Tesseract OCR

build passing build passing sw failing

coverity passed code quality: c/c++ A+ lgtm alerts 32 oss-fuzz fuzzing

license Apache-2.0 download all releases

Table of Contents

About

This package contains an **OCR engine** - `libtesseract` and a **command line program** - `tesseract`. Tesseract 4 adds a new neural net (LSTM) based OCR engine which is focused on line recognition, but also still supports the legacy Tesseract OCR engine of Tesseract 3 which works by recognizing character patterns. Compatibility with Tesseract 3 is enabled by using the Legacy OCR Engine mode (--oem 0). It also needs traineddata files which support the legacy engine, for example those from the tessdata repository.

The lead developer is Ray Smith. The maintainer is Zdenko Podobny. For a list of contributors see AUTHORS and GitHub's log of contributors.

Tesseract has **unicode (UTF-8) support**, and can **recognize more than 100 languages** "out of the box".

Tesseract supports **various output formats**: plain text, hOCR (HTML), PDF, invisible-text-only PDF, TSV and ALTO (the last one - since version 4.1.0).

You should note that in many cases, in order to get better OCR results, you'll need to **improve the quality of the image** you are giving Tesseract.

This project **does not include a GUI application**. If you need one, please see the 3rdParty documentation.

Tesseract **can be trained to recognize other languages**. See Tesseract Training for more information.

Brief history

Tesseract was originally developed at Hewlett-Packard Laboratories Bristol and at Hewlett-Packard Co, Greeley Colorado between 1985 and 1994, with some more changes made in 1996 to port to Windows, and some C++izing in 1998. In 2005 Tesseract was open sourced by HP. From 2006 until November 2018 it was developed by Google.

Major version 5 is the current stable version and started with release 5.0.0 on November 30, 2021. Newer minor versions and bugfix versions are available from GitHub.

Latest source code is available from main branch on GitHub. Open issues can be found in issue tracker, and planning documentation.

See **Release Notes** and **Change Log** for more details of the releases.

Installing Tesseract

You can either Install Tesseract via pre-built binary package or build it from source.

A C++ compiler with good C++17 support is required for building Tesseract from source.

Running Tesseract

Basic **command line usage**:

```
tesseract imagename outputbase [-l lang] [--oem ocrenginemode] [--psm pagesegmode]
[configfiles...]
```

For more information about the various command line options use `tesseract --help` or `man tesseract`.

Examples can be found in the documentation.

For developers

Developers can use `libtesseract` C or C++ API to build their own application. If you need bindings to `libtesseract` for other programming languages, please see the wrapper section in the AddOns documentation.

Documentation of Tesseract generated from source code by doxygen can be found on tesseract-ocr.github.io.

Support

Before you submit an issue, please review **the guidelines for this repository**.

For support, first read the documentation, particularly the FAQ to see if your problem is addressed there. If not, search the Tesseract user forum, the Tesseract developer forum and past issues, and if you still can't find what you need, ask for support in the mailing-lists.

Mailing-lists:

- tesseract-ocr - For tesseract users.
- tesseract-dev - For tesseract developers.

Please report an issue only for a **bug**, not for asking questions.

License

```
The code in this repository is licensed under the Apache License, Version 2.0 (the
"License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at
```

NOTE: This software depends on other packages that may be licensed under different open source licenses.

Tesseract uses Leptonica library which essentially uses a BSD 2-clause license.

Dependencies

Tesseract uses Leptonica library for opening input images (e.g. not documents like pdf). It is suggested to use leptonica with built-in support for zlib, png and tiff (for multipage tiff).

Latest Version of README

For the latest online version of the README.md see:

https://github.com/tesseract-ocr/tesseract/blob/main/README.md