LaTeX Generation from Printed Equations

Oct 29, 2015

Jim Brewer (jebrewer@stanford.edu)
James Sun (jsun2015@stanford.edu)

Motivation:

LaTeX is a powerful typesetting system that is extremely useful for technical documents. In particular, the ability to recreate high quality representations of mathematical equations greatly motivates its use in the engineering fields. However, once a document as been rendered in this format the underlying code to recreate it is inaccessible without the original producer's code. It may be desirable for a graduate student or professor to be able to reproduce existing equations from textbooks, technical papers, homework assignments, etc. into their own homework assignments, solutions, or papers. Recoding a lengthy equation is time consuming and prone to error. This project aims to make this process more accessible by allowing the user to take a photograph of a printed equation and produce the required LaTeX code to reproduce the equation.

Project Description:

The LaTeX Generation project aims to automatically generate LaTeX expressions for existing photographed printed equations. For feasibility, this project has the following limitations in scope:

1. The types of mathematical equations will be limited such that only certain common operators will be recognized. The design of this project should, however, be modular enough that further expressions can be added through additional work.
2. The equation will be assumed to be the primary data in the processed image, allowing for page edges or additional noise around the edges, and will not be extracted from an entire page.

This project can be divided into three main tasks: page optimization, character recognition and LaTeX compilation. The more involved task is predicted to be character recognition. This project will take in a photograph of a printed equation, correct for sub-optimal lighting conditions or a skewed image, binarize the image and perform character recognition with our subset of mathematical operators, and then convert the found characters and expression into LaTeX code. Given time this could possibly be extended to incorporate the EE368 Spr 15 project by Deo, Kong, and Weiman to decompose a document into its component parts of text, figure, etc. and extract equations from a full paged document.

Details:

Inputs: Photographs of printed mathematical expressions

Outputs: LaTeX code for the mathematical expression

Data: Images of mathematical expressions with original LaTeX that generated them for testing. Photographs of those same expressions printed.

Implementation: Matlab. Possibly OpenCV. No Android planned.

References

Akram, M.U.; Bashir, Z.; Tariq, A.; Khan, S.A., "Geometric feature points based optical character recognition," in *Industrial Electronics and Applications (ISIEA), 2013 IEEE Symposium on* , vol., no., pp.86-89, 22-25 Sept. 2013

Jianhong Xie, "Optical Character Recognition Based on Least Square Support Vector Machine," in *Intelligent Information Technology Application, 2009. IITA 2009. Third International Symposium on* , vol.1, no., pp.626-629, 21-22 Nov. 2009

Pradhan, A.; Pradhan, M.P.; Prasad, A., "An approach for reducing morphological operator dataset and recognize Optical Character based on significant features," in *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on* , vol., no., pp.1631-1638, 10-13 Aug. 2015

Qadri, M.T.; Asif, M., "Automatic Number Plate Recognition System for Vehicle Identification Using Optical Character Recognition," in *Education Technology and Computer, 2009. ICETC '09. International Conference on* , vol., no., pp.335-338, 17-20 April 2009

Heng-You Wang; Rui-Zhen Zhao; Jing-An Cui, "Fast and robust skew correction in scanned document images based on low-rank matrix decompositon," in *Machine Learning and Cybernetics (ICMLC), 2014 International Conference on* , vol.2, no., pp.883-887, 13-16 July 2014