung

Anwendungsfälle

Es gibt mehrere Möglichkeiten, in denen Texte in Dokumenten nicht als solche von Editoren und Suchmaschinen erkannt werden:

- Eingescannte Dokumente enthalten in der Regel nur Rasterdaten, also Bildpunkte.
- Engineeringtools (CAD) und Layoutanwendungen stellen Buchstaben als Linienzüge oder Flächen dar.
- **Bilder von Texten** sind wie Fotografien.

Damit Texte von Computern lesbar sind, müssen diese durch Zeichen mit einem eindeutigen Zeichencode aus einem Font gebildet werden. Sind diese Zeichencodes nicht korrekt, so sind die Texte zwar sichtbar, aber nicht automatisch zu finden.

Das Verfahren

In einem ersten Schritt werden die Dateien auf eventuell bereits vorhandene Texte untersucht. Alle enthaltenen Grafikelemente, die Texte enthalten oder darstellen können, wie Linienzüge, Bilder von Texten, gerasterte Texte, werden in ein **Rasterformat** umgewandelt. Diese **einheitlichen Daten** dienen als Basis für die OCR Erkennung. Die gefundenen Texte können jetzt als zusätzlicher Layer in der Ausgangsdatei eingelagert werden oder werden als separate Textdatei dem nachfolgenden Prozess zur Verfügung gestellt.

Integration

Die SEAL Systems Texterkennung ist als WorkingUnit für die DPF (Digital Process Factory®) erhältlich. Damit kann diese Funktionalität schnell in alle Abläufe integriert werden:

- Konvertierungsverfahren
- Freigabverfahren
- Einchecken von Dokumenten in DMS
- Altbestandskonvertierung

Eingangsformate

BMP, PCX, DCX, JPEG, **TIFF**, PNG, GIF, DjVu, **PDF** (bis 1.6)

Ausgangsformate

TIFF, PDF, PDF/A. Text in separater Text- oder XML-Datei oder als zusätzlicher Layer in PDF.

Systemumgebung

Windows Server 2003 / Windows XP / 2008 R2 / Windows 7

Produktcode

OCR-25PM, OCR-75PM, OCR-200PM, OCR-500PM, OCR-500T

Dr. Uwe Wächter ist Spezialist für Ihre Fragen rund um das Thema:

Konvertierung und PDF

Dr. Uwe Wächter
Tel +49 6154 637 372
uwe.waechter@sealsystems.de

SEALSYSTEMS
THE DIGITAL PAPER FACTORY

Lohmühlweg 4
91341 Röttenbach (Deutschland)

Tel. +49 9195 926-0
Fax +49 9195 1739
E-Mail: info@sealsystems.de
Web: www.sealsystems.de

OUTPUT MANAGEMENT
CORPORATE SOLUTIONS BY SEAL SYSTEMS

Wir beantworten gerne Ihre Fragen rund um die Erzeugung von Dokumenten aus einer Akte und deren Einsatzmöglichkeiten und Potentiale in Ihrem Hause

© 2016 SEAL Systems AG. PLOSSYS ist eingetragenes Warenzeichen der SEAL Systems AG. Andere in diesem Prospekt erwähnte Computer- und Softwarebezeichnungensind Handelsmarken und/oder Warenzeichen der entsprechenden Hersteller. Änderungen vorbehalten. Stand: 6. August 2016. V512-120327-0-de