

3-Heights™ OCR Enterprise Add-On

Das 3-Heights™ OCR Enterprise Add-On ergänzt mehrere Produkte der PDF Tools AG mit einer leistungsfähigen Texterkennung. Somit können bei der Konvertierung von Bildern, wie TIFF oder JPEG nach PDF oder PDF/A, oder bei der Konvertierung von PDF nach PDF/A im gleichen Schritt OCR Informationen hinzugefügt werden.

Bei der OCR Engine hat der Kunde die freie Wahl der Engine. Zur Verfügung steht zurzeit ABBYY FineReader 10 und Tesseract mit verschiedenen Lizenzierungsarten. Somit kann je nach Anforderung an Erkennungsrate, Durchsatz und Kosten eine passende Lösung gefunden werden.




Eigenschaften und Nutzen

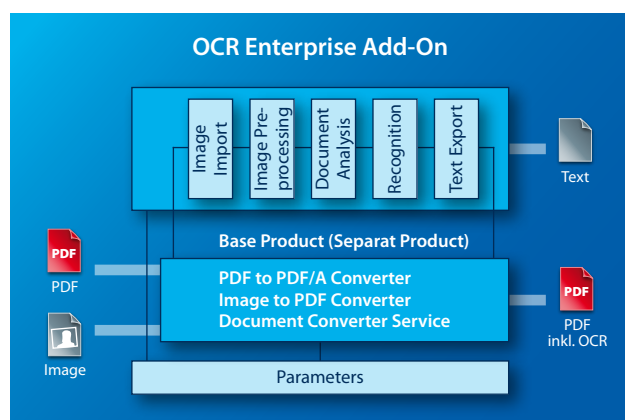
Das 3-Heights™ OCR Enterprise Add-On ist ein OCR Modul, das in Verbindung mit mehreren Produkten von PDF Tools AG zum Einsatz kommt.

Basierend auf der gewählten OCR Engine erkennt es Textinhalte und bettet diese als

Unicode Text in das PDF und PDF/A Dokument ein. Damit sind die PDF Dokumente im Volltext durchsuchbar. Zahlreiche Optionen in den Bereichen Bildmanipulation, Vorverarbeitung und Texterkennung ermöglichen einen optimal auf die Bedürfnisse abgestimmten Erkennungsprozess.

Produktvarianten

API	Shell	Service
		



Einsatzgebiete

Eingangspost

Beim Scannen der Eingangspost wird Text erkannt und kann in den Metadaten der Eingangsdokumente sowie in nachgelagerten Unternehmensprozessen, wie beispielsweise ERP und Workflow Systemen, verwendet werden. Dokumente im Eingangsstadium werden inklusive Texterkennung direkt archiviert. Zur vereinfachten Weiterverarbeitung wird eine Texterkennung in gescannten E-Mail Anhängen durchgeführt.

Archivierung

Texterkennung bei der Konvertierung von Archiven aus dem TIFF oder PDF Format ins standardisierte PDF/A Format. Umwandlung proprietärer Formate nach PDF/A und Einbettung der Texte. Erkennung von Informationen auf Indexierungsblättern und Übertragung in die Metadaten des Dokumentes.

Weitere Einsatzgebiete

- Scannen von E-Mail Anhängen
- Vorbereitung zur Archivierung
- Archivmigration

Formate

Eingangformate

Bestimmt durch das Basisprodukt:

- 3-Heights™ PDF to PDF/A Converter
 - PDF
- 3-Heights™ Image to PDF Converter
 - TIFF (Tagged Image File Format)
 - JPEG (Joint Photographic Expert Group)
 - PNG (Portable Network Graphics)
 - GIF (Graphics Interchange Format)
 - BMP (Window Bitmap)
 - EPS (Encapsulated Post Script)
 - JB2 (JBIG2, Joint Bi-level Image Experts Group)
 - JP2 (JPEG2000)
 - JPX (Extended JPEG2000)
 - PBM (Portable Bitmap File Format)
 - JIF (GIF Flate)
- 3-Heights™ Document Converter Service
 - Microsoft Office 2003 und 2007 Dokumente
 - Dokumente von älteren Microsoft Office Versionen
 - Einfacher Text
 - WordPerfect
 - HTML
 - Outlook (MSG)
 - PDF
 - Internet Mail Message Format
 - Bilddateien (TIFF, JPEG, PNG, JBIG2, JPX, GIF, BMP usw.)
 - ZIP und TAR Archive
 - Add-ins für kundenspezifische Formate

Zielformate

- PDF, PDF/A

Varianten und benötigte Basisprodukte

Benötigte Basisprodukte

Das 3-Heights™ OCR Enterprise Add-On kann mit folgenden Basisprodukten verwendet werden:

- 3-Heights™ PDF to PDF/A Converter
- 3-Heights™ Image to PDF Converter
- 3-Heights™ Document Converter Service

Plattformen

Betriebssysteme

- ABBYY: Windows 2000, XP, Vista, 7
Windows Server 2003, 2008,
2008 R2 – 32 und 64 Bit
- Tesseract: Windows, Linux, andere auf Anfrage

Schnittstellen und Sprachen

Die OCR Engine wird als separates Produkt installiert. Sie verfügt über keine sichtbaren Schnittstellen. Stattdessen wird sie direkt aus dem Basisprodukt angesprochen. Wenn das Basisprodukt eine API ist, so stehen dort entsprechende Funktionen oder Methoden zur Verfügung. Wenn das Basisprodukt ein Shell Tool oder ein Service ist, so gibt es entsprechenden Parameter.

Schnittstellen

Entsprechend dem Basisprodukt:

- API: C, Java, .NET, COM
- Shell Tool: Befehlszeile für Stapelverarbeitung
- Windows Service: Windows Dienst mit Überwachten Verzeichnissen

Leistungsmerkmale

- Verfügbare OCR Engines:
 - ABBYY FineReader OCR Engine 10
 - Tesseract OCR Engine 2
- Je nach OCR Engine werden bis zu 200 Sprachen unterstützt
- Erweiterte Unterstützung mit Wörterbüchern
- Lizenzschlüssel Migration (nur ABBYY 10)

Funktionen

Allgemeine Funktionen

- OCR Textinformation für PDF Dokumente hinzufügen
- Setzen der OCR Sprache(n), um die Erkennungsrate zu erhöhen
- Direkte Ansteuerung der OCR Engine oder Ansteuerung via Service zur Synchronisierung
- Erkennung von mehrsprachigen Dokumenten

Tesseract spezifische Funktionen

- Unterstützung von 8 Sprachen mittels Wörterbüchern
- Optimierte für gescannte Bilder mit einer Auflösung von 300 dpi
- Automatische Baseline Detektierung

ABBYY spezifische Funktionen

- Erkennung von fast 200 Sprachen bei Maschinen generierten Inhalten
- Erweiterte Unterstützung von fast 50 Sprachen mittels Wörterbüchern und morphologischen Werkzeugen
- Chinesisch, Japanisch, Koreanisch
- Alte europäischen Sprachen
- 2D Barcode
- Erkennung von Schreibmaschinenschriften
- Erkennung und Decodierung von 1D Barcodes
- Erkennung von Inhaltstypen (Bilder versus Texte)
- Schneller, präziser oder ausbalancierter Modus wählen
- De-Skewing: Automatische Ausrichtung von Bildern
- Bildreinigung: unerwünschte Artefakte werden erkannt und eliminiert
- Filterung von nicht relevanten Hintergründen
- Erkennung und Korrektur der Seitenorientierung
- Erstellung und Verwendung von Profilen, um obige Funktionen zusammenzufassen