# cpp11tesseract: An R Package for OCR Using Tesseract

Mauricio Vargas

2018-08-23

## Statement of Need

Optical Character Recognition (OCR) eases extracting text from images and PDFs, especially in fields like public policy that often deals with scanned government documents. Despite the availability of various OCR tools, many are sensitive to image resolution, have a monetary cost per document as these use cloud services, and struggle with non-English documents. `cpp11tesseract` leverages the power of Tesseract to address these challenges within the R ecosystem, providing high-accuracy text extraction for a wide range of documents and it allows to use personalized models for OCR.

## Summary

Tesseract is a popular open-source OCR engine currently maintained by Google that supports over 100 languages and is widely used for extracting text from images and PDFs (Tesseract 2024). `cpp11tesseract` is an R package that provides a simple interface to Tesseract-OCR, enabling R users to perform OCR on images and PDFs directly within their R scripts.

This package offers a range of features, including support for multiple languages, configurable options, and the ability to use personalized models for OCR. The last point is relevant, as (Hegghammer 2021) states: "Established OCR libraries such as Tesseract are highly sensitive to noise and often require extensive corpus-specific adaptations to render text accurately."

There are existing and well-maintained implementations in R for OCR, including packages to use Tesseract (Ooms 2024) and Amazon Textract (Kretch and Banker 2021). The main difference is that `cpp11tesseract` uses header-only C++ bindings to the Tesseract engine (Vaughan, Hester, and François 2024), which makes it faster and more efficient than previous implementations, and it does not depend on an internet connection. In restricted environments, such as the Niagara Cluster, this header-only approach can be a significant advantage as it

allows to install this package with a minimal number of dependencies including all the code to compile it by using vendoring.

As a way to contribute to open-source software, this package has resulted in different contributions to Ooms (2024), including the addition of new features and bug fixes, but we maintain different approaches regarding the concept and internal design.

## Key Features

### Simple Usage

Running OCR on an image is straightforward:

```
library(cpp11tesseract)

ocr("https://pacha.dev/cpp11tesseract/reference/figures/testocr.png")
```

### Multi-Language Support

To improve OCR accuracy for non-English documents, **cpp11tesseract** allows users to easily install additional language training data. For example, Romanian training data can be installed with:

```
tesseract_download('ron', model = "best")
```

### Ensembled Models

An important feature is the option to use personalized model and ensembles of similar languages to improve OCR accuracy. We also provide a guide to train your own models.

```
# assuming roncustom.traineddata is a trained model with a subset of your own
# documents in Romanian

pdfs <- list.files("documents")

languages <- "ron+ita+spa+fra+lat+roncustom"

for (p in pdfs) {
  parsed <- ocr(p, engine = tesseract(languages, datapath = getwd()))
  writeLines(parsed, gsub(".pdf", ".txt", p))
}
```

## Conclusion

`cpp11tesseract` is a robust R package for OCR, enabling R users to leverage the power of Tesseract-OCR directly within their workflows. Its support for multiple languages, configurable options, and ease of integration with other R tools make it an addition to the R ecosystem, especially for researchers and analyst that require to extract data from images and PDF documents. It can be useful for historical documents and it can help with large-scale text extraction where manual transcription can be slow.

## Acknowledgements

I am indebted to my advisor Mark Manger for understanding the value of developing a tool that facilitates the extraction of text from images and PDFs instead of dedicating the same time and effort to use regular expression to tidy scanned documents with other tools.

## References

Hegghammer, Thomas. 2021. "daiR: An R Package for OCR with Google Document AI." *Journal of Open Source Software* 6 (68): 3538. https://doi.org/10.21105/joss.03538.

Kretch, M., and A. Banker. 2021. "paws: Amazon Web Services Software Development Kit." https://cran.r-project.org/package=paws.

Ooms, J. 2024. "tesseract: Open Source OCR Engine." https://cloud.r-project.org/web/packages/tesseract/index.html.

Tesseract. 2024. "Tesseract OCR." GitHub. https://github.com/tesseract-ocr/tesseract.

Vaughan, Davis, Jim Hester, and Romain François. 2024. *Cpp11: A c++11 Interface for r's c Interface*. https://CRAN.R-project.org/package=cpp11.