Intro tesseract-OCR

Tesseract is an optical character recognition engine for various operating systems. It is free software, released under the Apache License, Version 2.0, and development has been sponsored by Google since 2006. In 2006 Tesseract was considered one of the most accurate open-source OCR engines then available.

Tesseract was in the top three OCR engines in terms of character accuracy in 1995. It is available for Linux, Windows and Mac OS X. However, due to limited resources it is only rigorously tested by developers under Windows and Ubuntu.

Tesseract up to and including version 2 could only accept TIFF images of simple one-column text as inputs. These early versions did not include layout analysis, and so inputting multi-columned text, images, or equations produced garbled output. Since version 3.00 Tesseract has supported output text formatting, hOCR[9] positional information and page-layout analysis. Support for a number of new image formats was added using the Leptonica library. Tesseract can detect whether text is monospaced or proportionally spaced.

The initial versions of Tesseract could only recognize English-language text. Tesseract v2 added six additional Western languages (French, Italian, German, Spanish, Brazilian Portuguese, Dutch). Version 3 extended language support significantly to include ideographic (Chinese & Japanese) and right-to-left (e.g. Arabic, Hebrew) languages, as well as many more scripts. New languages included Arabic, Bulgarian, Catalan, Chinese (Simplified and Traditional), Croatian, Czech, Danish, German (Fraktur script), Greek, Finnish, Hebrew, Hindi, Hungarian, Indonesian, Japanese, Korean, Latvian, Lithuanian, Norwegian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak (standard and Fraktur script), Slovenian, Swedish, Tagalog, Tamil, Thai, Turkish, Ukrainian and Vietnamese. V3.04, released in July 2015, added an additional 39 language/script combinations, bringing the total count of support languages to over 100. New language codes included: amh (Amharic), asm (Assamese), aze_cyrl (Azerbaijana in Cyrillic script), bod (Tibetan), bos (Bosnian), ceb (Cebuano), cym (Welsh), dzo (Dzongkha), fas (Persian), gle (Irish), guj (Gujarati), hat (Haitian and Haitian Creole), iku (Inuktitut), jav (Javanese), kat (Georgian), kat_old (Old Georgian), kaz (Kazakh), khm (Central Khmer), kir (Kyrgyz), kur (Kurdish), lao (Lao), lat (Latin), mar (Marathi), mya (Burmese), nep (Nepali), ori (Oriya), pan (Punjabi), pus (Pashto), san (Sanskrit), sin (Sinhala), srp_latn (Serbian in Latin script), syr (Syriac), tgk (Tajik), tir (Tigrinya), uig (Uyghur), urd (Urdu), uzb (Uzbek), uzb_cyrl (Uzbek in Cyrillic script), yid (Yiddish). [10] Tesseract can be trained to work in other languages too.

CUSTOMER SIGNATURE