Document Layout Analysis

By: Garrett Hoch
Document Layout Analysis Overview

- What is Document Layout Analysis
  - Geometric layout analysis
  - Logical layout analysis

- Why is it useful?
  - Done before OCR
  - Gives meaning to text
  - Databases

- Algorithm: Docstrum
Algorithm - Docstrum

1. Preprocessing
2. Detect centroids
3. Determine k nearest neighbors
4. Estimate skew of image
5. Estimate in line and between line spacing
6. Find lines of text
7. Find blocks of text
8. Bounding box calculation
1. Preprocessing

- Convert image to gray scale
- Threshold
- Salt and pepper noise
- Median Filter
- Morphological opening
2. Detect Centroids

- 8-connected components
  - `bwlabel`
- Calculate area and position of centroids
- Filter out large and small centroids
3. K-Nearest Neighbors

- 5 Nearest Neighbor
  - knnsearch
- Calculate Phase
- Calculate Distance
- Longest Part of Computation
4. Estimate Phase
5. Estimate inline and between line distance

- Based on the phase
  - Nearest neighbors that have phase around 0 degrees are inline
  - Nearest neighbors that have phase around 90 degrees are between line
6. Find text lines

- Threshold centroids based on phase
- Transitive Closure
- Linear regression
7. Find text blocks

- Each line is compared to each other
  - If it meets the criteria to be in block then add it to the block
  - Else start a new block

- Sort text lines by:
  - Approximately parallel
    - based on estimated phase
  - Perpendicular distance
    - Based on between lines distance
  - Overlap or parallel distance
    - Based on inline distance

- Customized based on a document to document basis
8. Bounding box

- From the Previous step bounding boxes are drawn for each text block.

- Based on the position and size of a box each box can be labeled as text, equation, equation number, section heading, and etc.
Similar and Dissimilar Document Structure
Discussion

• Pros
  • Can separate analysis into subsection for more accurate results
  • Analysis independent of skew

• Cons
  • The algorithm needs to be customized based on the document
    • Current area of research
  • Nearest neighbor computation is computational heavy

• Future work
  • Need to implement skew estimation
  • Explore more advanced techniques
  • Use in conjunction with OCR
References


Questions?