Google's Optical Character Recognition (OCR) software works for 248+ languages

Posted 18 Sep 2015 | Subhashish Panigrahi (/users/psubhashish) | 72 | 1 comment

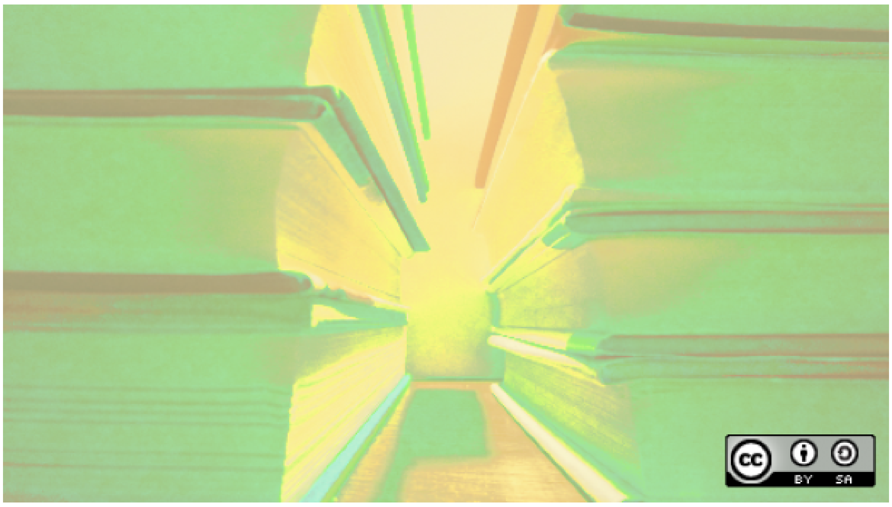


Image by :

Image by Kate Ter Haar. (https://www.flickr.com/photos/katerha/5020407401/in/photolist-8DCUdi-7i3R7t-7acJMF-6Xr2zN-cCq3d9-9k6cM-7i7Ky9-7vap8K-dLZA9U-7uNiS7-5hKBR3-7Ay9pH-5JaKKJ-dMTqKC-dMTqNo-q5c8gP-7uJCRF-6Vc3mh-HN51N-7i3R3P-7i7KLN-6BW9qM-8BdMFv-9ggca5-7i7Kth-oVAXz-4TiJLN-4TiKwd-3D7p16-8pZB2Q-5ZQ7Hn-mmgwkx-96kf8y-7ZKwDf-6Vc4C9-b3SXpF-96hcRH-6owcpY-64vPQy-7vedkj-7uJtMv-89D449-4fmB8N-7i3RcZ-s5uC3e-6p5sXJ-9mWtBt-cZhQH-6nkkba-8gXK1y) Modified by opensource.com. CC BY-SA 2.0. (https://creativecommons.org/licenses/by-sa/2.0/)

About the author

Subhashish Panigrahi - Subhashish Panigrahi (@subhapa (http://twitter.com/subhapa is the Asia Community Catalyzer at Mozilla's Participation team, and is transitioning from his role as Programme Officer of the Centre for Internet and Society (https://cis-india.org)'s Access To Knowledge program (https://meta.wikimedia.or Earlier with Wikimedia Foundation's India Program, he is an India based educator,

» More about me (/users/psubhashish)

» Learn how you can contribute (/participate)

Google's Optical Character Recognition (https://support.google.com/drive/answer/176692) (OCR) software now works for over 248 world languages (including all the major South Asian languages). It's quite simple and easy to use, and can detect most languages with over 90% accuracy.

The technology extracts text from images, scans of printed text, and even handwriting, which means text can be extracted from pretty much any old books, manuscripts, or images.

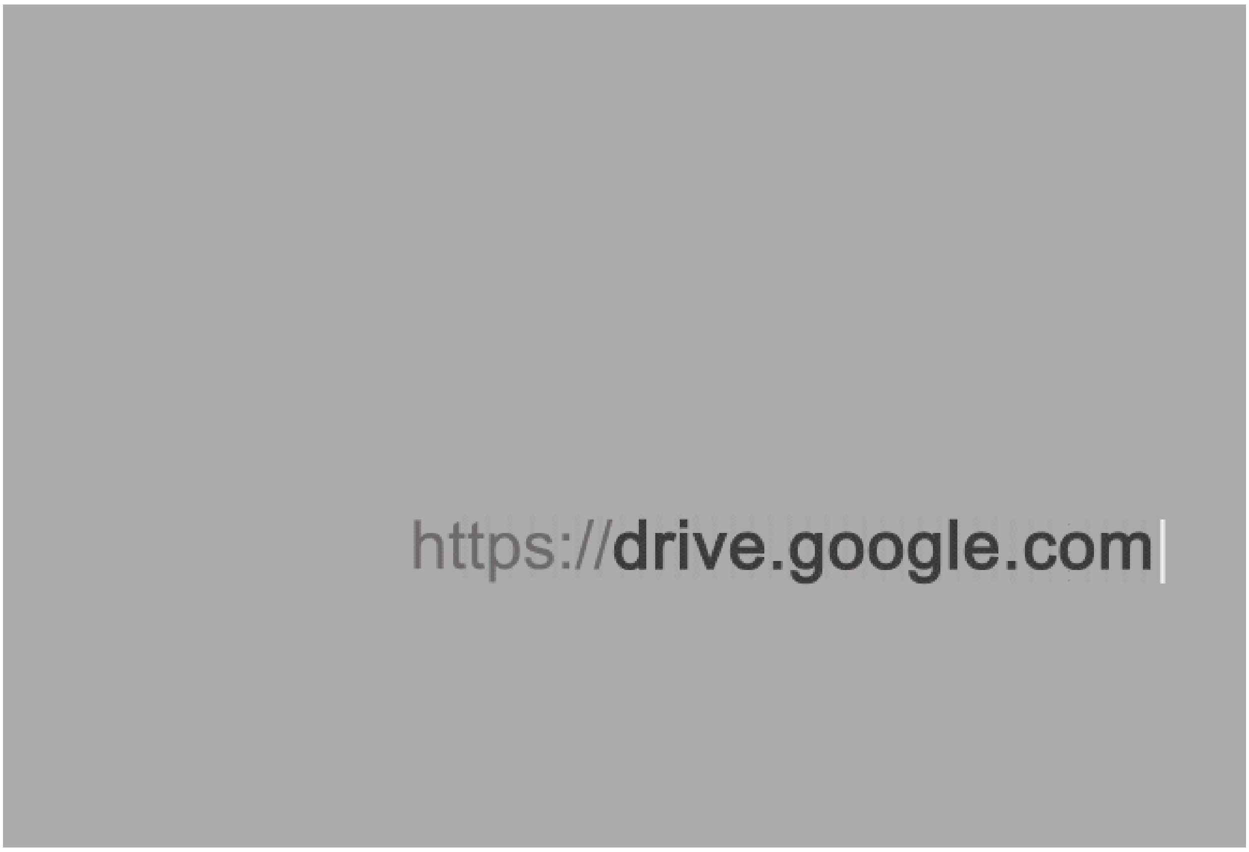
Google's OCR is probably using dependencies of Tesseract (https://en.wikipedia.org/wiki/Tesseract_(software)), an OCR engine released as free software, or OCRopus (https://en.wikipedia.org/wiki/OCRopus), a free document analysis and optical character recognition (OCR) system that is primarily used in Google Books (https://books.google.com/). Developed as a community project during 1995-2006 and later taken over by Google (https://code.google.com/archive/p/tesseract-ocr/), Tesseract is considered one of the most accurate OCR engines and works for over 60 languages. The source code is available on GitHub (https://github.com/tesseract-ocr).

The OCR project support page (https://support.google.com/drive/answer/176692) offers additional details on preserving character formatting for things like bold and italics after OCR in the output text:

> When processing your document, we attempt to preserve basic text formatting such as bold and italic text, font size and type, and line breaks. However, detecting these elements is difficult and we may not always succeed. Other text formatting and structuring elements such as bulleted and numbered lists, tables, text columns, and footnotes or endnotes are likely to get lost.

Tamil-language Wikimedian and Wikimedia India's program director Ravishankar Ayyakkannu said on Facebook (https://www.facebook.com/ravidreams/posts/10154278945453569)this after testing: "For some of the languages like Malayalam and Tamil, the OCR works with almost 100% accuracy, along with support in formatting like auto cropping, separating text by discarding images, and ignoring colored backgrounds." Native speakers of the following Indian lanaguages—Bangla, Malayalam, Kannada, Odia, Tamil, and Telugu—also commented on a Facebook post with feedback after testing the OCR.

However, for a few scripts like Gurmukhi (used to write Punjabi), the output after OCR is quite poor and results in gibberish text in different scripts.



A tutorial to convert text in Odia (Indian language) from a scanned image using Google's OCR. Designed by Subhashish Panigrahi. [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0/)](https://creativecommons.org/licenses/by-sa/4.0/)

Overall, this is quite a large leap for languages that have old texts that have not yet been digitized. Old and valuable text in many languages can now be digitized and shared over the internet using platforms like [Wikisource (https://wikisource.org/wiki/Main_Page)](https://wikisource.org/wiki/Main_Page).

Editor's note: Article has been updated based on community feedback. We changed "Google's OCR partly uses Tesseract, an OCR engine released as free software" to "Google's OCR is probably using dependencies of [Tesseract (https://en.wikipedia.org/wiki/Tesseract_(software))](https://en.wikipedia.org/wiki/Tesseract_(software)), an OCR engine released as free software, or [OCRopus (https://en.wikipedia.org/wiki/OCRopus)](https://en.wikipedia.org/wiki/OCRopus), a free document analysis and optical character recognition (OCR) system that is primarily used in [Google Books (https://books.google.com/)](https://books.google.com/)." If you have additional feedback on the article or technology, please let us know in the comments. -Rikki Endsley

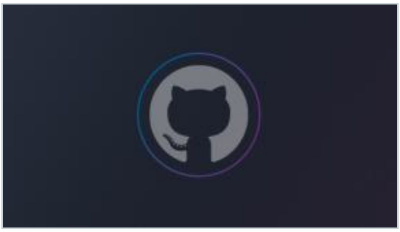
Tags:

Life (/tags/life), Google (/tags/google)

Recommended reading

A tour of Google's 2016 open source releases (/16/12/yearbook-tour-googles-2016-open-source-releases)

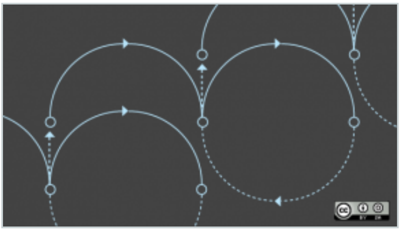
Win a 3D printer! Enter the 2016 open source holiday giveaway (/life/16/11/2016-holiday-gift-guide-sweepstakes)



How to build your code club on GitHub (/life/16/11/github-organizations)

Education management with Moodle: The beginning, middle, and today (/life/16/11/moodle-today)

Learn Perl with this temperature-conversion script (/life/16/11/learn-perl-temperature-conversion-script)



Why design and marketing matter and what to do about it (/life/16/11/all-things-open-keynotes-day-2-

1 Comments

Suraj Sh on 05 Nov 2015

how about devanagri or hindi/marathi. Nothing mentioned about this?

0 0

SIGN UP FOR OPENSOURCE.COM NEWS

Continue

Privacy Policy | Terms of Use | Contact | Meet the Team | Visit opensource.org |

Find us: