

TABLESNAP

Table Detection

ASHLEY BREA, ASHLEY KIM, DIANE PARK

OVERVIEW

- Originally wanted to be able to digitize handwritten data tables
- To work our way there, we will first focus with *table structure and object detection*



Object
Detection



Row and
Column
Detection



Run OCR on the
handwriting



Present a fully
digitized data
table



PREVIOUS SOLUTIONS

Deep learning for table detection and structure recognition: A survey
(Kasem et al., 2022)

Prasad et al. proposed **CascadeTabNet**, a *R-CNN* that identifies table areas along with individual table cell structures.

Siddiqui et al. created a network, a combined *R-CNN* and *FPN* that works on photos instead of just PDFs.

GENERAL TABLE DETECTION DATASET

Figure 5-6: 截至 2018 年 9 月末发行人主要参股公司情况

序号	被投资单位名称	主营业务	注册资本	出资比例		核算方式
				出资	实缴	
1	江苏五峰电子科技股份有限公司	电子科技	5,000.00	25.00%	权投注	
2	南京新联电子有限公司	电子科技	500.00	30.00%	权投注	
3	南京国城物业管理有限公司	物业管理	200.00	40.00%	权投注	

发行人主要参股公司情况介绍：

(1) 江苏五峰电子科技股份有限公司

江苏五峰电子科技股份有限公司成立于 2011 年 11 月 1 日，注册资本 5,000 万

Figure 5-6: 截至 2018 年 9 月末发行人主要参股公司情况

序号	被投资单位名称	主营业务	注册资本	出资比例		核算方式
				出资	实缴	
1	江苏五峰电子科技股份有限公司	电子科技	5,000.00	25.00%	权投注	
2	南京新联电子有限公司	电子科技	500.00	30.00%	权投注	
3	南京国城物业管理有限公司	物业管理	200.00	40.00%	权投注	

发行人主要参股公司情况介绍：

(1) 江苏五峰电子科技股份有限公司

江苏五峰电子科技股份有限公司成立于 2011 年 11 月 1 日，注册资本 5,000 万

Figure 5-6: 截至 2018 年 9 月末发行人主要参股公司情况

序号	被投资单位名称	主营业务	注册资本	出资比例		核算方式
				出资	实缴	
1	江苏五峰电子科技股份有限公司	电子科技	5,000.00	25.00%	权投注	
2	南京新联电子有限公司	电子科技	500.00	30.00%	权投注	
3	南京国城物业管理有限公司	物业管理	200.00	40.00%	权投注	

发行人主要参股公司情况介绍：

(1) 江苏五峰电子科技股份有限公司

江苏五峰电子科技股份有限公司成立于 2011 年 11 月 1 日，注册资本 5,000 万

Table 2.6: High online service performance relative to income

Country	Online Service Index	Income group
Rwanda	0.5718	Low
Colombia	0.7636	Upper Middle
Ethiopia	0.4547	Low
Kazakhstan	0.7680	Upper Middle
Morocco	0.6729	Lower Middle
Kenya	0.4252	Low
Sri Lanka	0.6535	Lower Middle
Malaysia	0.6772	Upper Middle
Tunisia	0.6378	Upper Middle
Mongolia	0.6742	Lower Middle

Table 2.7: Low online service performance relative to income

Country	Online Service Index	Income group
Equatorial Guinea	0.0115	High
Monaco	0.2205	High
Urya	0.0157	Upper Middle
Saint Kitts and Nevis	0.1339	High
San Marino	0.2756	High
Tuvalu	0.0394	Upper Middle
Barbados	0.2205	High
Algeria	0.0787	Upper Middle
Sao Tome and Principe	0.0079	Lower Middle

Table 2.6: High online service performance relative to income

Table 2.7: Low online service performance relative to income

Table 2.6: High online service performance relative to income

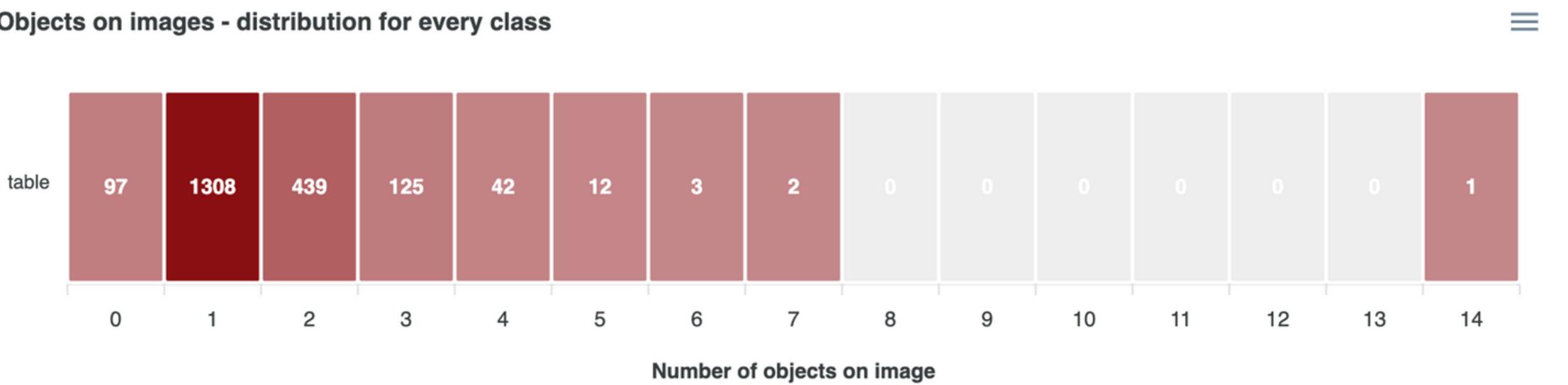
Country	Online Service Index	Income group
Rwanda	0.5718	Low
Colombia	0.7636	Upper Middle
Ethiopia	0.4547	Low
Kazakhstan	0.7680	Upper Middle
Morocco	0.6729	Lower Middle
Kenya	0.4252	Low
Sri Lanka	0.6535	Lower Middle
Malaysia	0.6772	Upper Middle
Tunisia	0.6378	Upper Middle
Mongolia	0.6742	Lower Middle

Table 2.7: Low online service performance relative to income

Country	Online Service Index	Income group
Equatorial Guinea	0.0115	High
Monaco	0.2205	High
Urya	0.0157	Upper Middle
Saint Kitts and Nevis	0.1339	High
San Marino	0.2756	High
Tuvalu	0.0394	Upper Middle
Barbados	0.2205	High
Algeria	0.0787	Upper Middle
Sao Tome and Principe	0.0079	Lower Middle

DATA EXPLORATION

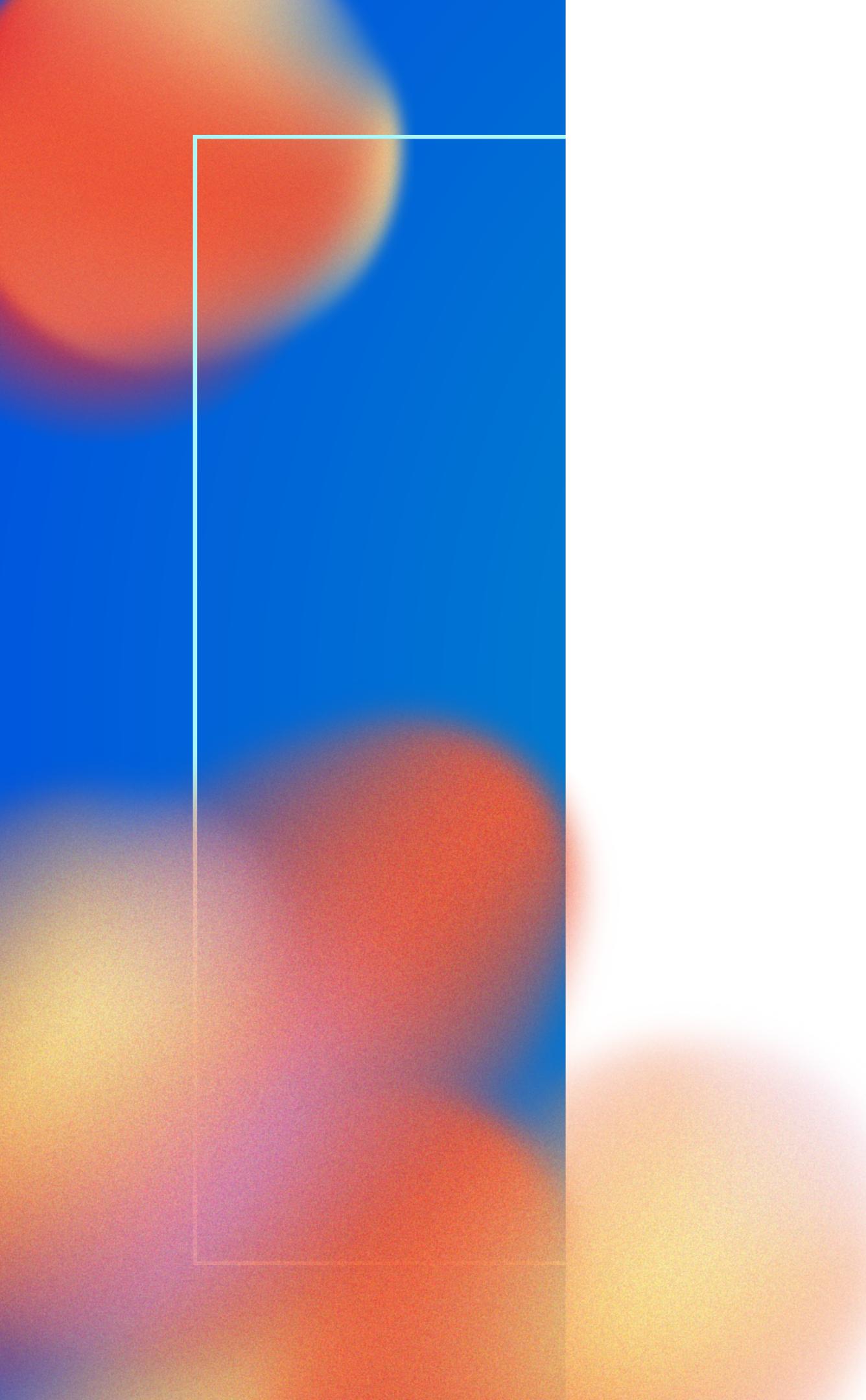
- A total of 2029 samples consisting of mixed file types, e.g. PNG, JPG, etc.



- However we only want to use the examples with 1 table per page, which reduces the set to 1308 samples
- The dataset is compiled of the `train_annotated.csv` and `train_folds.csv`, and `train.csv`

DATA CLEANSING

- Using `train_annotations.csv` and `train_folds.csv`
 - `train_annotations.csv`: boundingbox values in a 4-tuple
 - `train_folds.csv`: number of tables found in image
- Only want to focus on images with 1 table
 - Filtered data within `train_folds`, then within `train_annotations`
 - Left us with 1308 values for single tables



DATA PREPARATION

- X: Image ID
- Y: Bounding-Box values (`x_min`, `y_min`,
`bbox_width`, `bbox_height`)
- Split Data:
 - Training: 70%
 - Validation: 20%
 - Testing: 10%

CHALLENGES

- Identifying tables with no borders
- Using a small dataset so our training may struggle to converge.
- Different file formats: .jpg and .png files

THANK YOU!

ANY QUESTIONS ?