# Connected Components Based Layout Analysis Approach for Educational Documents

Ruiying Liu[1], Shenbao Yu[1], Fan Yang[1], Yinghui Pan[2] and Yifeng Zeng[3]*

[1]Department of Automation, Xiamen University,

[2]College of Computer Science and Software Engineering, Shenzhen University, [3]Department of Computer and Information Sciences, Northumbria University

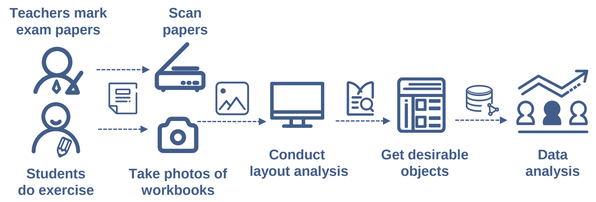
## Introduction

Layout analysis (page segmentation or page object detection), is a fundamental step of document image processing which aims to detect and categorize areas of interest on document images.

Several challenges cause tremendous hardship to propose a universally general method, thus currently most existing researches focus on one specific kind of document.

Existing researches have conducted layout analysis on various documents, but none for documents yielded from teaching.

We consider two practical application scenarios and propose a novel layout analysis system to 1)segment text and non-text areas of workbook pages and 2) extract regions of interest on exam papers. Our system is based on connected component (CC) analysis.




**Main contributions**:

- The first layout analysis system for exam papers and workbook pages
- Only basic digital image processing techniques to trade off efficiency and accuracy; convenient to be implemented in practice
- Experiments on images collected from real-world scenarios and satisfactory results

## Materials & Basis

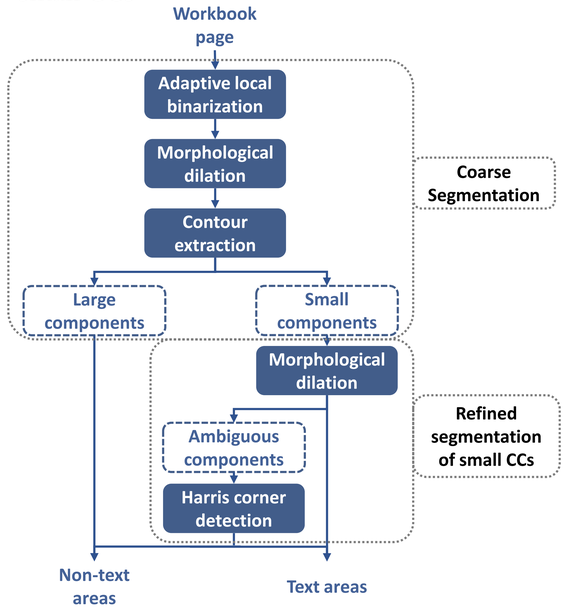
CC analysis is conducted with rule-based heuristics. There are four basic steps:

1. **Binarization** segments the input image into foreground contents and background, generating a binary image.
2. **Morphological operations** transform the binary images to underline desirable objects or filter out useless parts.
3. **Contour extraction** can obtain CCs (there is a one-to-one correspondence between contours and CCs).
4. **CC analysis** aims to analyze geometric features and spatial relations between different CCs to achieve layout analysis.

## Text Segmentation

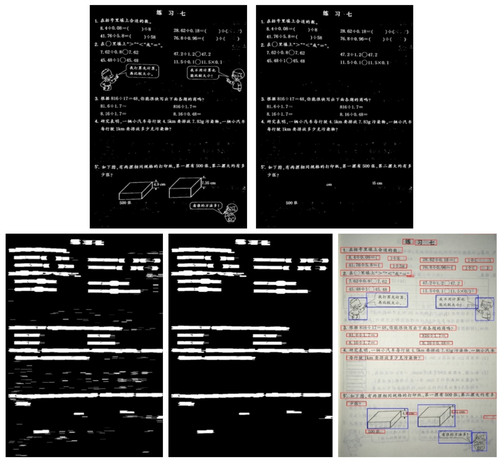
To segment text and non-text areas on workbooks:

- Used two-stage framework: 1) distinguish CCs with size information to get coarse segmentation, and 2) further extract texts from small CCs
- Added some special morphological transformations for better segmentation
- Utilized Harris corners only for identification of small CCs



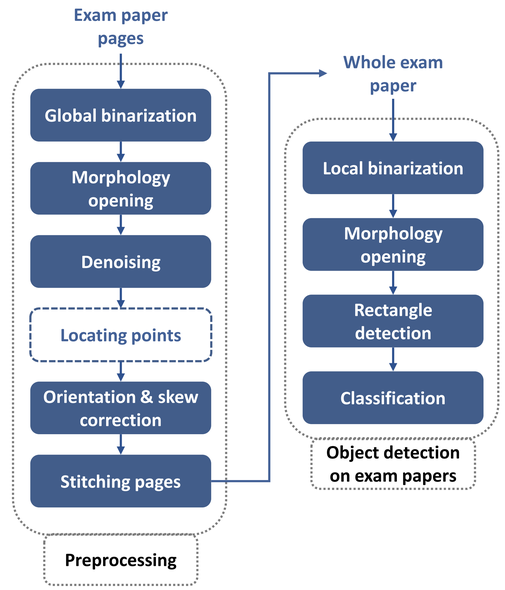

**A process example**:

1. Binarization and morphological dilation
2. Small CCs only (large CCs are classified into non-text areas)
3. Another morphological dilation to get text lines
4. Tiny noises elimination
5. Result: non-text areas and text lines.



## Object Detection

During the entire process from which specific regions of exam papers containing desirable information like student information can be extracted, the inputs experience two sub-modules.




**Preprocessing**

The input is two side pages(scanned), so need to correct them respectively and then stitch them together:

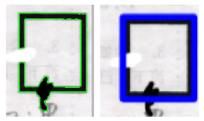
- Correction: rotates images to correct orientations and eliminates probable skews or tilts.
- Stitching: gets the whole exam paper for final detection

**Detecting Regions of Interest**

On exam papers, desirable information is surrounded by similar thick rectangles.

Steps:

1. Local binarization: the Gaussian thresholding
2. Extract thick CCs: transforms images by opening to filter out all thin lines with the structural element of which the side is $min/30$.
3. Rectangle detection: for incomplete bold box, combines polygonal approximation and corner analysis instead of using contours to extract the whole rectangular region successfully

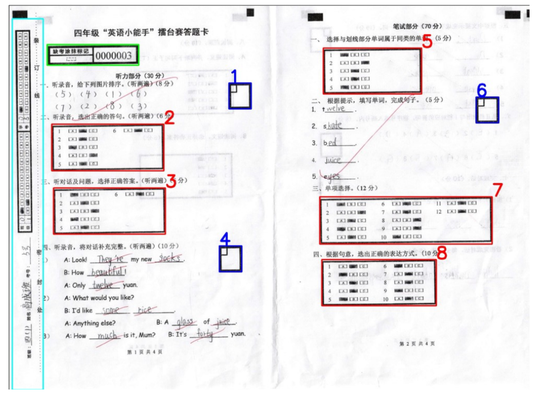


4. Classifies these boxes based on their geometric features and coordinates

## Experiment Results

**On exam papers**

- 662 marked papers covering multiple subjects
- Objects: student information region, exam paper information, answer sheets of objective questions and score boxes of subjective questions




- Precisions for all categories above 0.98

**On workbook pages**

- We took photos of 118 pages from primary school Math workbooks.
- From 111 pages, all text and non-text areas were successfully detected.
- Only some acceptable mistakes happened on other images.

## Future Works

1. Further detection on non-text areas containing texts
2. Handwriting detection, an interesting yet challenging task, in exam paper processing
3. A versatile and generic layout analysis system facilitated with sophisticated machine learning algorithms

## References

[1] Galal M Binmakhashen and Sabri A Mahmoud. Document layout analysis: a comprehensive survey. *ACM Computing Surveys (CSUR)*, 52(6):1–36, 2019.

[2] Nikos Vasilopoulos and Ergina Kavallieratou. Unified layout analysis and text localization framework. *Journal of Electronic Imaging*, 26(1):013009, 2017.

[3] Syed Saqib Bukhari, Faisal Shafait, and Thomas M Breuel. Improved document image segmentation algorithm using multiresolution morphology. In *Document recognition and retrieval XVIII*, volume 7874, page 78740D, 2011.