Image-based Methods for Character Recognition

TREO Talk Paper

Olusola Samuel-Ojo
Claremont Graduate University
Olusola.samuel-ojo@alumni.cgu.edu

Lorne Olfman
Claremont Graduate University
Lorne.olfman@cgu.edu

Efosa C. Idemudia
Arkansas Tech University
eidemudia@atu.edu

Abstract

Commercial optical character recognition (OCR) systems often fail to deliver sufficient results. Their processing algorithms are generally optimized toward large-scale commercially relevant collections. They provide inadequate outputs in recognizing unconstrained writings, blurry images, infrequent image classes and similarities between character images. Though OCR is widely used and breakthrough results have been reported in areas such as speech recognition, the accuracy of word and character recognition is as good as that of a second-grade child. Thus, in this review, we address the question of whether there are essential characteristics of current high performing OCR algorithms that may support large-scale processing of related and unrelated document image batches. We review OCR systems to understand their algorithms and processes. Then we identify insights concerning the combination of parts that may produce the most effective improvement impact. We select and discuss in detail system candidates that are representative of high performing algorithms, which may provide guidance for customization, development and improvement projects. These system candidates include Tesseract OCR and hidden Markov Model (HMM) based recognition systems. Using the Tesseract algorithm, HMM and messaging theories as the theoretical review lenses, we present a new framework for constructing OCR workflows together. We also present key performance metrics that may be used to assess OCR systems. With these artifacts, practitioners in the surveillance industry, for example, may combine those most effective processes to improve recognition metadata and results, and optimize batches of both related and unrelated document images as shown in both Figures 1 and 2.

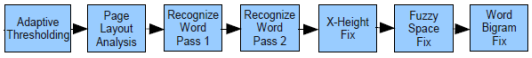

Figure 1. Tesseract architecture (Smith 2013).

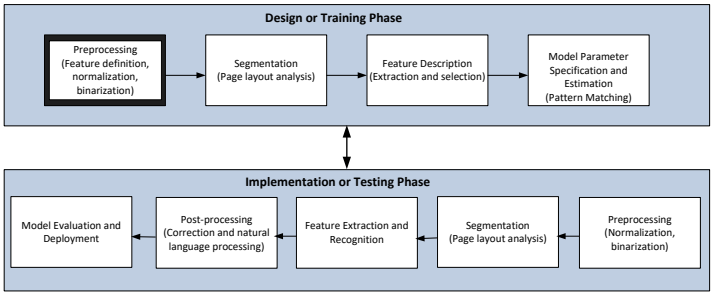

Figure 2. Analytics Framework for OCR tasks.

Index Terms— automatic number plate recognition, extraction, optical character recognition, pattern recognition, performance metrics segmentation.

References (optional)

R. W. Smith, "History of the Tesseract OCR engine: what worked and what didn't," in Proc. SPIE 8658, Document Recognition and Retrieval XX, 865802, 2013.