

# Optical character recognition

From Wikipedia, the free encyclopedia

**Optical character recognition** (**optical character reader**, **OCR**) is the mechanical or electronic 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 (for example the text on signs and billboards in a landscape photo) or from subtitle text superimposed on an image (for example from a television broadcast).<sup>[1]</sup> It is widely used as a form of information entry from printed paper data records, whether passport documents, invoices, bank statements, computerised receipts, business cards, mail, printouts of static-data, or any suitable documentation. It is a common method of digitising printed texts so that they can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as cognitive computing, machine translation, (extracted) text-to-speech, key data and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision.

Early versions needed to be trained with images of each character, and worked on one font at a time. Advanced systems capable of producing a high degree of recognition accuracy for most fonts are now common, and with support for a variety of digital image file format inputs.<sup>[2]</sup> Some systems are capable of reproducing formatted output that closely approximates the original page including images, columns, and other non-textual components.

## Contents

- 1 History
  - 1.1 Blind and visually impaired users
- 2 Applications
- 3 Types
- 4 Techniques
  - 4.1 Pre-processing
  - 4.2 Character recognition
  - 4.3 Post-processing
  - 4.4 Application-specific optimisations
  - 4.5 OCR Tools
- 5 Workarounds
  - 5.1 Forcing better input
  - 5.2 Crowdsourcing
- 6 Accuracy
- 7 Unicode
- 8 See also
- 9 References
- 10 External links

## History

Early optical character recognition may be traced to technologies involving telegraphy and creating reading devices for the blind.<sup>[3]</sup> In 1914, Emanuel Goldberg developed a machine that read characters and converted them into standard telegraph code. Concurrently, Edmund Fournier d'Albe developed the Optophone, a handheld scanner that when moved across a printed page, produced tones that corresponded to specific letters or