R&D : OCR

-Dhanushkumar.R & Harisudhan.S

Optical character recognition or **optical character reader** (**OCR**) is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene photo.

Types

- Optical character recognition (OCR) – targets typewritten text, one glyph or character at a time.
- Optical word recognition – targets typewritten text, one word at a time (for languages that use a space as a word divider). Usually just called "OCR".
- Intelligent character recognition (ICR) – also targets handwritten printscript or cursive text one glyph or character at a time, usually involving machine learning.
- Intelligent word recognition (IWR) – also targets handwritten printscript or cursive text, one word at a time. This is especially useful for languages where glyphs are not separated in cursive script.

Techniques

Pre-processing

OCR software often pre-processes images to improve the chances of successful recognition. Techniques include:

- De-skewing – if the document was not aligned properly when scanned, it may need to be tilted a few degrees clockwise or counterclockwise in order to make lines of text perfectly horizontal or vertical.
- Despeckling – removal of positive and negative spots, smoothing edges
- Binarization – conversion of an image from color or greyscale to black-and-white (called a binary image because there are two colors). The task is performed as a simple way of separating the text (or any other desired image component) from the background.
- Line removal – Cleaning up non-glyph boxes and lines
- Layout analysis or zoning – Identification of columns, paragraphs, captions, etc. as distinct blocks. Especially important in multi-column layouts and tables.
- Line and word detection – Establishment of a baseline for word and character shapes, separating words as necessary.
- Script recognition – In multilingual documents, the script may change at the level of the words and hence, identification of the script is necessary, before the right OCR can be invoked to handle the specific script.

- Character isolation or segmentation – For per-character OCR, multiple characters that are connected due to image artifacts must be separated; single characters that are broken into multiple pieces due to artifacts must be connected.
- Normalization of aspect ratio and scale

Text recognition

There are two basic types of core OCR algorithm, which may produce a ranked list of candidate characters

- *Matrix matching* involves comparing an image to a stored glyph on a pixel-by-pixel basis; it is also known as *pattern matching*, *pattern recognition*, or *image correlation*. This relies on the input glyph being correctly isolated from the rest of the image, and the stored glyph being in a similar font and at the same scale. This technique works best with typewritten text and does not work well when new fonts are encountered. This is the technique early physical photocell-based OCR implemented, rather directly.
- *Feature extraction* decomposes glyphs into "features" like lines, closed loops, line direction, and line intersections. The extraction features reduces the dimensionality of the representation and makes the recognition process computationally efficient. These features are compared with an abstract vector-like representation of a character, which might reduce to one or more glyph prototypes. General techniques of feature detection in computer vision are applicable to this type of OCR, which is commonly seen in "intelligent" handwriting recognition and most modern OCR software. Nearest neighbour classifiers such as the k-nearest neighbors algorithm are used to compare image features with stored glyph features and choose the nearest match

Software such as Cuneiform and Tesseract use a two-pass approach to character recognition. The second pass is known as adaptive recognition and uses the letter shapes recognized with high confidence on the first pass to better recognize the remaining letters on the second pass. This is advantageous for unusual fonts or low-quality scans where the font is distorted (e.g. blurred or faded)

Post-processing

OCR accuracy can be increased if the output is constrained by a lexicon – a list of words that are allowed to occur in a document. This might be, for example, all the words in the English language, or a more technical lexicon for a specific field. This technique can be problematic if the document contains words not in the lexicon, like proper nouns. Tesseract uses its dictionary to influence the character segmentation step, for improved accuracy
The output stream may be a plain text stream or file of characters, but more sophisticated OCR systems can preserve the original layout of the page and produce, for example, an annotated PDF that includes both the original image of the page and a searchable textual representation.

Here are some of the various types of OCR:

- Printed Text OCR: This is the most common type of OCR and is used to extract text from printed documents, such as books, magazines, newspapers, and printed forms.

- Handwritten Text OCR: Handwritten text recognition OCR is used to convert handwritten text into digital text. This is useful in applications like digitizing historical documents, recognizing handwritten forms, and converting handwritten notes into editable text.

- Machine-printed OCR: This type of OCR is optimized for printed text that is generated by machines, such as computer-printed documents, receipts, and invoices.

- Cursive OCR: Cursive OCR focuses on recognizing cursive handwriting, which is a more challenging task compared to printed or block letters.

- Script-Specific OCR: Some OCR systems are designed to recognize specific scripts or languages, such as Chinese characters, Arabic script, or Cyrillic script. These OCR systems are tailored to the unique characteristics of the script.

- Bank Check OCR: This specialized OCR is used by banks to recognize the handwritten or printed information on checks, including the account number, amount, and signature.

- License Plate OCR: License plate recognition OCR is employed for reading vehicle license plates. It's often used in toll collection, parking management, and law enforcement.

- Form OCR: Form recognition OCR is used to extract data from structured forms, such as surveys, questionnaires, and application forms. It can identify and capture specific fields like names, addresses, and checkboxes.

- Invoice OCR: Invoice OCR automates the extraction of information from invoices, including vendor details, invoice numbers, dates, and line item data. It's widely used in accounts payable automation.

- Receipt OCR: Receipt OCR is used to extract data from retail and restaurant receipts, helping with expense tracking, accounting, and tax purposes.

- ID Card OCR: ID card recognition OCR is employed in various applications, such as border control and identity verification, to extract information from government-issued ID cards, passports, and driver's licenses.

- Bank Statement OCR: This OCR type is used in financial institutions to digitize and extract data from bank statements, facilitating account reconciliation and financial analysis.

- ❖ E-book OCR: E-book OCR is used to convert scanned pages of printed books into digital text for creating e-books or making printed content searchable.
- ❖ Document OCR: Document OCR is a general-purpose OCR used to extract text and data from a wide range of documents, including business documents, contracts, and manuals.
- ❖ Mobile OCR: Mobile OCR applications are designed for smartphones and tablets, enabling users to capture text using the device's camera and convert it into editable text or perform actions like translation.

This **comparison of optical character recognition software includes:**

- OCR engines, that do the actual character identification
- Layout analysis software, that divide scanned documents into zones suitable for OCR
- Graphical interfaces to one or more OCR engines

Name	Founded year	Latest stable version	Release year	License	Online	Windows	Mac OS X	Linux	BSD	Android	iOS	Programming language	SDK?	Languages	Fonts	Output Formats	Notes
Google Drive OCR or Google Cloud Vision			2015	Proprietary	Yes	Browser	Browser	Browser	Unknown	?	?	Unknown	Yes	200+	All fonts	text	Google blog post[1][2]
Tesseract	1985	5.2.0	2022	Apache	No	Yes	Yes	Yes	Yes	?	?	C++, C	Yes	100+[3]	Any printed font	Text, ALTO, hOCR,[4] PDF, others with different user interfaces[5] or the API	Created by Hewlett-Packard; under further development by Google[6]
ABBYY FineReader	1989	16	2022	Proprietary	Yes	Yes	Yes	No	Yes	Yes	Yes	C/C++	Yes	192[7]	All fonts	DOC, DOCX, XLS, XLSX, PPTX, RTF, PDF, HTML, CSV, TXT, ODT, DjVu, EPUB, FB2[8]	ABBYY also supplies SDKs for embedded and mobile devices. Professional, Corporate and Site License Editions for Windows, Express Edition for Mac.[9]
E-akshara yan	2010					Yes	No	Yes	No	?	?			14		RTF, TXT, BRL	
Asprise OCR SDK	1998	15	2015	Proprietary	Yes	Yes	Yes	Yes	Yes	?	?	Java, C#,VB.NET, C/C++/Delphi	Yes	20+[10]	?	Plain text, searchable PDF, XML[11]	Java, C#, VB.NET, C/C++/Delphi SDKs for OCR and Barcode recognition on Windows, Linux, Mac OS X and Unix.[12]
AnyDoc Software	1989	?	?	Proprietary	No	Yes	No	No	No	?	?	VBScript	?	?	?		Works with structured, semi-structured, and unstructured documents.
CuneiForm	1996	1.1	2011-04-19	BSD variant	No	Yes	Yes	Yes	Yes	?	?	C/C++	Yes	28	Any printed font	HTML, hOCR, native, RTF, TeX, TXT[13]	Enterprise-class system, can save text formatting and recognizes complicated tables of any structure
Dynamsoft OCR SDK	2003	8.2	2012	Proprietary	Yes	Yes	No	No	No	?	?	C/C++	Yes	40+[14]	?	PDF, TXT	

OmniPage	1970s	19.2	2015	Proprietary	Yes	Yes	Yes	Yes	No	?	?	C/C++, C#[15]	Yes	125[16]	Machine and handprinted fonts	DOC/DOCX XLS/XLSX PPTX RTF PDF PDF/A Searchable PDF HTML Text XML ePUB MP3	Product of Nuance Communications
Microsoft Office OneNote 2007	2011	?	2007	Proprietary	No	Yes	No	No	No	?	?	?	?	?	?		
GOCR	2000	0.52[17]	2018-10-15	GPL	Yes[18]	Yes	Yes	Yes	Yes	?	?	C	?	20+	?		
Ocrad	?	0.26[19]	2017-03-31	GPL	Yes	No	Yes	Yes	Yes	?	?	C++	Yes	Latin alphabet	?		Command line
SmartScore	1991	10.5.8	2015-07	Proprietary	No	Yes	Yes	No	No	?	?	?	?	?	?		For musical scores
Microsoft Office Document Imaging	?	Office 2007	2007	Proprietary	No	Yes	No	No	No	?	?	?	?	?	?		Uses OmniPage[citation needed]
Puma.NET	?	?	2009-10-29	BSD	No	Yes	No	No	No	?	?	C#	Yes	28	Any printed font		.NET OCR SDK based on Cognitive Technologies' CuneiForm recognition engine. Wraps Puma COM server and provides simplified API for .NET applications
ReadSoft	?	?	?	Proprietary	No	Yes	No	No	No	?	?	?	?	?	?		Scan, capture and classify business documents such as invoices, forms and purchase orders integrated with business processes.
Scantron	?	?	?	Proprietary	No	Yes	No	No	No	?	?	?	?	?	?		For working with localized interfaces, corresponding language support is required.
OCRFeeder	2009-03	0.8.3	2014-12-22	GPL	No	No	No	Yes	No	?	?	Python	?	?	?		Features a full user interface and has a command-line tool for automatic operations. Has its own segmentation algorithm but uses system-wide OCR engines like Tesseract or Ocrad
OCRopus	2007	1.3.3	2017-12-16	Apache	No	No	Yes	Yes	Yes	?	?	Python	?	All languages using Latin script (other languages can be trained)	Normal Latin script and Fraktur (other scripts can be trained)	TXT, hOCR,[20] PDF[21]	Pluggable framework under active development, used for Google Books

Keras OCR:

Description:

Keras OCR is an OCR library built on top of the Keras deep learning framework. Keras is a popular high-level neural networks API written in Python. Keras OCR is designed to perform optical character recognition tasks using deep learning models.

Features:

Provides pre-trained deep learning models for OCR tasks.
Allows fine-tuning or training custom models for specific OCR requirements.
Integrates with image preprocessing and post-processing techniques for better accuracy.
Can handle tasks such as text detection and recognition.

Use Cases: Keras OCR is suitable for a wide range of OCR applications, including printed text recognition, handwritten text recognition, and document analysis.

Paddle OCR:

Description:
Paddle OCR is an OCR library developed by Baidu's PaddlePaddle team. It is built on the PaddlePaddle deep learning platform, which is an open-source platform for deep learning. Paddle OCR is designed to provide efficient and accurate OCR capabilities.
Features:
Offers pre-trained models for various OCR tasks, including text detection and text recognition.
Supports both traditional and simplified Chinese character recognition.
Includes tools for data preprocessing and evaluation.
Can be used for tasks like document scanning, text extraction, and more.
Use Cases:
Paddle OCR is particularly useful for applications that involve Chinese text recognition and OCR tasks in general, including document digitization and image-to-text conversion.
These libraries provide developers with pre-built models and tools for OCR tasks, making it easier to implement OCR in their applications. They are part of the broader ecosystem of deep learning frameworks and tools designed to solve optical character recognition challenges.

Tesseract OCR:

Description:
Tesseract is an open-source OCR engine developed by Google. It is one of the most widely used OCR libraries and supports multiple languages.
Features:
Supports text recognition for various languages.
Can be integrated with various programming languages through wrappers.
Allows training custom models for specific OCR tasks.

Pytesseract (Python Wrapper for Tesseract):

Description:
Pytesseract is a Python wrapper for the Tesseract OCR engine, making it easy to use Tesseract's OCR capabilities in Python applications.
Microsoft Azure Cognitive Services OCR:

Description:
Microsoft Azure offers OCR as a part of its Cognitive Services suite. It provides cloud-based OCR capabilities with high accuracy and language support.
Features:
Supports printed text recognition.
Integrates with other Azure services for broader document processing tasks.
Offers a REST API for easy integration into applications.

ABBYY FineReader:

Description:

ABBYY FineReader is a commercial OCR software with advanced features for document scanning and recognition.
Features:
Supports OCR for printed and handwritten text.
Offers features like layout retention, table recognition, and PDF conversion.
Suitable for document management and digitization tasks.

GOCR:

Description:
GOCR is an open-source OCR engine that is designed for simplicity and ease of use.
Features:
Provides basic OCR capabilities for printed text.
Suitable for simple OCR tasks where ease of installation and use is a priority.

OCRopus:

Description:
OCRopus is an OCR system developed by Google. It is designed for large-scale OCR processing and document analysis.
Features:
Provides a collection of OCR tools and utilities.
Supports layout analysis, text recognition, and more.
Can be used for batch processing of documents.

Amazon Textract:

Description: Amazon Textract is a cloud-based OCR service provided by AWS. It is designed for extracting text and data from documents.
Features:
Supports structured data extraction from forms and tables.
Offers machine learning-based text extraction.
Integrates with other AWS services for document processing workflows.

Evaluation

016 analysis of the accuracy and reliability of the OCR packages Google Docs OCR, Tesseract, ABBYY FineReader, and Transym, employing a dataset including 1227 images from 15 different categories concluded Google Docs OCR and ABBYY to be performing better than others.