# What is OCR Technology: A Step By Step Guide for Data Extraction

Think of something like how you can move from paper to digital documents to save your time?

How can you move data of tons of papers to a hard drive or even on the cloud?

[Optical character recognition technology](https://pdf.abbyy.com/what-is-ocr/) makes it easier for us to translate scanned physical files into editable digital documents.



Imagine you have a paper document, for example, magazines, brochures, PDF contracts sent by your colleague through email.

Obviously, a scanner alone is not enough for this information to make it editable in any text editor.

All you can do with a scanner is create a snapshot of an image or document, which is just a collection of black-and-white or color dots.

To extract and divert data from scanned documents, camera images, or image-only PDF files, you need OCR software to extract the letters from the images, and then convert them to words.

Now you can edit the content or information of the document.

OCR uses technology to perceive and change printed characters from handwritten scanned documents into electronic text that can be easily accepted and recognized by a computer.

The primary OCR operation involves checking text and translating characters into code that can be used to process data.

OCR is sometimes called text recognition and it combines software and hardware used to convert physical files into machine-readable text.

Hardware such as dedicated wiring boards or optical scanners is used to read the text. It also helps to copy the text of the physical document.

On the other hand, the software is responsible for advanced working.

The software can use artificial technology or machine learning to implement advanced recognition techniques, for example, recognizing the handwriting styles or different languages.

OCR is mostly used in converting the historical files to PDF documents.

The resulting electronic form of a document can be edited and formatted by any person using ordinary text editors.

## How OCR extract data?

* The first step in OCR recognition involves a scanner to process the physical document.
* After reading and copying all the pages, OCR transforms the document into a black-and-white or two-colored version.
* The scanned document is analyzed for the presence of dark and light areas.
* Dark areas of the file are recognized as characters and light areas are identified as background.
* The dark areas are then processed to find letters, numbers, and symbols.

Existing recognition software and programs can have different algorithms to recognize characters.

However, they all include targeting one character, word, or block of text.

We are going to discuss two basic algorithms for this purpose.

### Algorithm1

Recognized material is processed using examples of text formats and various fonts.

### Algorithm2

Recognition is done based on the use of feature detection rules regarding the characteristics of a specific number or letter.

With this detection feature, the program evaluates the document content according to rules about how a number or letter is generated.

For example, the letter “A” can be identified as two diagonal lines intersecting with a horizontal line in the middle.

When the character is recognized, it is transformed into an [ASCII code](https://www.w3schools.com/charsets/ref_html_ascii.asp#:~:text=ASCII%20is%20a%207%2Dbit,Z%2C%20and%20some%20special%20characters.) that can be read by a machine or computer.



## Application of OCR

* Scanning the printed documents in a version that can be edited with a common text editor.
* Indexing the printed material for the search engines.
* Data entry and automated processing.
* Translate documents into a text that can be read aloud for visually disabling users.
* Data extraction and transferring it to accounting software like receipts and invoices.
* Archiving of historical information for example, in newspapers and magazines.
* Placing of important documents in an electronic database.
* Sorting letters for mail delivery.
* Recognition of characters on a license plate with the help of speed camera and software.
* Providing the facility of the search for scanned books.

Another very useful and most used application of optical character recognition is an image to text conversion.

There are numerous sites on the internet that provide the image to text converter which helps to extract the information and data from the image.

One of them is [prepostseo](https://www.prepostseo.com/image-to-text) which provides you the facility to extract text from different formats such as PNG, JPG, BMP, GIF, JPEG, and TIFF.



It is very easy to use as you just have to upload the picture and click on the submit button. It will do the rest of the work.

It also gave us the facility to upload the URL of online image and extract information out of it.

After uploading the image and clicking on the submit button, you will get an instant result.



You can see HCF & LCM was written on the picture and image to text converter extracted this text correctly.

It was just simple; you can extract a large amount of data as well.

In the end, you can download the text in .txt file or copy it to your clipboard as well.

Suppose you want to digitize an image text or a printed agreement. You can type it several times to correct the wrong print.

Alternatively, you can use a scanner or digital camera and Optical Character Recognition software such as an image-to-text converter, so you can convert all your desired content to digital format in minutes.

## Last words

Before the advent of optical character recognition technology, the only way to digitize paper media was to reprint the text.

This process was time-consuming and often resulted in printing errors. Using OCR saves time, helps eradicate errors, and reduces effort.

Additionally, technology allows you to perform actions that are not available to physical copies, for example, it can use compression in zip files, highlight keywords, post documents on the website, attach them to email.