Home

Jump to bottom

These wiki pages are no longer maintained. All pages were moved to tesseract-ocr/tessdoc.

The latest documentation is available at https://tesseract-ocr.github.io/.

Introduction

Tesseract is an open source text recognition (OCR) Engine, available under the Apache 2.0 license. It can be used directly, or (for programmers) using an API to extract printed text from images. It supports a wide variety of languages.

Tesseract doesn't have a built-in GUI, but there are several available from the 3rdParty page.

Installation

There are two parts to install, the engine itself, and the training data for a language.

Linux

Tesseract is available directly from many Linux distributions. The package is generally called **'tesseract'** or **'tesseract-ocr'** - search your distribution's repositories to find it. Thus you can install Tesseract 4.x and its developer tools on Ubuntu 18.x bionic by simply running:

```
sudo apt install tesseract-ocr
sudo apt install libtesseract-dev
```