parsel

WHITE PAPER

Optical Character Recognition: Digitising the world

Optical Character Recognition technology (or OCR) is bridging the gap between the physical and digital world by enabling the smooth conversion of printed text to machine-readable format. It has applications at the corporate and retail/individual levels and its use cases are expanding rapidly.

By Rohit Kumar
Financial Analyst
Tellimer

Consider being on an international trip to a country where language is a barrier, Google Translate can transform your camera into a live translator where you can see pretty much all the information written on signboards, products, invoices, etc., in your own language – OCR is the technology enabling this.

And, if you are an iPhone user, the latest iOS 15 has built-in OCR technology, meaning you can take a picture of any document and directly copy text and characters in the image – do try that, if you haven't already.

Here, we look at how this revolutionary technology works, the service providers that are available in the market, the limitations of the technology and its real-life applications across different sectors.

Optical Character Recognition technology (or OCR) is bridging the gap between the physical and digital world by enabling the smooth conversion of printed text to machine-readable format.

Understanding OCR technology

What is OCR? Optical Character Recognition is a technology that recognises text within a digital image. This enables the rapid transformation of printed/handwritten documents into machine-readable text and allows analytical tools to process the data.

History of OCR: The roots of OCR date back a century, when Austrian engineer Gustav Tauschek invented a mechanical device called the 'Reading Machine' in the 1920s capable of reading text in printed documents. Emanuel Goldberg also invented a 'Statistical Machine' around the same time, capable of reading and converting characters into telegraphic code. The technology evolved over time, with the first text-to-speech device developed in the 1960s to help the blind. Today, it has reached a point where most types of text and data can be converted into a machine-readable format with high accuracy.

How does the technology work?

OCR technology broadly works in a three-step model, as follows.

1. Pre-processing: This step makes changes to the image to make it as easy to read as possible. Some of the techniques applied here include rotating, aligning, 'de-speckling' (removing spots, etc.) and converting from colour to a binary image, where the only two colours are black and white.

2. Character recognition: Two approaches are generally used for text recognition after the image has gone through the pre-processing stage:

- *Feature detection (see Fig 1.):* This is a new approach in OCR technology. Here, the text character is decomposed into features such as lines and strokes; then, the algorithm identifies the character by analysing the features that comprise it. As different fonts/handwriting have differing features for the same character, this technique uses artificial intelligence (AI) processes, where different neural networks are used to recognise patterns.

- *Matrix matching (see Fig 2.):* This is an older approach, and converts the image of the character into a binary matrix, where white pixels represent 0 and black pixels represent 1. The system then identifies patterns within

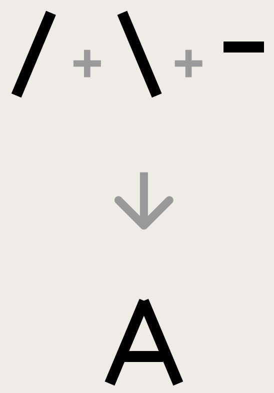

Fig 1.
Feature detection

the binary matrix and matches them with its database, returning the character that is statistically most common.

3. Post-processing: After the OCR systems recognises and coverts the text into machine-readable form, the output is then analysed for any errors to improve the accuracy. This step uses natural language processing (NLP) techniques to minimise errors.

There are various methods, including lexicon error correction (comparing a word against a dictionary of similar words) and grammar/semantic error correction (context-based correction performed with statistical language models, this is the same technology used for predictive text in our electronic devices).

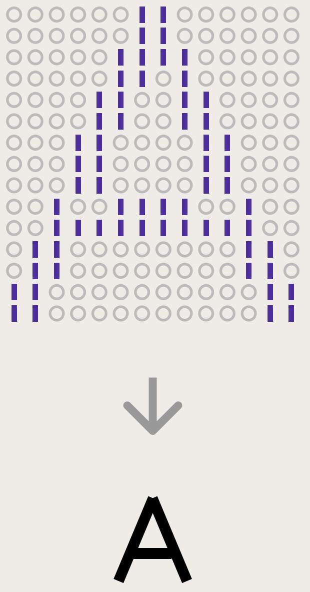

Fig 2.
Matrix matching

The limitations of OCR technology

OCR cannot work independently: OCR is an exciting technology that recognises and converts printed/handwritten characters into machine-readable form, but its output is, for the most part, unstructured raw data. Therefore, combining OCR with other modern technologies like AI, machine learning and NLP is necessary to achieve accurate and useable results.

It is not a one-size-fits-all solution: As OCR is usually combined with machine-learning technologies and involves rigorous training of the OCR model, the system's design is the key to the accuracy and reliability of the output. For example, some OCR-based systems (Parsel, for example) are built specifically to extract tabular information from documents such as annual reports and invoices, etc., while others are designed to process health records or identity documents.

OCR is still not 100% accurate: The accuracy of OCR has increased significantly but there are still instances where the output is not completely accurate.

The accuracy of OCR has increased significantly but there are still instances where the output is not completely accurate.

Some of the problems OCR model encounter include:

- *Coloured backgrounds* – it can be difficult to differentiate between text and coloured backgrounds
- *Blurry text* – it is challenging for an OCR system to analyse patterns in blurry text
- *Skewed documents* – it is difficult for OCR-based systems to analyse characters if they are not aligned
- *Unidentified font types* – any font types not in the system's database would be challenging for it to recognise
- *Similar characters* – some characters are very similar in shape (for example, '0' and 'O'), which presents problems
- *Handwritten text* – everyone has different handwriting and OCR systems may sometimes be unable to correctly capture the patterns within a particular style.

OCR market overview

As use cases of OCR multiply across the globe, the number of service providers offering OCR solutions is rising. By and large, there are two types of OCR software or engines: open-source engines and proprietary software/solutions.

Open-source OCR engines: These systems are available publicly and allow other technology companies to use and modify the code and enable them to build their own targeted solutions by training AI models. Some examples of open-source OCR engines include:

- Tesseract
- OCRopus
- Kraken
- GOCR
- A9T9

Proprietary OCR solutions. These OCR solutions are available on a use-only basis and cannot be modified. They also include software solutions that are built on top of an open-source engine, but with a targeted use case built from proprietary AI training models. There are a variety of OCR software/platforms on the market, including some from big cloud providers such as Google, Amazon and Microsoft. Examples include:

- Parsel
- AWS Textract
- Google Cloud Vision API
- Microsoft Azure OCR
- ABBY FineReader

There are a variety of OCR software / platforms on the market, including some from big cloud providers such as Google, Amazon and Microsoft.

How OCR is transforming different industries

OCR technology has applications in many industries and has significantly reduced time and human efforts in variety of tasks, as well as improving the experiences of consumers. Below, we discuss some of the use cases of OCR across different sectors.

Banking. Banks can digitise various document used in risk management such as financial statements, mortgage papers, contracts, etc, using OCR. In addition, banks can automate account opening and know-your-customer (KYC) procedures using OCR, which can be used to read a customer's identity documents. One other use case for the banking industry is digital cheque clearing – OCR can make it possible to process handwritten checks.

Healthcare. OCR can be of a great use in automating health records. Millions of medical claims are processed every year, which involves a lot of manual processing, as medical records and bills are often printed or handwritten. OCR can help immensely in speeding up the process of filing and assessing these medical claims.

Logistics. There is a great deal of manual data entry and document processing in the logistics sector with regards to invoicing, cargo specification and other relevant paperwork. OCR can easily automate the bulk of this manual processing with great accuracy.

Accounting. Corporate accounts departments deal with enormous amounts of paperwork, including invoices and

OCR tech has applications in many industries and has significantly reduced time and human efforts in variety of tasks, as well as improving the experiences of consumers.

receipts that need to be recorded. In many companies, this process is still manual and requires significant human effort. Invoice-scanning OCR technology can automate the processing of invoices with high accuracy and speed.

Invoice-scanning OCR technology can automate the processing of invoices with high accuracy and speed.

Financial Research. My own field of financial research involves analysing vast amounts of data, mostly in numerical form. This data can be contained in large PDF documents from corporates, governments and multilateral agencies. OCR technology can significantly reduce the time it takes for financial analysts to extract data and update their financial and economic models.

Public institutions. Government departments, such as revenue, excise, immigration, education, etc., often deal with significant amount of paperwork, too. Huge physical store rooms contain old files and data, but government departments across the globe are gradually digitising their databases – OCR helps expedite this process.