What is OCR (Optical Character Recognition)?

Optical Character Recognition (OCR) is the process that converts an image of text into a machine-readable text format. For example, if you scan a form or a receipt, your computer saves the scan as an image file. You cannot use a text editor to edit, search, or count the words in the image file. However, you can use OCR to convert the image into a text document with its contents stored as text data.

Why is OCR important?

Most business workflows involve receiving information from print media. Paper forms, invoices, scanned legal documents, and printed contracts are all part of business processes. These large volumes of paperwork take a lot of time and space to store and manage. Though paperless document management is the way to go, scanning the document into an image creates challenges. The process requires manual intervention and can be tedious and slow.

Moreover, digitizing this document content creates image files with the text hidden within it. Text in images cannot be processed by word processing software in the same way as text documents. OCR technology solves the problem by converting text images into text data that can be analyzed by other business software. You can then use the data to conduct analytics, streamline operations, automate processes, and improve productivity.

How does OCR work?

The OCR engine or OCR software works by using the following steps:

Image acquisition

A scanner reads documents and converts them to binary data. The OCR software analyzes the scanned image and classifies the light areas as background and the dark areas as text.

Preprocessing

The OCR software first cleans the image and removes errors to prepare it for reading. These are some of its cleaning techniques:

- Deskewing or tilting the scanned document slightly to fix alignment issues during the scan.
- Despeckling or removing any digital image spots or smoothing the edges of text images.
- Cleaning up boxes and lines in the image.
- Script recognition for multi-language OCR technology