Decomposing Documents into \LaTeX from its Geometrical and Logical Layout

Members:
Vincent Deo (vdeo@stanford.edu)
Terry Kong (tckong@stanford.edu)
Maisy Wieman (mwieman@stanford.edu)
Android Device: Yes

Motivation: One popular method of producing aesthetically pleasing PDF documents including diverse contents is LaTeX. LaTeX is a low-level markup and programming language that allows high flexibility for producing equations, tables, and overall document structure. Once a document is compiled from LaTeX, there is no way in general to alter the document. Also, LaTeX is appreciated by the scientific community for allowing typesetting of mathematical formulae much faster than WYSIWYG interfaces, but however still far less efficient than hand-writing. This project aims to simplify the generation of LaTeX source code with an application that disassembles documents into hierarchized blocks and outputs the appropriate LaTeX code to reproduce a document with contents organized as the input's.

Goals: This project will extract the structure of formatted documents through segmentation techniques. Images of printed text acquired from a camera will be analyzed to generate the LaTeX code with the same layout as the original document. Preprocessing will be applied to the original image, and this step will include binarization of the input image and correction for any skew. A top-down approach will be used for the segmentation, which first decomposes a document into blocks, then classifies and interprets the blocks based on logical layout (eg. paragraphs, sections or subsection) and content (such as text, images or equations). Line and character recognition (for example, using the Tesseract OCR engine) will be implemented, and, if time permits, will be used to analyze equations in the document.

References:

[1] Nazemi, A., Murray, I., & McMeekin, D. A. (2014). Practical segmentation methods for logical and geometric layout analysis to improve scanned PDF accessibility to Vision Impaired. *International Journal of Signal Processing, Image Processing and Pattern Recognition*.

[2] Impedovo, S., Ottaviano, L., & Occhinegro, S. (1991). Optical character recognition—a survey. *International Journal of Pattern Recognition and Artificial Intelligence*, *5*(01n02), 1-24.

[3] Baird, H. S. (1992). Anatomy of a versatile page reader. *Proceedings of the IEEE*, *80*(7), 1059-1065.
[4] Simon, A.; Pret, J.-C.; Johnson, A.P., "A fast algorithm for bottom-up document layout analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* , vol.19, no.3, pp.273,277, Mar 1997. doi: 10.1109/34.584106
[5] Cattoni, R., Coianiz, T., Messelodi, S., & Modena, C. M. (1998). Geometric layout analysis techniques for document image understanding: a review. *ITC-irst Technical Report*, *9703*(09).