Decomposing LaTeX Documents from Geometrical and Logical Layout

Vincent Deo, Terry Kong, Maisy Wieman

Department of Electrical Engineering, Stanford University

Overview

LaTeX is a low-level markup and programming language that produces aesthetically pleasing PDF documents with diverse content. This project aims to analyze the layout of mobile images of LaTeX documents to simplify the generation of LaTeX source code. Input is captured from an Android phone, and the results are computed on a server.
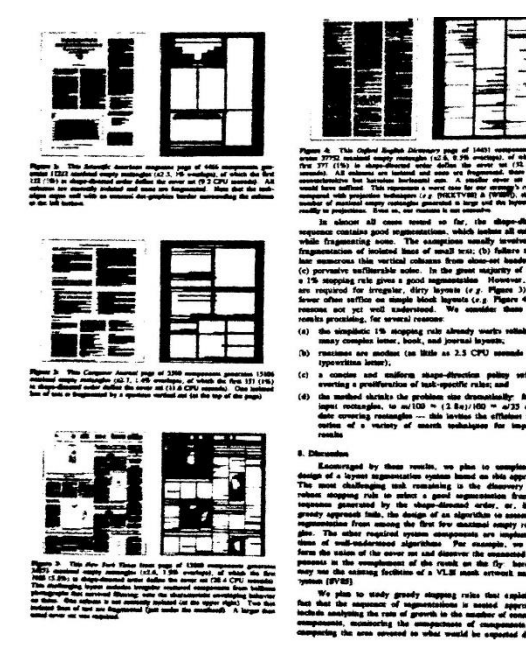
Methodology

Preprocessing
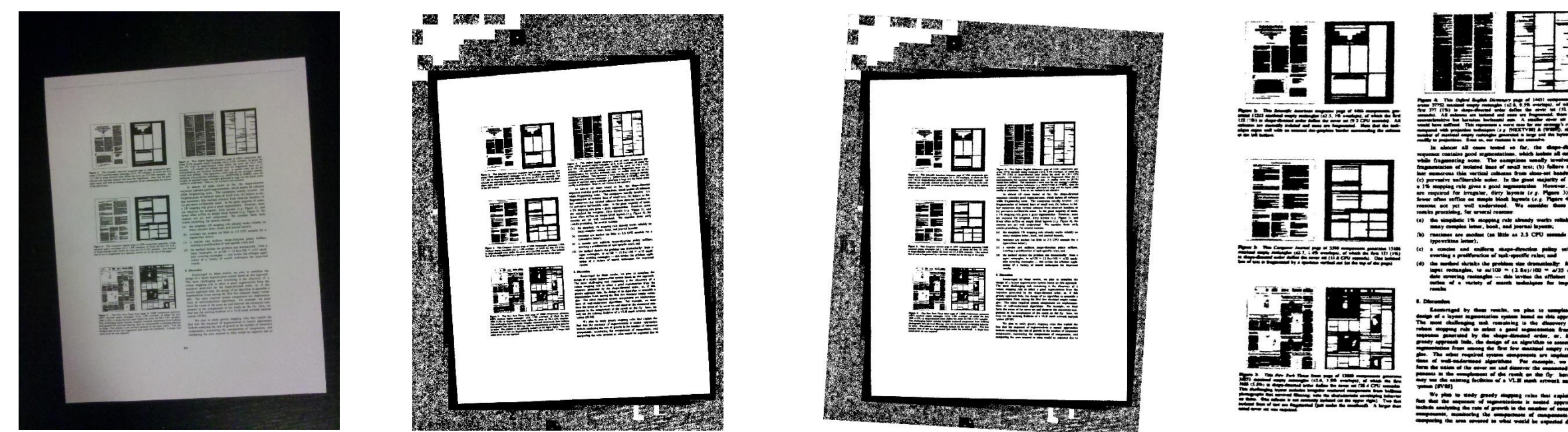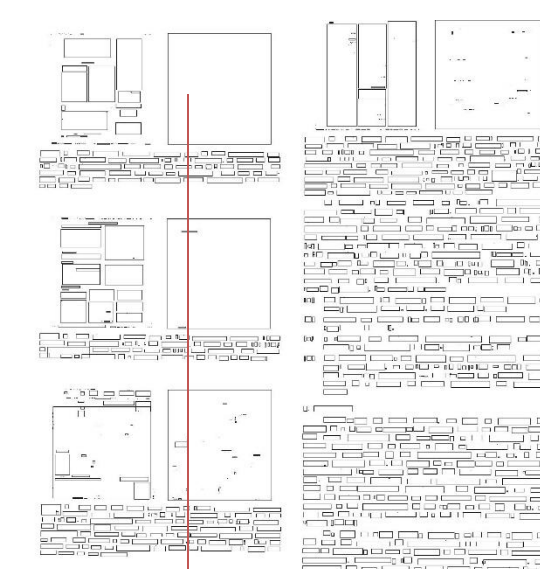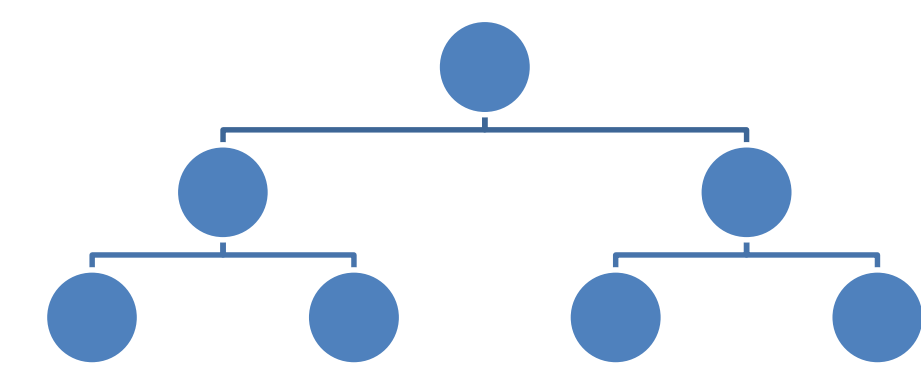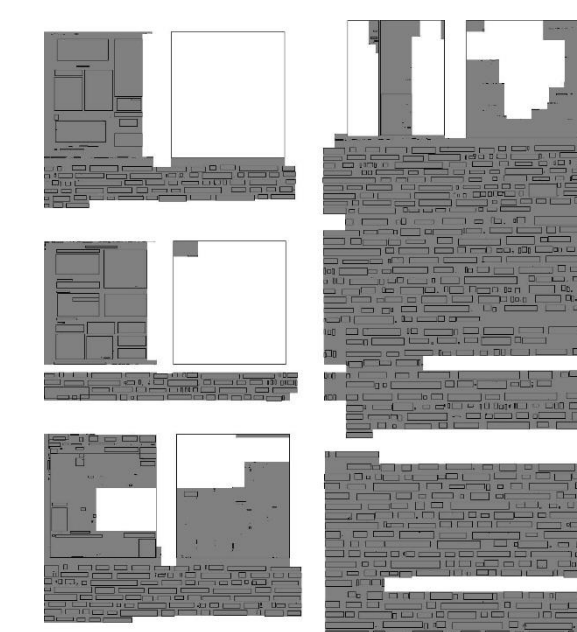

Input
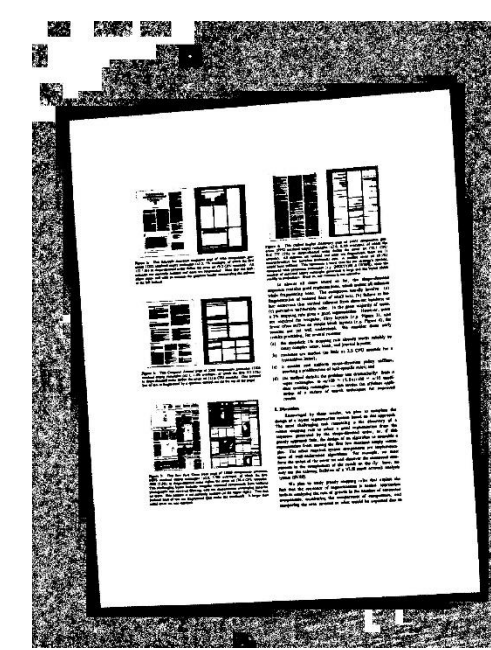
Adaptive Binarization
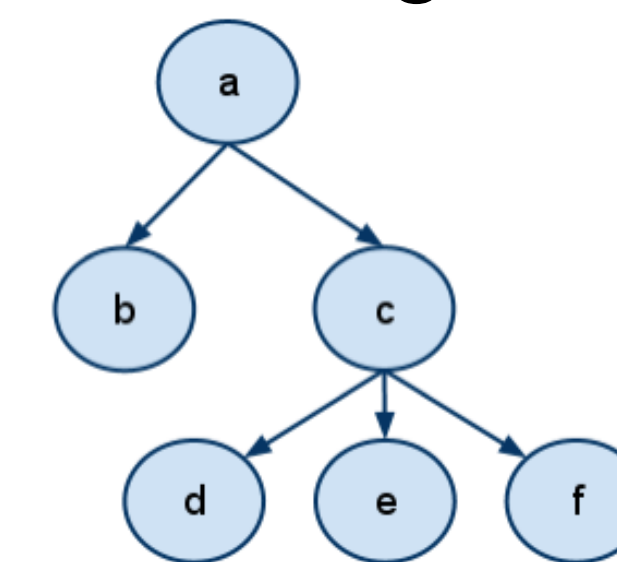

Skew Correction

Margin Removal

Find Maximal White Rectangles

Scan Image for black CCs

Create and balance binary tree of white rectangles

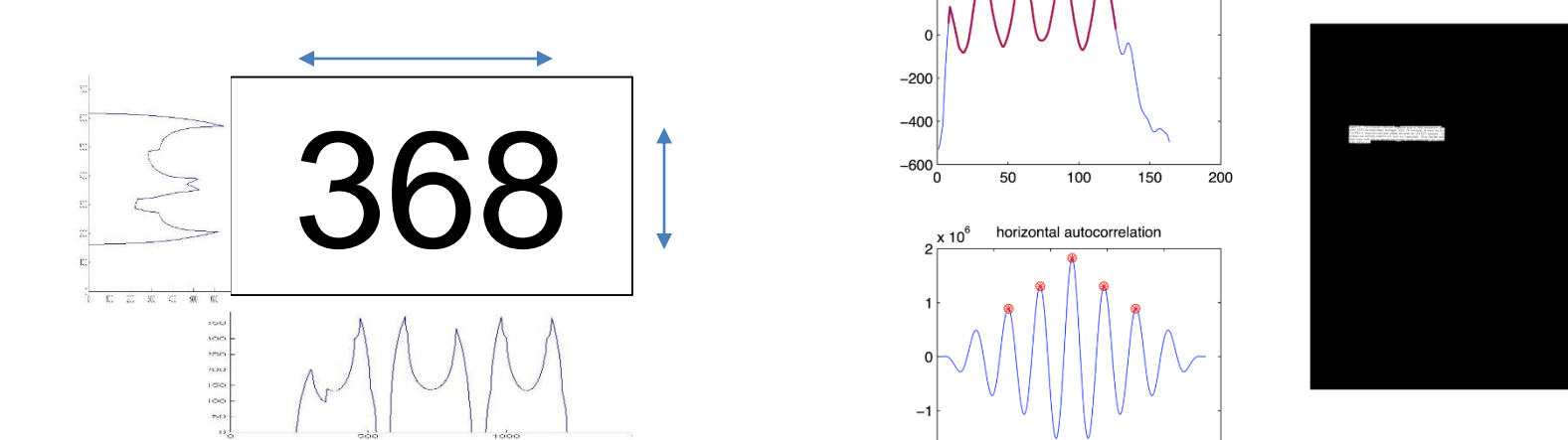Largest 100 Rectangles

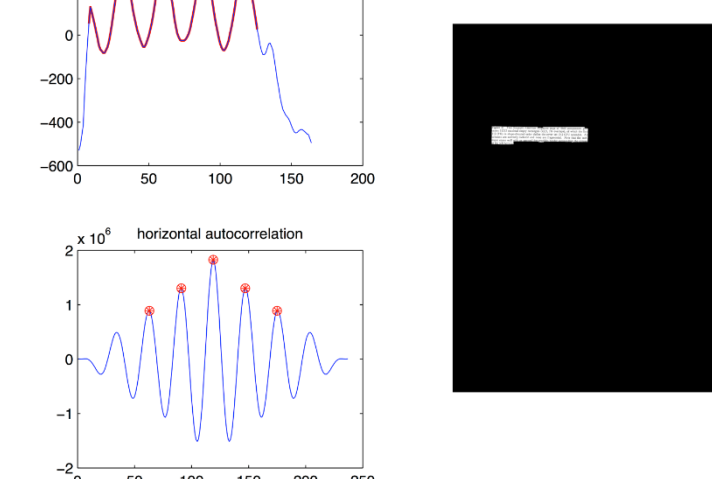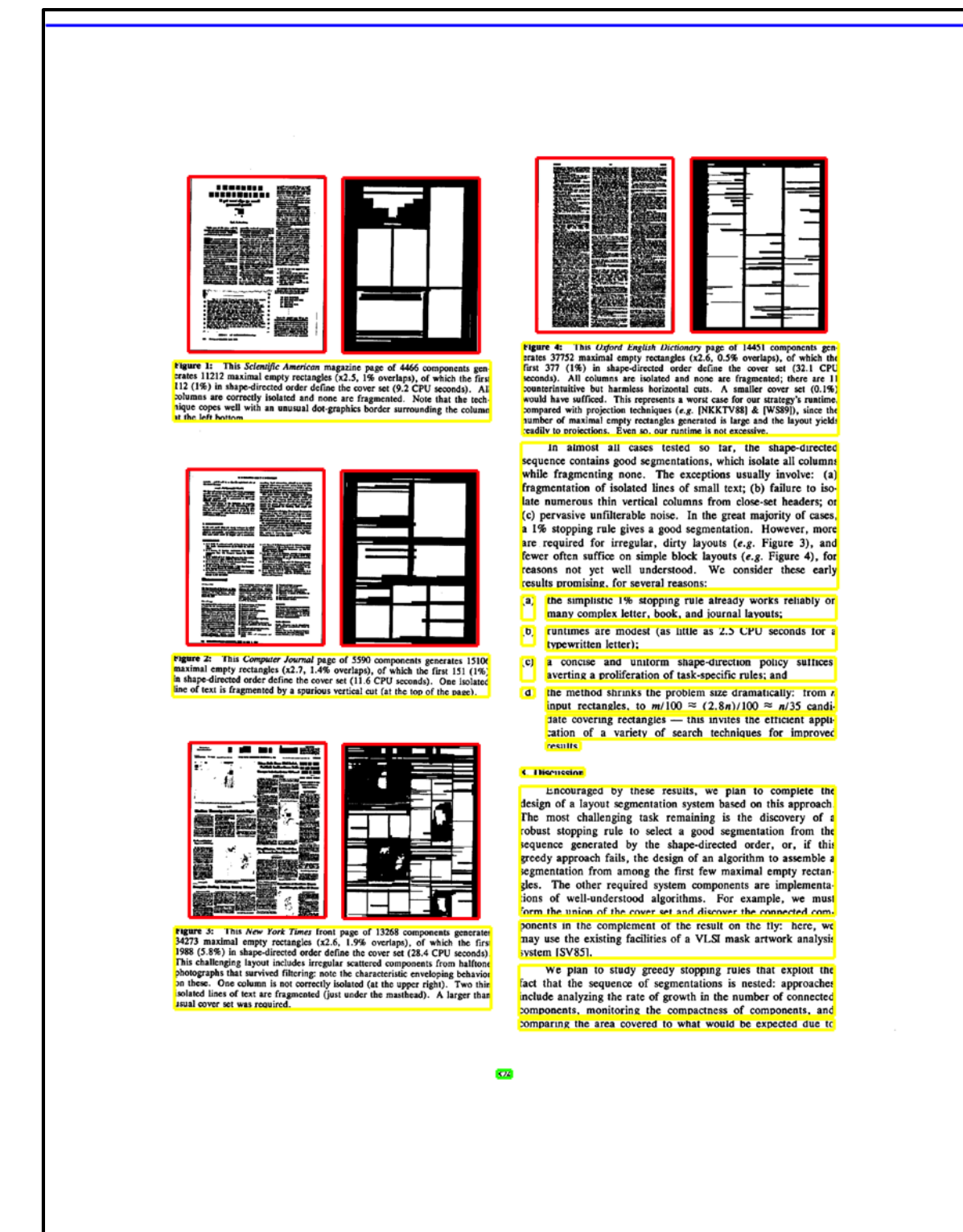**Classify Regions**

Determine CC Statistics

Use Horizontal Projections

Global and Local Feature Comparison

Results

Color	Type
Red	Figures
Yellow	Text
Green	Page No.
Blue	No type

Future Work

1. Use the results of layout decomposition to generate code for a LaTeX template for the given document
2. Add text recognition to generate the entire LaTeX document

References

H.S. Baird, S.E. Jones and S.J. Fortune, "Image Segmentation by Shape-Directed Covers," *Proc. Int',l Conf. Pattern Recognition,* pp. 820-825, 1990.