doi:10.20396/revpibic.v0i0.id

XXVI Congresso de Iniciação Científica Unicamp

17 a 19 de outubro Campinas | Brasil

Evaluation of OCR Free Software Applied to Old Books

Pedro H. Barcha Correia*, Gerberth Adín Ramírez Rivera

Abstract

This project compares state-of-the-art Free Software Optical Character Recognition (OCR) programs. Particularly, their results over old books pages were evaluated. Moreover, in order to optimize the recognition for this kind of data input, methods that are not implemented in the programs were proposed and their results were analyzed as well.

Key words:

Optical Character Recognition, Old Books, Free Software

Introduction

OCR (Optical Character Recognition) is the automated process of recognizing text blocks in images.

This research evaluates the two programs regarded as the state-of-the-art in the Free Software OCR area: Tesseract and OCRopus. In particular, their results were evaluated over an old books pages dataset, that we created[1] and provided to the programs' communities.

This project aims at optimizing the recognition for old books, hence allowing people and libraries to transform public domain physical books into eBooks and PDFs. Additionally, the results obtained here are of great value for the programs' communities.

Results and Discussion

Our evaluation metric was the CER (Character Error Rate), which is the percentage of characters that were incorrectly recognized by the OCR process.

Binarization

Several binarization methods were tested; the one with the best results is shown on Chart 1.

We noticed that OCRopus' original method does not work well for pages with pictures, as it does not bleach them, leading the processing steps to find false positives.

Chart 1. OCR results using the original programs' binarization methods and using the Minimum method.

	Original Binarization	Minimum
Tesseract CER	2,71 %	2,71 %
OCRopus CER	7,91 %	6,17 %

Noise Reduction

OCRopus and Tesseract do not have a denoising step. To find out which were the best solutions to enhance noisy old books pages, several filters were tested to images containing, separately, Gaussian, Salt and Pepper and Speckle-like . We had the best results for the Gaussian noise reduction, described in Chart 2.

Chart 2. OCR results for the dataset with gaussian noise before and after the bilateral filter was applied.

	Noisy Dataset	Bilateral Filter
Tesseract CER	24,13 %	4,37 %
OCRopus CER	44,26 %	12,20 %

Training

A second dataset of old books pages was created and used to train OCRopus (which uses LSTM Neural Networks) and Tesseract (polygonal approximation approach). Only the model generated by the former yielded CER reduction (of 0,32%), for the main dataset.

This shows how powerful LSTM Neural Networks are for the OCR task. This approach in currently being implemented by Tesseract's community.

Post-Processing

Bassil and Alwani[3] claim that OCR transcriptions can be post-processed by spell checking it with online search engines (as google's “did you mean...”, for instance). The transcribed text should be segmented into 5 words blocks and queried to the engine. A program was developed for this purpose[2], however the results for the dataset were not positive, as suggested by Bassil and Alwani. A possibly good alternative to be tested in future works would be to chop the transcribed text according to its punctuation, thus preserving the sentence's context.

Conclusions

The results obtained from this project were particularly helpful to enhance OCRopus performance on old books pages and thus of great value to its community. Its CER was reduced by 42%, accounting all of the improvements proposed. The only step that the program had better results than Tesseract was for the training step.

Although the online post-processing experiment did not yield satisfactory results, it might turn out to be a good solution by using different approaches.

Finally, the results shown by Tesseract of 2,71% CER suggest that it might soon be a reliable tool to turn physical public domain book into open access eBooks.

Acknowledgement

Professor Neucimar Leite, who mentored me in the first part of this project is gratefully acknowledged. Thanks for embracing and motivating me so many times during my academic life.

[1] BARCHA, Pedro. Old-books-dataset. Available at: <https://github.com/PedroBarcha/Old-Books-Dataset>.

[2] BARCHA, Pedro. Context-spelling-correction. Available at: <https://github.com/PedroBarcha/context-spelling-correction>.

[3] BASSIL, Youssef; ALWANI, Mohammad. OCR Post-Processing Error Correction Using Google Web 1T 5-Gram Data Set. USA. Pp 14-25. 2012.

(cc) BY-NC-ND