How far has Optical Character Recognition Softwares have come.[*]

Evaulation of OCR software

Sampson Zhao

25 April 2022

Abstract

Optical Character Recognition (OCR) has been used since the 1930s as a way to recognize standardized text from a photo. This technology has greatly developed throughout the ages, and is now widely used in data entry from printed data, like passports, invoices, etc. Though the earlier versions of OCR required the text to have a set font for recognition, modern iterations of this has developed enough to recognize multiple fonts as well as different writing systems. This project focuses on the accuracy of modern OCR systems, like blank and blank, and will compare the accuracy of the produced text to the originals. (conclusion) Though the softwares…

Contents

[*] Code and data are available at: https://github.com/Sampsonzhao/OCR-final.

1 Introduction

OCR was first developed as a technology in the 1910s, as it was primarily used in tandem with reading devices for the blind (**OCRfirst?**), with the main premise of OCR being using some optical device to capture characters on a document, and transform it into a manner where computer software can then recognize the text. Though earlier iterations of OCR were used in the development of text documents to audio readers for the blind or visually impaired, the technology is now widely used as a shortcut in data entry, as documents like passports, cheques, and other printed documentations, can be quickly and accurately transformed, effectively digitizing the document to allow for editing and searching electronically. Though earlier iterations of the software required documents to be in particular fonts for the software to recognize the text, more modern and advanced systems have developed significantly, to the point where these softwares are able to recognize texts in a variety of fonts within the same documen, as well as recognizing different writing systems, like Cyrillic, Arabic, etc. (List of most of the available languages that are recognized by Google's OCR software here).

OCR development has been progressing fairly well, as the accuracy of the software from 19th century to the 20th century has seen an increase of 81% accuracy to 99% accuracy. Though this is comparing the developments of the technologies that these softwares are based off of, there was a fundamental shift in the style of recognition. OCR developed from a character-to-character recognition style to a more pattern recognition style, signifying the advancements of computational power, as effectively they changed from a character-to-character recognition to more of a line-by-line recognition. Though the recognition of text improved drastically back then, there were multiple variables that caused the 1% of texts to fail, and that is related to the pre-processing of the image. Factors like wrinkles, slanting, normalization of ratio and scales, were all factors that attributed to failed OCR reads, these were considered artifacts that the software could not recognize, effectively showing garbage code as a result. Though these factors have improved since the early 2000s, OCR is still continuing on the development of optimizing the overall recognition software.

Ultimately, this paper looks into the development of OCR software that are open access today, effectively looking at the accuracy of the text that is produced by the OCR software, when compared to the original text. We are able to use this information to track the accuracy of these software, which would be useful in digitizing older texts. Here, we would look at the difference in processed and unprocessed images of text documents from different titles and compare the accuracy of OCR, as well as testing the limits of OCR. The continued development of OCR aims to do this as one of it's main applications, as it would allow for further digitization of older texts, allowing more of this information to be share on the web. As many older texts have only photo renditions of the book online, OCR can be used to scan these pictures and create a searchable and interactable version of the text.

While many old texts show faded or blurred text, due to the age of the documents themselves, this posed an issue for both old and newer versions of OCR, as the pre-processing of the image would result in the software returning errors alongside the text. There are tweaks that are currently in development, as development of application-specific OCR has come up as one of the main projects developed by government entities, like recognizing license plate numbers or invoice specifics. These developments allow for the specialization of the software, which trims down on the processing time needed for the software to recognize these texts. The hope is that further development of application-specific OCR and effectively transition to further optimizing the recognition of older texts, allowing for the continued digitization of historical texts.

2 Data

2.1 Acknowledgement

2.2 Where this data comes from

2.3 Understanding the Data

3 Discussion

3.1 Problems with OCR

While OCR is a very convenient piece of software that allows us to quickly translate information from a photo into an interactable piece of digital text, there are still many flaws associated to the software itself.

3.2 Progress with OCR

3.3 Further Development of OCR alongside the development of A.I.

4 Conclusion

5 Apendix

6 References