Project Proposal
LaTeX Generation from Printed and Handwritten Equations

ANSHUL VERMA
#1004730703

October 7^{th}. 2019

Summary

The project aims to convert images of printed and handwritten equations in to LaTeX code. So that one can add the equations in a reference file or an handwritten equations directly into their LaTeX code to generate their document.

Motivation

LaTeX is an important tool for everyone working in academics these days to generate and recreate high quality technical documents. It's use is highly recommended in all the engineering fields because it allows high quality representations of mathematical equations. But it is impossible to recreate the underlying code of a document rendered using LaTeX without the original producer's code. Thus even after referencing academic papers, textbooks and other sources people still need to manually code the equations in their documents, this process is extremely time consuming and prone to error. Thus it can be a desirable solution for students and researchers to have a tool which easily adds equations to their LaTeX document from images of equations.

Methodology

The project involves three major tasks Classification, Optical Character Recognition and LaTeX compilation. The data provided will firstly be used to build a classifier to classify the random input into printed or handwritten equations. For this purpose a basic classifier like GMM(Gaussian Mixture Model) or SVM(Support Vector Machine) classifier will be used. This classifier will be used to classify the test data. Then the more challenging task of Optimal Character Recognition will be performed. This task is extremely challenging because there are several factors like alignment and lighting of the image, size of the equation in the image and clarity of the handwritten equation which will all affect character recognition hugely. Then the final step of generating the LaTeXcode for the recognition result will be a more straight forward task.

Implementation and Data set

The majority of this project will focus on **OCR**(Optical Character Recognition) using CNN[1],[2] which will be implemented in Python. Major libraries which will be used for **OCR** are pytesseract, PyTorch and OpenCV.

There are two major data source available for this very purpose of printed and handwritten equation to LaTeXcode. First one being Im2LaTeX-100k, which is a prebuilt dataset published by a Harvard paper. It has $\approx 100k$ images of printed equations and their LaTeX code. Second is CHORME data-set from CHORME 2011 competition which consists $\approx 10k$ images of handwritten equations and their LaTeXrepresentation.

A small test-data will be generated manually to check the global accuracy on a totally random data and all the data available in the data sets will be used for training and validation.

References

[1] Guillaume Genthial and Romain Sauvestre. Image to latex. 2016.

[2] Wei Zhang, Zhiqiang Bai, and Yuesheng Zhu. An improved approach based on cnn-rnns for mathematical expression recognition. In *Proceedings of the 2019 4th International Conference on Multimedia Systems and Signal Processing*, pages 57–61. ACM, 2019.