# Package ‘cpp11tesseract’

October 22, 2024

**Type** Package

**Title** Open Source OCR Engine

**Version** 5.3.2

**Description** Bindings to 'Tesseract':
a powerful optical character recognition (OCR) engine that supports over 100 languages. The engine is highly configurable in order to tune the detection algorithms and obtain the best possible results.

**License** Apache License (>= 2)

**URL** https://pacha.dev/cpp11tesseract/

**BugReports** https://github.com/pachadotdev/cpp11tesseract/issues

**SystemRequirements** Tesseract >= 4.0.0 (libtesseract-dev /
tesseract-devel) and Leptonica (libleptonica-dev /
leptonica-devel). On Debian you need to install the English and
other languages training data separately (e.g.
tesseract-ocr-eng or tesseract-ocr-spa).

**Imports** pdftools (>= 1.5), curl, digest

**LinkingTo** cpp11

**RoxygenNote** 7.3.1

**Suggests** magick (>= 1.7), spelling, knitr, tibble, rmarkdown

**Encoding** UTF-8

**VignetteBuilder** knitr

**Language** en-US

**NeedsCompilation** yes

**Author** Jeroen Ooms [aut] (<https://orcid.org/0000-0002-4035-0289>),
Mauricio Vargas Sepulveda [aut, cre]
(<https://orcid.org/0000-0003-1017-7574>),
Munk School of Global Affairs and Public Policy [fnd]

**Maintainer** Mauricio Vargas Sepulveda <m.sepulveda@mail.utoronto.ca>

**Repository** CRAN

**Date/Publication** 2024-10-22 13:40:02 UTC

# Contents

---

`cpp11tesseract-package`

*Open Source OCR Engine*

---

## Description

Bindings to 'Tesseract': a powerful optical character recognition (OCR) engine that supports over 100 languages. The engine is highly configurable in order to tune the detection algorithms and obtain the best possible results.

## Author(s)

**Maintainer**: Mauricio Vargas Sepulveda `<m.sepulveda@mail.utoronto.ca>` (ORCID)

Authors:

- Jeroen Ooms `<jeroen@berkeley.edu>` (ORCID)

Other contributors:

- Munk School of Global Affairs and Public Policy [funder]

## See Also

Useful links:

- `https://pacha.dev/cpp11tesseract/`
- Report bugs at `https://github.com/pachadotdev/cpp11tesseract/issues`

## ocr *Tesseract OCR*

### Description

Extract text from an image. Requires that you have training data for the language you are reading. Works best for images with high contrast, little noise and horizontal text. See tesseract wiki and our package vignette for image preprocessing tips.

### Usage

```
ocr(image, engine = tesseract("eng"), HOCR = FALSE)

ocr_data(image, engine = tesseract("eng"))
```

### Arguments

| | |
|---|---|
| image | file path, url, or raw vector to image (png, tiff, jpeg, etc) |
| engine | a tesseract engine created with `tesseract()`. Alternatively a language string which will be passed to `tesseract()`. |
| HOCR | if `TRUE` return results as HOCR xml instead of plain text |

### Details

The `ocr()` function returns plain text by default, or hOCR text if hOCR is set to `TRUE`. The `ocr_data()` function returns a data frame with a confidence rate and bounding box for each word in the text.

### Value

character vector of text extracted from the image

### References

Tesseract: Improving Quality

### See Also

Other tesseract: `tesseract()`, `tesseract_download()`

### Examples

```
# Simple example
file <- system.file("examples", "testocr.png", package = "cpp11tesseract")
text <- ocr(file)
cat(text)
```

## tesseract *Tesseract Engine*

### Description

Create an OCR engine for a given language and control parameters. This can be used by the ocr and ocr_data functions to recognize text.

### Usage

```
tesseract(
  language = "eng",
  datapath = NULL,
  configs = NULL,
  options = NULL,
  cache = TRUE
)

tesseract_params(filter = "")

tesseract_info()
```

### Arguments

`language` string with language for training data. Usually defaults to eng

`datapath` path with the training data for this language. Default uses the system library.

`configs` character vector with files, each containing one or more parameter values. These config files can exist in the current directory or one of the standard tesseract config files that live in the tessdata directory. See details.

`options` a named list with tesseract parameters. See details.

`cache` speed things up by caching engines

`filter` only list parameters containing a particular string

### Details

Tesseract control parameters can be set either via a named list in the `options` parameter, or in a `config` file text file which contains the parameter name followed by a space and then the value, one per line. Use `tesseract_params()` to list or find parameters. Note that that some parameters are only supported in certain versions of libtesseract, and that invalid parameters can sometimes cause libtesseract to crash.

### Value

no return value, called for side effects

no return value, called for side effects

list with information about the tesseract engine

### See Also

Other tesseract: `ocr()`, `tesseract_download()`

### Examples

```
tesseract_params("debug")
```

---

## `tesseract_download` *Tesseract Training Data*

---

### Description

Helper function to download training data from the official tessdata repository. On Linux, the fast training data can be installed directly with yum or apt-get.

Helper function to download training data from the contributed tessdata_contrib repository.

### Usage

```
tesseract_download(
  lang,
  datapath = NULL,
  model = c("fast", "best"),
  progress = interactive()
)

tesseract_contributed_download(
  lang,
  datapath = NULL,
  model = c("fast", "best"),
  progress = interactive()
)
```

### Arguments

| | |
|---|---|
| `lang` | three letter code for language, see tessdata repository. |
| `datapath` | destination directory where to download store the file |
| `model` | either `fast` or `best` is currently supported. The latter downloads more accurate (but slower) trained models for Tesseract 4.0 or higher |
| `progress` | print progress while downloading |

### Details

Tesseract uses training data to perform OCR. Most systems default to English training data. To improve OCR performance for other languages you can to install the training data from your distribution. For example to install the spanish training data:

- tesseract-ocr-spa (Debian, Ubuntu)
- `tesseract-langpack-spa` (Fedora, EPEL)

On Windows and MacOS you can install languages using the tesseract_download function which downloads training data directly from github and stores it in a the path on disk given by the `TESSDATA_PREFIX` variable.

### Value

no return value, called for side effects

no return value, called for side effects

### References

tesseract wiki: training data

tesseract wiki: training data

### See Also

tesseract_download

Other tesseract: `ocr()`, `tesseract()`

Other tesseract: `ocr()`, `tesseract()`

### Examples

```
# download the french training data

  tesseract_download("fra", model = "best", datapath = tempdir())


if (any("fra" %in% tesseract_info()$available)) {
  french <- tesseract("fra")
  file <- system.file("examples", "french.png", package = "cpp11tesseract")
  text <- ocr(file, engine = french)
  cat(text)
}
# download the polytonic greek training data

  tesseract_contributed_download("grc_hist", model = "best", datapath = tempdir())


if (any("grc_hist" %in% tesseract_info()$available)) {
  greek <- tesseract("grc_hist")
  file <- system.file("examples", "polytonicgreek.png", package = "cpp11tesseract")
  text <- ocr(file, engine = greek)
  cat(text)
}
```

# Index