0
CC-
SXML
acks

Ac-
knowl-
edg-
ments

screenonly printonly anonsuppress

Orville "El" Anderson, and10393@umn.edu,

This paper looks at Optical Character Recognition(OCR) as a means to improve usability and accessibility of Scanned Documents. The focus is on how to compare OCR implementations, the weaknesses of OCR, and how we can create tests to specifically target those weaknesses.

Optical Character Recognition

Orville "El" Anderson
and10393@umn.edu
Division of Science and Mathematics
University of Minnesota, Morris
Morris, Minnesota, USA

Abstract

Keywords: optical character recognition, scanned documents, visual noise, datasets

1 Introduction

Scanned documents are images of physical documents typically made with a scanner or camera. These documents are widely used because they are easy to create and share. Scanned documents have some hidden disadvantages compared to [native online documents]. The most notable of these disadvantages are larger storage requirements, longer loading times, and the inability to search within the text.[1]

While manually processing(typyign up) doccuments is a valid way to digitize it, it does not scale well. This is why we have Optical Character Recognition(OCR), a technology specifically made to extract text from images.

2 Background

OCR is made of three main steps (plus acquisition and output). [2]

2.1 Document Layout Analysis

Document layout analysis, DLA, is a general pre-processing step. (Besides taking high quality scans to begin with), this step has the largest impact of accuracy. The purpose of this step is to identify what part of the image is text and what is not. This step exists to remove any visual noise generated when scanning(or handling) the physical document and is sometimes used to remove tables and images from the page.

Pre-processing (aka cleaning) a document is one of the most important steps to ensure accuracy. (along with just taking good scans...). here you remove page borders, and rotate the image if applicable.

2.2 Text Line Detection

this step consists of breaking up blocks of text into lines of text, then into individual words, then letters.

After the individual characters have been identified, we go through and look at each of the boxes we have drawn around the characters and mark which pixels we think are text and which are not.

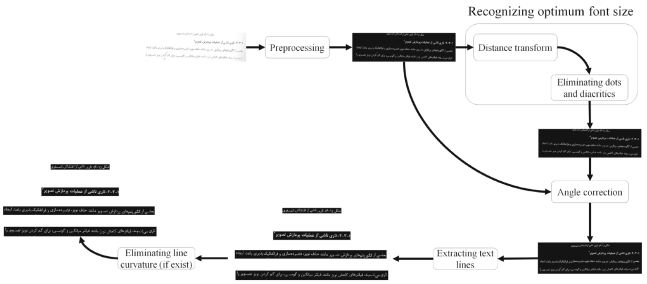

Figure 1. stages of TLD

2.3 Recognition

We then take the mega-cleaned characters and try to identify them. There are many ways to do this, but it boils down to matching your letter to a whole bunch of other letters and seeing which one is closest.

One very common method for character recognition is to use a Deep Learning. (woo magic). The notable thing about this method is that it must be intentionally fed examples of text to train.

2.4 Post-Processing

Post-processing is an optional, but very common step for OCR. This is glorified spellcheck for the output. It generally improves accuracy, but will write over spellings from the source document(misspellings and unconventional spellings included.)

2.5 Output

The output of OCR is given in many different ways. One of the more popular outputs is as the accuracy percentage.

The output of OCR is a text document [**? ? ?**] [inline]Formula?

3 Challenges

There are three main categories of things that make scanned documents harder to digitize.

3.1 Layout

There exists many ways to format a paper.

[1] also you can't use a screen reader on them

[2] soemthign to note here is that the input for OCR is generally a .pdf, .jpg, ... and the output can be .html, .pdf, ...

3.2 Alphabet

Most OCR are trained on the Latin Alphabet. These models are not automatically applicable to other alphabets. As this paper[**?**] says about Arabic, "Finally, character and TLD approaches cannot be universally applied to scripts with different languages. This challenge is particularly pronounced in languages such as Persian and Arabic, where text characters can take the form of connectors or non-connectors. Connector letters are affixed to both pre- and post-letters to form words, and diacritic marks are commonly used in Arabic text—both of which can introduce complexity to TLD techniques." This inherently gives all other languages a bit of a disadvantage. Other alphabets also have nuances that make it harder to directly apply OCR to. A common example is Arabic. (there's dots) [**?** **?**]

A more universal example of this trait is cursive.

3.3 Visual Noise

There's so many ways to scan a document badly. [**?**]

4 Results

Results - some specialized datasets exist [**?** **?**]

5 Conclusion

Conclusion - While it's ideal to have one golden benchmarking set, that's hard (because of the reasons outlined above), so now we have specialized ones. Acknowledgments

Thanks.