ind.senz OCR engines for Indian languages

How Does OCR Work?

OCR (optical character recognition) programs are "computer typists": You give a scanned or photographed image of a text to the program, and the OCR converts this text image into digital, editable text.

At the core of most OCR engines, similar mathematical and linguistic techniques are used. However, they need to be adapted to each specific script and language to guarantee a better performance.

Our OCR programs process a page of text in three steps:

(1) A text image is opened and cleaned (preprocessing).

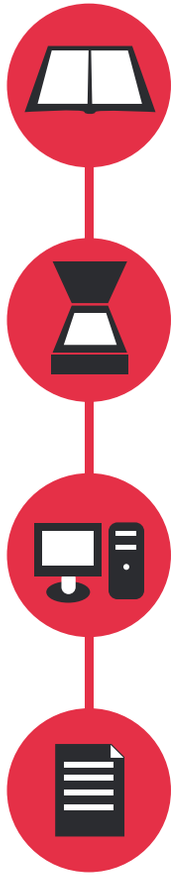
(2) The engine analyzes the image, detects the text in the image, and tries to "read" it.

(3) The program uses built-in knowledge about the language to improve the results (postprocessing).

When and How to Use OCR?

OCR is a drop-in replacement of a human typist. It is useful for any kind of conversion from printed into digital documents (books, newspapers, printed files, etc.). Use OCR as a readymade plugin in your digitization workflow – and combine it with a human check of the results, if needed!

The OCR engine for Hindi is also available as an API, which allows a more flexible integration into your workflow.



http://www.indsenz.com

How to produce high-quality output?

Although OCR techniques have made rapid advances in the last decades, OCR engines still cannot compete with the human capability of reading and understanding text. To obtain high-quality digitization results, you may consider a few tweaks:

Image quality is one of the central factors for a successful conversion. Low image resolution, bad paper quality, or broken and bloated letters decrease the output quality of an OCR engine. Therefore, we recommend using greyscale images of well printed pages scanned with at least 250 dpi.

Language and content: If the OCR "knows" all words found in a text, it is able to correct errors made during recognition. Therefore, a good coverage of words is essential for obtaining high-quality results. The program achieves best accuracy rates for texts written in standard Hindi, and dealing with usual topics. Mixed language documents are currently not covered by the ind.senz engines.

Contact:

oliver.hellwig@indsenz.com

skype: o.hellwig7

http://www.indsenz.com

ind.senz

Dr. Oliver Hellwig

Heiterwanger Weg 3

12209 Berlin

Germany